FITTING MODELS (CENTRAL TENDENCY)

SOMETHING IN THE MIDDLE

Statistical models #116
From the dead #116
Why do we need statistical models? #117
Sample size #118
The one and only statistical model #120
Central tendency #124
The mode #126
The median #128
The mean #129
The ‘fit’ of the mean: variance #135
The fit of the mean #136
Estimating the fit of the mean from a sample #141
Outliers and variance #148
Dispersion #149
The standard deviation as an indicator of dispersion #149
The range and interquartile range #151
Key terms #155
JIG:SAW’s puzzles #155
Fitting models (central tendency)

Somewhere in the middle

Statistical models

From the dead

Why do we need statistical models?

Sample size

The one and only statistical model

Central tendency

The mode

The median

The mean

The 'fit' of the mean: variance

The fit of the mean

Estimating the fit of the mean from a sample

Outliers and variance

Dispersion

The standard deviation as an indicator of dispersion

The range and interquartile range

Key terms

JIG:SAW’s puzzles

4
I stopped in my tracks and turned to face the cat from Professor Pincus's office. He was licking his chest in that over-purposeful way that cats do when they feel awkward. I had managed to convince myself that I had probably imagined the cat talking in Professor Pincus's office, but unless I’d lost my mind twice it really could talk. I wondered whether Alice knew that she worked somewhere that had a talking cat? If she did then why hadn’t she mentioned it to me? If I met a talking cat at our rehearsal studio the first thing I would do is tell her. I’d call her now if she hadn’t been wiped from the planet.

The cat stared at me, waiting for my reaction. How could he talk, and why was he claiming to be a dead scientist? If you were going to claim to be a dead scientist wouldn’t you pick one who was a bit more popular than the reality prism dude? I had so many questions to ask, but no time to get side-tracked by them.

‘That’s impossible,’ I eventually said, ‘Milton Grey died five years ago.’ I was aware that this wasn’t the most impossible aspect of the situation, but I had to start somewhere.

The cat’s rate of cleansing rapidly increased before he stopped and looked up. ‘Officially I did, but only because the alternative was to try to explain to everyone that I had become a cat. It transpires that the WGA are reluctant to issue passports and other legal documents to cats, and you can forget anything involving a fingerprint or retina scan. Without official documents I am effectively dead, but with Catherine’s help I exist as a cat. It has considerable benefits: I get a lot more time for research, because only a mad person would email a cat.’

If someone told me about a cat that could email, I’d definitely message it, I thought to myself. How could a cat be Milton Grey, I wondered? Had his brain been transplanted? That seemed unlikely: a cat’s skull was too small for a human brain. Maybe the brain had been shrunk? Was it some kind of mind merge? Was Milton Grey’s human body walking around somewhere with the
mind of a cat inside it? I needed to establish whether I could trust him, and to do that I needed to know his story, which he was happy to tell.

‘After the reality prism nonsense I found myself stripped of my position within the School of Physics here at the University of Elpis, and my reputation made me unemployable. The WGA offered me a position in their Scientific Division, where I helped them to develop technology to rebuild society. Did you know that my colleague Roediger and I invented steam fusion?’ The cat paused to take in the awe that he expected from me at this revelation. He looked offended when he got none, and licked his chest to comfort himself. Steam fusion drove so much of the modern world that I took it for granted. I’d never considered what lay beneath it, or how hard it might have been to invent.

‘I was quite happy at the WGA,’ Milton continued, ‘but as their power increased I felt that they were losing their humanitarian principles, so I resigned. It was not easy to convince the WGA that the person they believed brought down society should be given his freedom. They allowed me to go because of my loyal service, but they monitored me constantly: my freedom was an illusion. Worse than that, even the passage of time had not weakened society’s ire, and universities still stayed well away from me. I suspect the WGA applied some pressure in that respect. Catherine – Professor Pincus – gave me a lifeline. Although known for my work in physics, for many years I pursued a side interest in genetics. I published an obscure paper on how – theoretically at least – principles of physics could be applied to genetic problems. Catherine was fascinated by the potential of the theory and offered me an opportunity to pursue it at the Beimeni Centre. Catherine put her neck on the line to employ me: in doing so she put the Centre under the WGA’s spotlight. I owe her a great deal for her faith, and her unwavering principle to prioritize knowledge and humanity. The first fruits of my work were what I jokingly called the “gene mixer”, which was a device for repairing damaged cells; for example, cancer cells. It was a simple enough idea: you turn the damaged cells into programmable matter and then give them a “program” based on healthy cells.’

‘That sounds the same as something of Alice’s that I read.’

‘Alice’s work is much more advanced than my gene mixer.’

So he did know Alice. Why hadn’t she mentioned him to me?

‘The gene mixer was crude, difficult to control, and too powerful – a blunt instrument that rampaged through an organism affecting any gene in its path. We have subsequently discovered that you need the human mind to control the process, but at the time, the gene mixer had promise. We had very encouraging preliminary results, but we needed to test it on a human. You cannot go sticking humans in gene mixing machines, not even ones with cancer who you are trying to help. If it goes wrong you might make the disease worse or even kill them. Fortunately, there was one person we knew who had a recent diagnosis of cancer who was willing to take that risk.’

I panicked that it was Alice, before realizing that he was talking about the past, a time before Alice would have worked there. I asked who it was.

‘Me. I stepped under the machine, gathered a dish of healthy cells and prepared to have my diseased cells replaced with programmable cells that would take on the behaviour of those healthy cells.’

‘What happened?’

‘My cat, Erwin, happened. The dish of healthy cells looked a lot like a dish of milk. Like most cats, Erwin was fond of milk. At the point of transmission he jumped onto the table, lapped up the healthy cells, and became caught in the transfer beam. The machine transformed my cells to programmable matter, took Erwin’s healthy cat cells and implanted that template into my cells. I became a cat.’
'Over the wall, man! That sucks!' 'On the plus side, my cancer was cured.'

I blurted out a barrage of questions: What happened to the cat? Did Milton's personality change at all? Did the cat's? Milton looked a little upset at my concern for his cat over him. 'Erwin is fine – the transfer was in one direction, so he was unaffected – apart from discovering, much to his disgust, that there was another cat in the house. Cats are not willing cohabiters, so after a few weeks he wandered off to find a new owner. The most surprising thing was that CAT scans – he paused and smiled at the pun – 'showed that physically my brain had shrunk, but also had become incredibly dense to retain everything within my human mind. I did pick up some cat habits, but on the whole it is a good life: people tickle my chin a lot more than they used to.'

Professor Grey looked at me with vulnerable eyes and I instinctively stroked his chin, pulling away as I realized I was fondling a professor's beard. I still wasn't sure whether to buy his story, but I couldn't deny the evidence of my eyes: he was a talking cat. Why would he help me, though, and how could I trust someone who had destroyed society? I asked him whether he felt bad about the Reality Revolution. 'I created the technology, but does that make me responsible for how people used it or their inability to deal with the consequences? Many consider me the modern-day Robert Oppenheimer …' 'Who?' He gave me a look of disgust, '... the scientist in charge of the creation of the first atomic bomb. I would argue that, unlike Oppenheimer, the consequences of my work were not obvious to me or anyone else. Do I feel bad for inventing the reality prism? No. It was a device that could have been used for great good. Do I regret how people used reality prisms? Yes, but I blame them, not myself. Do I think society is a worse place for the Reality Revolution? For many years it became a more humane place, of that I am certain, but as through all of history it is only a matter of time before someone tries to take advantage of humanity.' Milton looked sad. As detached as he felt from the consequences of his work, he didn't sound evil. I asked him why he wanted to help. 'I have a reputation that makes my apparent death quite convenient for Catherine. It is bad enough that the Beimeni Centre employed the “notorious” Professor Grey, inventor of the reality prism and destroyer of modern civilization, but society forgives them because they seemingly killed me as part of a bizarre genetics experiment. The world believes that justice has been served. Imagine the furore if anyone discovered that instead of being dead I was in fact a rather lovable cat. The Centre would be investigated and probably shut down for good. As I mentioned, the WGA kept a particularly close eye on things while I was working here. Since my “death” they have lost interest in us, and Catherine prefers it that way. I, however, do not, because my activities are restricted. Before I actually die I want to be remembered for something good, I want to erase the false belief in my part in destroying society. As I told you, my work with the gene mixer was getting close to a device that could cure a great many fatal diseases. Perhaps if I can bring that work to fruition the world will forgive me – I can die a hero, not a villain. I can't do that while I'm trapped here, only knowing about the work being done in Catherine's lab. I am forever grateful for all that she has done for me, but I need to expand my horizons. To do that I need someone I can trust. Can I trust you, Zach?' It was the same question I'd wanted to ask him. 'Of course, I'm a good guy, but why do you think I need your help?' 'Sooner or later you are going to decide that you need to investigate JIG:SAW – '
‘Not after what the Prof said,’ I interrupted.
‘True, only a fool would go there, but sooner or later the fool inside will get the better of you. Catherine is right, it will destroy you; but with my help, you have hope.’

4.1 STATISTICAL MODELS

4.1.1 From the dead

I picked Milton up, put him in my backpack and headed swiftly out of the building. I sheepishly passed the front desk hoping that the head there wouldn’t have a change of heart about having let us in and call security. She didn’t, but as we moved away from the building my pace quickened until I had broken into a run. I decided to head back to Janus, the area where I lived, and go to Occam’s café. Janus was the liberal part of town; if I could take a talking cat anywhere, it was there. Janus was quite a way from Veritas, where the Beimeni Centre was, so I hailed a bubble. Personal transport like cars were a thing of the past; instead the city provided clockwork fusion-powered ‘bubbles’ that hovered around the city. These transparent spheres could be hailed through memoryBank or a Proteus and arrived almost instantly. Once inside, you told it where to go and it hovered you there. We were soon at Occam’s.

It was early evening and I was starving. Occam’s was a minimalist café that invented ‘parsimoniaus cooking’, in which only essential ingredients are used. It tasted better than it sounded. Even in the ‘open-minded’ part of town I was worried that taking a cat to a café might stretch people’s tolerance, so I sat on a long leather couch in one of the many private booths where we’d be less conspicuous. I placed my backpack beside me and Professor Grey wriggled out.

‘I’ll have a tuna steak please,’ he said. ‘Now, on to business. Tell me what you know about Alice.’

Hearing him speak, in public, made me suddenly self-conscious and scared that someone would discover my talking cat. I understood how Professor Pincus must have felt earlier on when Milton had walked into her office. I told Milton about how Alice had disappeared, how I couldn’t contact her, and about my visit to Dr Genari.

‘Who is this Genari chap?’ Milton asked. As he did, I remembered how I had seen a cat outside Genari’s window when I was there. Now I thought of it, the cat was ginger. So that was how Milton could send me messages relating to what Genari was saying: he must have been listening outside. Why would he have known about me? Why would he have followed me? Did Alice send him? Also, Milton said he needed me to help him escape from the Beimeni Centre, but if he was at Genari’s then he must be free to come and go from the Centre as he pleased – not the prisoner he made himself out to be. Why was he lying to me? I decided to play along for now rather than confront him: there must be a reason why and the best way to find out was to play dumb.

I told Milton my theories about Alice, about how at first I thought she’d left me, but that after visiting Genari I feared that she was in trouble. I told him how Professor Pincus had helped me to look at data to test this hypothesis, and that she had seemed disappointed when the data we looked at didn’t support that idea – as if she’d hoped to convince me that Alice had left. Revisiting the last 36 hours or so made me feel anxious and panicky: why couldn’t I just speak to Alice? I wanted to hear her voice, to have her reassure me that she was alright and that this was
just some silly misunderstanding. I called her Proteus again and the same cold voice as before told me that her number did not exist.

Milton watched me with interest. I couldn’t make out what he was thinking, but you never can with a cat. Finally, he said, ‘It doesn’t sound as though Catherine, Professor Pincus, really did a good job of evaluating the evidence from Alice: she just looked at general data about what women look for in a man. Perhaps what Alice looks for is a nincompoop like you.’

‘We did look at her score on the Relationship Assessment Scale,’ I replied. ‘Oh, and what note is this?’ I sang an F#.

‘A C?’ Milton guessed.

‘No, it’s an F# – nincompoop.’

There was an awkward silence while the cat considered his options. Having considered them, he carried on as though nothing had happened. ‘You have more data from Alice than you showed Catherine?’ he asked.

I showed Milton all of the information from Alice’s file. He took it all in, swatting at my diePad screen and screwing up his face in concentration. He licked his lips, again looking like he was trying to decide what to do.

‘What you have here’, he said, ‘are competing hypotheses: she has left you and doesn’t want you to find her, or she is in trouble and someone else doesn’t want you to find her.’ With horror I suddenly had the thought that perhaps Milton and Catherine were the people who didn’t want me to find her. ‘What you need to do is to learn some statistics and put these hypotheses to the test.’

The cat had to be kidding. I hated maths at college, and I wasn’t about to voluntarily enrol on a stats module taught by a talking cat in a café. ‘Why would I want to do that?’ I asked.

‘Because statistics is cool, that’s why. Now listen up, you have two choices. I can send you away to read a book, like the now legendary, genre-defining Discovering Statistics … series by Andy Field …’

‘Alice reckons those books are rubbish,’ I interrupted.

‘In which case, it’s choice number two: listen to me. Oh, and one more thing: nobody can find out who I really am, from now on you refer to me only as Milton. Now go and order my tuna.’

4.1.2 Why do we need statistical models?

Upon my return Milton began his lesson. ‘Scientists are interested in the true state of the world.’

‘Yeah, Alice mentioned this to me the night before she vanished. She said that scientists want to find out the truth, but because they can’t usually get data from the entire population they work with smaller samples from that population.’

‘Ah, you know the basics already. That is excellent, but you need to know where statistical models fit into all of this. I’ll begin with an analogy. You remember the Beimeni Centre building: it looks like three cogs on a spiralling tower of human DNA?’

‘Oh, is that what it was supposed to be? I thought it was just a weird spirally coggy thing.’

Milton raised his eyebrow. ‘Do you think that the architect who designed it just turned up one day with a truckload of materials and built it?’

‘I doubt it. Don’t they do drawings and models and stuff?’

‘Purr cisely. It is expensive and impractical to build the “full building”, so instead they build a scaled-down version. This model can differ from the real building in several ways – it will be smaller for a start – but the architect or engineer will build a model that resembles the real one as closely as
possible. Once the model is ready, it can be used to estimate what would happen to the real building. For example, they might place it in a wind tunnel to see whether the building is likely to withstand the winter Elpis winds, or they might simulate an earthquake to see if the structure remains intact. To get accurate predictions about the real building, it is vital that the model resembles it as closely as possible, otherwise the architect’s predictions about the real building will be wrong.

‘Scientists do much the same. They are trying to make predictions about something in the real world – it might be a psychological, societal, biological or economic process to which they do not have direct access. Because they cannot access the process directly they gather data and construct small-scale models of the process and use them to predict how these processes operate under more general conditions. Just like the architect, we want our small-scale model to resemble the real situation as closely as possible so that the predictions we make about the real world are accurate; the statistical model we build must represent the data collected (the observed data) as closely as possible. The degree to which a statistical model represents the data collected is called the fit of the model.’

Milton started scratching out a picture in the wooden table with his claw. The noise set my teeth on edge and made me really uncomfortable, but not as uncomfortable as I felt about the criminal damage he was-inflicting on the table. ‘Look,’ he continued, ‘I’ve drawn three models that an architect might have created before building the Beimeni Centre. The first model is an excellent representation of the real building: it is a good fit. If the architect uses this model to make predictions about the real building then, because it so closely resembles reality, these predictions ought to be accurate: if the model collapses in a strong wind, then there is a good chance that the real building will too. The second model has some similarities to the real world: the model retains some of the basic structural features, but there are some big differences too (e.g., two of the floors are missing). This model has moderate fit: there are some similarities to reality but also some important differences. If the architect makes predictions about the real world based on this model, these predictions could be inaccurate or even catastrophic. For example, the model survives a simulated earthquake but the real building doesn’t. The final model bears few structural similarities to the real building – for a start, it is missing the whole top half of the building – and is a poor fit. Any predictions based on this model are likely to be completely inaccurate.’

‘Is what you’re saying that because we can’t get data from the whole population and fit a statistical model to it, we instead use samples, and we look at what happens within those samples?’

‘Yes, we fit a statistical model to the data from the sample.’

‘And that model can be a good or bad fit to the data?’

‘Yes, and a good or bad fit to the reality in the population! It’s very important to know how well the model fits the data.’

‘Is this because – as Alice told me – samples vary from each other; even if you measure the same thing you can get different results in different samples?’

‘That’s true, but also not all samples are equal; we can have greater confidence in some. Now, where’s my tuna steak?’

### 4.1.3 Sample size

The waiter arrived with our food and placed both plates in front of me, assuming that I wasn’t crazy enough to buy café food for a cat. The smell made me realize that I hadn’t eaten since the morning, and I savaged my food.
4.1 Statistical models

Figure 4.1  Fitting models to real-world data (see text for details)

‘Try not to spit food at me, it is undignified,’ Milton said as he licked a strand of purry cat dribble from his chin. He picked at his tuna, continuing the conversation as he chewed. ‘I have another analogy for you: imagine you are interested in a phenomenon that is trapped inside a cardboard box.’
Milton scratched an image into the table. It was astonishing that a cat could draw so well. Scientists weren’t known for their artistry and I found myself wondering whether Milton’s cat might have had some artistic flair that had been wasted on a cat and could now be unleashed with the benefit of a human mind. Milton clawed my shirt to bring my attention back to him.

‘The population is this big cardboard box,’ he said, pointing at a well-drawn cardboard box. ‘You can’t see inside the box, but you know that it contains an interesting phenomenon. You really want to see that phenomenon, but you cannot because it is hidden in the box. A very handsome talking cat appears with a suitcase of magical discs that he calls samples. When you stick these discs on any solid object it allows you to see through that object for a brief time before the disc vanishes. After admiring the cat’s handsomeness for some considerable time, you think, “Hmm, I could stick one of those discs on the side of the box and see what is inside”. It turns out that these magical discs come in three different sizes: a small sample, a medium sample and a large one. First, you take a small sample and stick it to the side of the box. You can make out a little of what is in the box, but not much. That sample vanishes, so you take another small one and stick it to a different part of the box. Again, you can see a little, but what you can see is quite different from when you looked before because the hole is so small and you stuck it to a different part of the box. You take a final small disc and stick it to another different part of the box. You can again see a little of what’s inside, but what you can make out is different from the other two times. You end up being not too sure what is in the box because each time you looked you saw something different. You now take three medium-sized samples. The first one gives you a much better view inside the box than the small sample: because it is bigger you can see more of what is inside the box. When you place the second and third samples on different parts of the box, you get a different view each time, but there is quite a lot of overlap with what you saw with the other two samples (because each sample shows you a reasonable amount of the inside of the box). Finally, you use the three large samples in the same way. These show you almost the entire inside of the box and you get a very precise view of its contents: it is the staggeringly handsome cat! With the other two large samples it doesn’t matter where you place them on the box – you get a very similar view of the phenomenon, because each time you can see almost the entire inside of the box.’

‘So what you’re saying is that small samples don’t give you a good “view” on the phenomenon in the population, and different small samples are likely to show different things, but with big samples you get a better “view” of the phenomenon that you’re trying to study and what you “see” is likely to be similar across these big samples.’

‘That is exactly my point.’

‘So, how large is a large sample?’

‘That very much depends on the situation, but for now, just focus on the idea that with samples bigger is better. The same cannot be said for dogs.’

4.1.4 The one and only statistical model

I understood what the cat was saying about architects and models, but I didn’t really get what he meant by a statistical model. I mean, you can’t build a bridge out of statistics, can you? I asked him to explain.

‘A statistical model is an equation that describes the phenomenon of interest.’
I hate equations. I was never any good at maths at college – I used to get Alice to help me out. Her name triggered a sick feeling; I wished I hadn’t eaten.

'I will keep it simple for you and purrhaps you will understand.' Milton scratched an equation into the table. 'Most statistical models are a variation on this one equation. This equation means that for a given person, \( i \), we can predict their score on an outcome variable from the model we choose to fit to the data plus some amount of error that is associated to that prediction (for person \( i \)).

\[
\text{outcome}_i = (\text{model}) + \text{error}_i
\]  

(4.1)
‘For example,’ Milton continued while adapting the equation, ‘I noticed some Relationship Assessment Scale (RAS) scores in Alice’s file. Imagine that we asked a sample of 100 people to fill in the RAS so that we had a score for each person. The outcome would be the RAS score for person \( i \), so the equation means that for a given person, \( i \), we can predict their score on the RAS from the model we choose to fit to the data plus some amount of error that is associated to that prediction (for person \( i \)).’

\[
\text{RAS}_i = (\text{model}) + \text{error}_i
\]  

(4.2)

‘Why is there error? Isn’t that a bad thing?’

‘Error is inevitable because by fitting a model we simplify the data to give it structure and order, to make sense of it, to reduce it down to a summary that encapsulates a truth about the state of the world. The price of reducing the raw data to a summary is that it will not perfectly reflect every entity in the sample: for some, the model will predict their RAS score very accurately (the error associated with that prediction is small), but for others the prediction might be terribly inaccurate, and there will be a large error associated with predicting that score.’

‘That does make sense, but what exactly is the “model” in the equation?’

‘The “model” will change depending on what you’re trying to achieve. Ultimately the word “model” is replaced with an equation that you believe summarizes the pattern of data. Sometimes that can be as simple as a single value that summarizes scores, but other times the model might need to be more complicated. However, no matter how long the equation that describes your model, you can close your eyes and replace its hideous, stress-inducing form with the word “model”, which is much less scary. The main thing to understand is that whatever model we choose to represent the phenomenon of interest, there will be error in prediction. Obviously, we want to fit a model where this error is as small as it can possibly be.’

This equation did seem less scary than when I was in college. Perhaps it helped hearing it from a cat, or perhaps he was trying very hard to keep me engaged. I wondered why he cared about me understanding – surely there were easier ways to help me. If this really was the reality prism guy, surely he had other inventions, or he could just split reality and show me where Alice was. I asked him whether we could use a reality prism to find Alice, but his terse reply that I pay attention to the statistics suggested I had broached a subject that he didn’t want to talk about.

‘Statistical models usually, but not always, contain variables and parameters,’ he continued a little grumpily.

‘Alice told me about these the night before she disappeared. Variables are things that we have measured and that vary across entities in the sample. I can’t remember what parameters are.’

My response got him back on side and Milton adopted a kinder tone. ‘Not to worry, it’s a tricky concept. Parameters are (usually) constant values believed to represent some fundamental truth about the relations between variables in the model. Unlike variables, which are measured, parameters are estimated from the data. That probably does not mean a lot to you at the moment, but I will explain soon enough. Let’s look at the simplest possible model you might fit.’ Milton scratched another equation into the table. ‘In this model we are estimating the outcome from a single parameter, which I have denoted with the letter \( b \).’

\[
\text{outcome}_i = (b_0) + \text{error}_i
\]  

(4.3)
Why $b$? Why not $\rho$ for parameter?'

Statisticians try to confuse students by using different symbols for different parameters. For example, $\bar{X}$ tends to be used for the average, which is also called the mean; there’s something called a correlation for which they use the letter $r$ or Greek symbol rho, $\rho$; and in a model called the linear model, they use $b$ or the Greek letter beta, $\beta$. Statisticians like Greek letters, but I thought you’d enjoy the simplicity of using the letter $b$.'

'Won't that upset the statisticians?'

'Probably, but they are easily upset.'

'Why does it have a zero next to it?'

'I put it there to remind you that we are predicting the outcome from zero other variables, that is, just from a single parameter. Can you order me a latte with lactose-free milk please?'

The waiter looked at me as though I was something unpleasant on his shoe as I asked for lactose-free milk in a latte. I guess they weren’t used to dealing with cats’ lactose intolerances, or perhaps he just didn’t like cats eating at the table. Milton got his drink, though, and explained things while his lapping tongue sprayed me.

'The model I have just described is the simplest possible one: we summarize the population using a single parameter, a single value that summarizes the outcome variable. A lot of the time it is more interesting to see whether we can summarize an outcome variable by predicting from scores on another variable. We usually denote predictor variables with the letter $X$; therefore, our model will look like this.' Milton scratched on the table again. 'In this model, we are predicting the value of the outcome for a particular entity ($i$) from its score on the predictor variable ($X$). The predictor variable has a parameter ($b$) attached to it, which tells us something about the relationship between the predictor ($X$) and outcome.'

\[ \text{outcome},_i = (b_0 + b_1 X_i) + \text{error},_i \] (4.4)

'Why is the $b_0$ still there?'

'We are building up the model, so the $b_0$ is a parameter estimate that tells us the overall levels of the outcome if the predictor variable was not in the model, and the $b_1$ is a parameter estimate that tells us about the relationship between the predictor variable and the outcome.'

'So we could, like, predict relationship assessment scores from how long the couple had been together?'

\[ \text{relationship satisfaction},_i = (b_0 + b_1 \text{length},_i) + \text{error},_i \] (4.5)

Purrfectly true. We could see whether your assessment of your relationship depends upon the length of the relationship. If you did that you could replace the outcome ($Y$) and predictor ($X$) in the model with the names of these variables. We could even take that a step further and add another predictor variable to the model. Let’s say we measured how much effort the couple put into their relationship. Now we’re predicting the value of the outcome for a particular entity ($i$) from its score on two predictor variables ($X_1$ and $X_2$). Each predictor variable has a parameter ($b$) attached to it, which tells us something about the relationship between that
predictor and the outcome. Again, to help you to see how this works, we could replace some of the letters with the variable names.

\[ \text{outcome}_i = (b_0 + b_1 X_{1i} + b_2 X_{2i}) + \text{error}_i \]  
\[ \text{relationship satisfaction}_i = (b_0 + b_1 \text{length}_i + b_2 \text{effort}_i) + \text{error}_i \]

'Sweet, you’re summarizing the outcome with other variables that you have measured.'

'Purrccisely, and the beauty of it is that we could carry on expanding the model with more predictor variables. In each case they will have a parameter that quantifies the connection between that variable and the outcome variable.'

'I think I understand all of this, but how can a letter like \( b \) quantify the phenomenon we’re studying. It’s a letter, letters can’t quantify anything.'

'Ah, dear human, the letter is just a useful way to represent the parameter in general terms. It helps us to show the form of the model, but we need to replace each letter with a number.' His eyes bulged with excitement at the mention of numbers.

'How do we do that?'

'We use the sample data to estimate the value of the model parameters.'

Estimating seemed a bit wishy-washy, surely we needed something better than an estimate. I must have been screwing my face up because Milton almost read my thoughts.

‘Remember that we are interested in drawing conclusions about a population to which we did not have access. In other words, we want to know what our model might look like in the whole population, and to do that we want to know the values of the parameters of that model. The problem is that we cannot calculate the parameters in the population because we did not measure the population, we measured only a sample. What we can do instead is to use the sample data to estimate what the population parameters are likely to be. You will hear statisticians use the phrases “estimate the parameter” or “parameter estimates” a lot because it reflects the fact that when we calculate parameters based on sample data they are only estimates (i.e., a “best guess”) of what the true parameter value is in the population. Shall we look at some simple models?’

Every atom in my body wanted to say no. I had survived this long without equations, so why would I want to have more of them now? I also thought that there surely had to be an easier way to find out where Alice was and to talk to her. That’s all I needed, just to talk to her, and to find out what was going on. However, I had a feeling Milton had other plans, and for now at least, I perhaps needed to go along with him.

4.2 CENTRAL TENDENCY

As I suspected he would, Milton ploughed on. ‘After I walked into Catherine’s office, you discussed with her a score that reflected Alice’s satisfaction with your relationship: the Relationship Assessment Scale (RAS). I noticed when you showed me the information you had from Alice’s file that she completed this measure not just once, but many times.’

He was right. I had shown the Professor only Alice’s score in the first week of her counselling sessions.
4.2 Central tendency

“You have done well to listen to my explanations, but I am certain that you are wondering how this helps you to resolve your little problem. Let us use Alice’s ratings to try to see whether we believe the hypothesis about her absence reflecting an over-elaborate exit strategy from her relationship with you.”

Now we were getting somewhere. I leaned in to listen closely to what the cat had to say. I got out my diePad and we noted Alice’s RAS scores over a 10-week block of her counselling sessions:

32, 30, 28, 30, 30, 29, 31, 29, 31, 11

Milton looked surprised. “The score of 11 on the last week is really low.” I looked surprised back at him and he explained. “Most weeks Alice rated your relationship quite highly – close to the top of the scale – but then in the week before her disappearance, last week, it fell to close to the bottom of the scale. Remember that the lowest score on this scale is 7, so she rated your relationship about as low as it was possible to rate it, whereas the previous 9 weeks she rated it close to the top. A score that is very different from others is called an outlier, and if we have one or more of these it can affect the models that we fit. This is one reason why it is important to look at frequency distributions of your data, which Catherine explained to you, because they can show up unusual scores.”

“Should we ignore it if it is unusual?”

“Well, it is true that it can affect our model in bad ways – I can show you some examples in a moment – so for the sake of the model sometimes we try to reduce the impact of these outliers. However, often these scores are telling us something interesting. For example, isn’t it interesting
that the week before Alice disappears, her relationship rating is low? She was not happy with you
and then she vanishes. Read into that what you will.

I didn’t like the knowing look that Milton was giving me, but his argument made sense. The
pain in my face must have been obvious because his tone softened.

‘Let us not jump to conclusions before we have looked at some simple statistical models.’

He said it as though this were some treat that would cheer me up. It wasn’t.

‘The simplest model we can fit to these data is one that tries to summarize them in terms of a
single parameter. A popular choice would be a parameter that measures central tendency: a value
that indicates the central point of a distribution of scores. A “typical” score, you might say. If we
have data from the entire population we could compute this value directly, but if we have only a
sample then we can estimate the population value from the sample data.’

4.2.1 The mode

‘Purrhaps the simplest way to quantify the centre of a distribution is to use the score that occurs
most frequently, which is called the mode,’ Milton continued. ‘This value is easy to spot in a his-
togram because it will be the tallest bar, and in a tabulated frequency distribution it will be the
score with the largest frequency. To take your mind off of Alice, draw the frequency distribution
of her RAS scores and tell me which score has the biggest frequency.’

Figure 4.3  Alice’s RAS scores over 10 weeks

CHECK YOUR BRAIN: Create a frequency
distribution of Alice’s RAS scores.
I quickly drew up the table; it didn’t take my mind off of Alice. The value of 30 had a frequency of 3, which was greater than any other value. I gave this answer to Milton.

‘Clawrect, 30 is the mode of these scores. The mode is useful because it represents the most popular response, and unlike other measures it will always be a value that actually occurred in the data. It does have a downside though, which is that it can often take on several values. Had Alice given another score of 29 then both 29 and 30 would have had a frequency of 3; there would be two modes.’ Milton started playing with my diePad, twisting and turning its cogs with his paws until the screen displayed an image. ‘This image shows an example of a distribution with two modes (there are two bars that are the highest), which is said to be bimodal, and three modes (data sets with more than two modes are multimodal). Another problem is that if the frequencies of certain scores are very similar, then the mode can be influenced by only a small number of cases.’

I was curious about whether this mode had been affected by Alice’s unusual score: the low score of 11, which the cat had pointed out. He suggested that I calculate the mode without that score and when I did I found that the mode was still 30: that unusual score had not had an impact.

**Table 4.1**  Frequency distribution of Alice’s RAS scores over 10 weeks (excluding categories with zero frequencies)

<table>
<thead>
<tr>
<th>RAS score (X)</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>2</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td>29</td>
<td>2</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ N = \sum f = 10 \]

**Figure 4.4**  Examples of bimodal (left) and multimodal (right) distributions
4.2.2 The median

Milton sat on my diePad, obscuring the data from my view. I asked him what he was doing. He looked puzzled, as though I'd somehow asked him an obvious and idiotic question. ‘It’s warm on here,’ he said.

‘But I can’t see the screen.’

‘Your human needs are insignificant. The diePad is warm. I need warmth. You would think that I would not, what with having a fur coat, but there are many interesting things about cats that you do not realize until you are one. One of them is that we expend an incredible amount of energy keeping the fabric of reality together.’

Milton wasn’t making any sense. It was odd that he sometimes referred to me as human, as though he’d never been one, or as if he was proud to be a cat. I wondered whether his personality really hadn’t changed. He had an air of superiority about him, which annoyed me, I wondered whether the human Professor Grey had had that arrogance too or whether he inherited it from his cat. My thoughts were soon interrupted.

‘Pay attention,’ he said abruptly, ‘or you will never find Alice. There are other ways to summarize the central point of a distribution. For example, we can look for the middle score when the scores are arranged in ascending order; this value is known as the median. To calculate the median for Alice’s RAS scores, we first arrange these scores into ascending order: 11, 28, 29, 29, 30, 30, 30, 31, 31, 32. Next, we find the position of the middle score by counting the number of scores we have collected \( n \), adding 1 to this value, and then dividing by 2. With 10 scores, this gives us \( (n + 1)/2 = (10 + 1)/2 = 11/2 = 5.5 \). Then, we find the score that is positioned at the location we have just calculated. So, in this example we find the score five and a half positions along the list:

‘Don’t be crazy! There is a fifth score and a sixth score, but there isn’t one in between.’

‘Just checking that you have stopped daydreaming. I know how humans like to sleep – they are always dozing off, not like us cats. What a value of 5.5 means is that the median is halfway between the fifth and sixth scores. To get the median we add these two scores and divide by 2. In this example, the fifth and sixth scores in the ordered list are both 30. We add these together \( 30 + 30 = 60 \) and then divide this value by 2 \( 60/2 = 30 \). The median score that Alice gave was, therefore, 30.

‘Is the median affected by the weird score of 11?’

\[ \text{Median} = (30 + 30)/2 = 30 \]

\textbf{Figure 4.5} When the data contain an even number of scores, the median is the average of the middle two values when you order the data.
Let us see, shall we? If we remove the score of 11 we have nine scores left, so the median is even easier to calculate. We find out the middle score as we did before, but now we have nine scores, so this gives us \((n + 1)/2 = (9 + 1)/2 = 10/2 = 5\). The median is the score at position 5 in the ordered list, which is a value of 30. It seemed as though the median wasn’t affected at all.

4.2.3 The mean

Milton did some more scratching on the desk and said, “A measure of central tendency that you will have heard of is the mean, sometimes referred to as the arithmetic mean, but more commonly just called the average. To calculate the mean we add up all of the scores and then divide by the total number of scores we have.

\[
\text{Population mean} \quad \mu = \frac{\sum_{i=1}^{n} x_i}{N}
\]

\[
\text{Sample mean} \quad \bar{X} \text{ (or } M) = \frac{\sum_{i=1}^{n} x_i}{n}
\]

I noticed he had written two equations that looked the same apart from a couple of symbols. I asked him why.

“Very observant. Yes, they are the same, but some of the symbols are different, depending on whether you’re calculating the mean of an entire population of scores (the population mean) or of just a sample of scores (the sample mean). I mentioned earlier that statisticians like to use lots of symbols, and the population mean is represented by the Greek letter mu (\(\mu\)), which is a Greek lower-case \(m\), which makes it a fairly sensible choice to represent a mean. The sample mean is represented by an \(X\) with a bar on top of it, or sometimes the letter \(M\). Also notice that when we have an entire population of scores we typically use a capital \(N\) to represent the ‘Number of scores’, but when we have only a sample we use a lower-case \(n\) instead. You are right, though, both versions of the equation are mathematically the same: the top half of the expression on the
right-hand side of the equation means ‘add up all of the scores’, and the bottom half means divide this total by the number of scores you have added up.’

‘I don’t understand how you can talk about a population or sample of scores when we have data from a single person: Alice.’

‘Good question,’ he said with a look of satisfaction as though he should get some credit for my asking a sensible question. ‘You can define samples and populations in different ways according to the question you want to address. A population of scores doesn’t have to be scores from different entities, and neither does a sample. With the RAS scores we are interested in the population of scores representing Alice’s relationship satisfaction; we could define this as the ratings she would give every week of your entire relationship. In which case, the scores we have for the 10 weeks that she completed the RAS are a sample of her relationship satisfaction. Let’s calculate the mean for this sample. First, we add up all of the scores. Milton again etched the table with his claw. ‘We then divide by the number of scores, in this case 10.’

$$\sum_{i=1}^{n} x_i = 32 + 30 + 28 + 30 + 30 + 29 + 31 + 29 + 31 + 11 = 281$$ \hspace{1cm} (4.9)

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{281}{10} = 28.1$$ \hspace{1cm} (4.10)

I stared silently at his calculation, trying to work out what this all meant for my relationship with Alice. Milton charged on, oblivious to my thoughts.

‘Alice’s mean assessment of your relationship is 28.1. This value is not one that we observed in the actual data, which illustrates the fact that we are fitting a statistical model: we have reduced the data to a summary that does not perfectly represent the scores we observed. Think back to the equation I drew for a simple model.’ For the umpteenth time the table top was vandalized. ‘The model is the value of the mean, so we can replace b with the symbol for the mean, and the outcome we want to predict is the RAS score at a particular time, so we can replace “outcome” with “RAS” to remind us what we are predicting. We have RAS scores at different times, so the i symbolizes a particular point in time that Alice completed the RAS. What does this model mean?’

outcome; = (b) + error;  
RAS; = \bar{X} + error;  
RAS; = 28.1 + error; \hspace{1cm} (4.11)

‘Is it that we predict that Alice’s RAS score at a particular time will be the mean, 28.1, but that there is some error attached to that prediction?’

The cat looked pleased. ‘You can understand equations when you put your mind to it. You can see this error in prediction, because in the first week Alice rated your relationship as 32, but the model predicts 28.1, so it underestimates the actual score that Alice gave.’

‘I looked at Alice’s scores. It seemed to me that the model underestimated most of them: eight of the ten scores were greater than 28. How was that possible if the model is supposed to represent
the centre of the distribution? I asked Milton, who said that it was because of the weird score. He called it an outlier – the rating of 11 that Alice gave in the last week. He then said that to see the impact that this score had had, I should compute the mean without including that score.

**CHECK YOUR BRAIN:** Compute the mean but excluding the score of 11.

Milton was still using my diePad as a heat pad, so I grabbed a napkin and pen and tried to add the numbers. The total was 270, and there were nine scores, so I needed to divide 270 by 9. That's the sort of sum my diePad could tackle in nanoseconds if it didn't have a big ginger cat on it. After consulting The Head I had an answer. ‘It’s 30,’ I said. Milton said nothing. He was curled into a ball on my diePad, gently purring. I poked him with my finger. He raised his outer eyelid slightly and made a token attempt to biff my hand away with his paw. I poked him harder. Milton yawned and sat up looking very disgruntled. ‘What is it? I’m trying to sleep.’

‘The mean without the outlier is 30.’

Milton got up from my diePad, revealing an image on the screen that he had somehow created while sleeping on it. ‘Look,’ he said ‘you can see the outlier clearly in this histogram. It is a score that is a long way from the others, and the effect it has is to shift the mean towards it. This is because the mean is the balance point of the data.’ I noticed that he had also written me a message.

Dear Human,

The mean is the balancing point of a distribution, which means that adding (or taking away) scores will change the mean unless the score you add (or remove) equals the value of the mean itself. Alice's RAS scores (excluding the outlier) are shown on the left of Figure 4.7; the mean of 30 is the score that keeps these values in balance. If you add a score to the data that is equal to the mean then the mean will not change (Figure 4.7, top row). If you add a score that is greater than the mean then it tips the scores to the right, and the mean shifts that way too (i.e., increases) to compensate and restore balance (Figure 4.7, middle row). Conversely, if you add a score that is smaller than the mean then it tips the scores to the left, and the mean also shifts that way (i.e., decreases) to compensate and restore balance (Figure 4.7, bottom row). A similar thing happens if you remove scores from the data: removing a score less than the mean will increase the mean, and removing a score greater than the mean will reduce it.

4.2 CENTRAL TENDENCY
In a similar vein, if you change the data by a fixed value, then the mean changes by that value too. For example, if you add 10 to every score then the mean of the new scores will be the mean of the old scores plus 10, similarly it would be 10 less if you subtract 10 from every score; if you multiple all of the scores by a value (e.g., 10) then the new mean will be the old mean times that value, and if you divide scores by a constant (e.g., 10) then the new mean will be the old mean divided by the constant (Figure 4.8).

Best fishes,
Milton

---

**Figure 4.7** The mean is the balancing point of a distribution

**Figure 4.8** The effect on the mean of changing scores by a constant value

*Milton's Meowsings 4.1* Things that affect the mean
Milton asked me what I noticed about the mean if we exclude the outlier. I looked at the image he had created. It seemed to me that without the outlier, the mean split the data exactly in two: the distribution was symmetrical about the mean. I said this to Milton.

‘Purr cisely. The mean divides the data in two, which makes it a reasonable summary of the data as a whole, but if we include the outlier, the mean decreases. What do you notice about this new mean in relation to the scores?’

‘It’s like I said earlier: the only scores that aren’t above the mean are 11 and 28: the other eight scores are above the mean.’

‘Very good. The outlier has made the mean less representative of the data.’

‘Why was the mean affected by the outlier, but the median and mode weren’t?’

‘Another good question,’ he replied in a slightly patronizing tone. ‘It is because the mean, median and mode measure the “typical” score in different ways. The mean defines it in terms of distance from the centre, so a score a long way from the centre can throw the mean off. The median measures it as the score at the centre, and the mode measures it as the most frequent score. The mean aims to split the data into two equal halves, but the mean does not try to do this, which is why you can end up in a situation in which two scores fall below the mean and eight above it – that would not happen with the median because it is based on splitting the data into equal halves. The median balances the data so that half of the scores are above it and half below: the distance of each score from the centre is not considered; only whether it is above or below the centre.’

‘So the median is better than the mean?’

‘Not necessarily. The mean has many useful features. It uses every score in the data, and so is representative. But most important is that it tends to be quite stable across samples; that is, if you took several samples and measured the same thing in them, you would find that the mean ought to be relatively similar in the different samples – this is less true of the median.’

**Figure 4.9** Outliers can affect the mean

4.2 CENTRAL TENDENCY
Dear Human,

When considering which measure of central tendency to use, you need to think about the type of data you have (Section 2.3.3). If you have nominal data then you can only use the mode (because the median assumes order in the data, and the mean is based on the distances between scores which doesn’t make sense for nominal data). With ordinal data the median is appropriate (because it is based on the scores being ordered) but the mean is not (because it is based on distances between scores and it assumes that the distance between scores is the same at all points along the scale, which is true of interval data, but not ordinal data). With interval or ratio data the mean is appropriate and so is the median (because on an interval scale scores are still ordered).

The other thing to think about is extreme scores or skew in the distribution. The median is less affected by extreme scores at either end of the distribution than the mean. Remember that with our example data, the extreme score (11) did not affect the median but it dragged the mean down. Skewed distributions (Section 3.2) represent a situation in which there are a minority of scores that are a fair bit smaller (negative skew) or larger (positive skew) than the rest. As such, skewed distributions will also affect the mean (negative skew reduces the mean, positive skew inflates it).

Best fishes,
Milton
‘That was strange,’ said Milton.

‘It’s been happening a lot recently. People collapse and don’t remember who they are. I was reading about it yesterday: they say there are no known cases of it happening to Clocktorians, only Chippers. The WGA is suggesting a Clocktorian revolution, you know, Clocktorians using some sort of drug to attack Chippers. I don’t buy it.’

‘Why not?’

‘It doesn’t make any sense. We don’t want to convert everyone else to be a Clocktorian, we just don’t want to have ID chips. We don’t care if other people have them. It’s convenient for the WGA too because they are using it to insist that ID chips become compulsory.’

‘I think their public relations office will have fun trying to convince people to implant themselves with something that apparently makes you more likely to forget who you are. Anyway, shall we get back to Alice’s data?’
I shrugged. ‘If you want to find out what has happened to her then we must – you decide.’

Milton was trying to act as though he didn’t care what I did, but he’d just invested a lot of time telling me about models, and I got the impression that he was relying on my need to find Alice to keep me hooked into whatever game he was playing. He was right: I am nothing without Alice, and I willingly continued.

4.3.1 The fit of the mean

‘Good decision,’ the cat said knowingly. ‘You understand now how the mean is a model, and earlier I told you that it is important to know whether a model is a good or bad fit.’

I nodded. ‘We can assess how well the mean, or any other model, “fits” the data by looking at the errors in prediction. Remember, when we use the mean as a model of the typical score, we write the model like this.’ Milton pointed to one of his equations on the table. ‘We can rearrange this equation to calculate the error.’ He started scratching into the wood again.

\[
\text{outcome}_i = (\text{model}) + \text{error}_i
\]

\[
\text{error}_i = \text{outcome}_i - \text{model}
\]

\[
\text{error}_i = \text{RAS}_i - \mu
\]

‘The error for a given observation is the value observed minus what the model predicted. If we are predicting particular RAS scores from the average RAS score, then it is the difference between the particular RAS score and the mean. Imagine we are interested only in the fit of the mean to the scores we have, so we use \( \mu \) (remember it is just a Greek \( m \)) to represent the mean.’

‘That’s what I said earlier, that the model underestimated Alice’s first RAS score because she scored 32 but the mean was only 28.1.’

‘Exactly. We can write your observation as an equation, in which we say we are predicting Alice’s score in week 1, from the mean (our model). Look,’ he said as he destroyed some more table, ‘the error will be 32–28.1, or 3.9.’

\[
\text{error}_{\text{week}1} = \text{outcome}_{\text{week}1} - \text{model}
\]

\[
\text{error}_{\text{week}1} = \text{RAS}_{\text{week}1} - \mu
\]

\[
\text{error}_{\text{week}1} = 32 - 28.1
\]

‘Can I let you into a secret about statisticians?’

‘They are all evil?’

‘Oh, so it is not a secret?’ Milton joked. ‘Legend has it that back in ancient Greece the early pioneers of modern mathematics solved the problem of time travel, but they shared it only with other mathematicians. Mathematicians have been secretly enjoying the privilege of time travel ever since.’
‘How do they do it?’ I asked, keen to discover the secret so that I could go back to before the revolution and watch some of my favourite bands.

‘I do not know. Rumour has it that it involves thinking about some really hard equations.’ I wondered whether Milton was lying and did know, but then wouldn’t he have gone back in time to prevent his accident? ‘In any case,’ he continued, sensing my distracted mind, ‘the great statistical minds used to gather across time to meet annually at a Victorian café in Soho, London, called Gossett’s Tea. They were all there – Fisher, Pearson, Neyman, Gauss, Bayes, and Florence Nightingale. The festivities would kick off with a bun fight between Bayes, Fisher, Neyman and Pearson, after which they would discuss life. They could not understand why all of the other people in the world had wives and husbands, and friends, and lived fulfilled lives. They were also confused about why other people didn’t wear glasses and have white beards – especially Florence. They concluded that of all the earthly experiences – friendship, love, a smooth chin – none could compare with the knowledge of statistics. They vowed to keep this experience for themselves, and drew up a pact to obfuscate statistics. They decided that the best way to achieve this was to come up with multiple names for the same thing.’

‘Weirds. Alice told me that was because of a guy called Confusius who invented a confusion machine.’

‘Don’t be ridiculous, that is completely implausible – it is because of time-travelling statisticians. Now pay attention. Error is one of these words, because the error in prediction from a model is sometimes known as error, and sometimes as deviance or deviation, and other times as residual. These terms have subtly different uses, but mean essentially the same thing. The deviance is the value of the outcome minus the value predicted from the model, and when the model is the mean, it is the observed score minus the mean.’

\[
\text{error}_i = \text{outcome}_i - \text{model}_i
\]

\[
\text{deviance}_i = x_i - \mu
\]  

(4.14)

Milton continued. ‘If we know that this is the error for a particular score, how might we work out how much error there is overall?’

I thought about this question; it seemed that you could work out the error attached to each score and then add them up, but I was scared of making a fool of myself, so I said nothing.

‘You are very quiet, I assume that you don’t know the answer. I will tell you: if we want to know the total error or total deviance then we could add up the deviances for each data point.’ I kicked myself for not having the balls to give Milton an answer, but I was also pleased that I had been right. ‘In equation form, we can scratch this into the table as follows.’

\[
\text{total error} = \sum_{i=1}^{n} (\text{outcome}_i - \text{model}_i)
\]

\[
\text{total deviance} = \sum_{i=1}^{n} (x_i - \mu)
\]  

(4.15)

‘The sigma symbol (\(\sum\)) simply means –’

4.3 THE ‘FIT’ OF THE MEAN: VARIANCE
‘– add up what comes after, and what comes after is the deviances, so it means add up all of the deviances,’ I interrupted.

Milton looked impressed. He had a much more expressive face than the average cat, which I guessed was because it was being controlled by a human mind. ‘Let us see what happens when we calculate the deviances for Alice’s RAS scores. We will ignore her final score of 11 for now.’ Milton sat on my diePad again and wriggled around a bit, before stepping off to reveal an image . I didn’t even want to think about how he did that. ‘This diagram shows Alice’s RAS scores over the nine weeks, and also her mean RAS score that we calculated earlier on. The line representing the mean is our “model”, and the circles are Alice’s observed scores. The vertical lines that connect each observed value to the mean value are the errors or deviances of the model for each observed score. The first week, Alice’s score was 32, but the model predicts a score of 30 so the error for week 1 is 2. This error is a positive number, and represents the fact that our model underestimates Alice’s actual score. For week 2, the model predicts 30 (again) and the observed score is 30 which means that the error is 0. For week 3, Alice scored 28 but the model predicts a score of 30, so the error is –2; the negative number tells us that our model overestimates Alice’s actual score. You can see the errors for the other weeks on the diagram.’

‘Yeah, you can see how the model sits in the middle of the data, so sometimes it overestimates and sometimes it underestimates the actual scores.’

‘Yes, so what do you think might happen when we add up these errors, or deviances, as you suggest?’

‘It tells us the total error?’

Figure 4.10  Graph showing the difference between Alice’s observed RAS scores each week, and the mean RAS score
The result was surprising: the total was 0, as though there was no error at all, but that couldn’t be true because I could see in the diagram that the mean was different from the observed scores; there was error. Milton explained.

‘Odd, is it not? It happens because the mean is at the centre of the distribution, and so some of the deviances are positive (scores greater than the mean) and some are negative (scores smaller than the mean). Consequently, when we add the scores up, the total is zero.’

‘Can’t we just ignore whether the error is positive or negative, then?’ I asked.

Another impressed look from Milton. ‘A sensible idea, and there is nothing wrong with doing that, but generally we square the deviances, which has a similar effect (because a negative number multiplied by another negative number becomes positive). We can add these squared deviances up to get the sum of squared errors, SS (often just called the sum of squares); unless your scores are all exactly the same, the resulting value will be bigger than zero, indicating that there is some deviation from the model (in this case the mean). We can express this like so; the first equation shows the general principle that we can apply to all models, and the second one shows the equation specifically for the mean.’

\[ \text{sum of squared errors (SS)} = \sum_{i=1}^{n} (\text{outcome}_i - \text{model}_i)^2 \]

‘So, as before, the sigma symbol means “add up all of the things that follow”, but what follows now is the squared errors.’

‘Yes. Can you do this for our nine scores? I will draw a table for you.’ Milton again sat on my diePad, wriggled around for a few seconds, then got up and stepped away, revealing a table.

The first column is Alice’s scores. The second column reminds you of the model, in this case the mean. The third column shows the deviance or error for each score. Note that when you add these values the sum is zero. The final column shows the deviance, or error, squared. Note that
Fitting models (central tendency)

Table 4.2  Table showing the deviations of each score from the mean

<table>
<thead>
<tr>
<th>RAS score ($x_i$)</th>
<th>Mean ($\mu$)</th>
<th>Deviance ($x_i - \mu$)</th>
<th>Deviance squared ($x_i - \mu)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>30</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>29</td>
<td>30</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>30</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>30</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>30</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

\[
\sum_{i=1}^{n} x_i - \mu = 0 \\
\sum_{i=1}^{n} (x_i - \mu)^2 = 12
\]

unlike the raw deviances, none of the values are negative. Therefore, when we add these up we get a positive number, 12. What do you think would happen to this total if we found another 10 scores and added them to the table?

‘Would the sum of squared errors get bigger because we’re adding up more squared deviances?’ I felt I was getting the hang of this now.

‘Supurrb. It would get bigger, so although we can use the sum of squares as an indicator of the total deviance from the model, its size will depend on how many scores we have in the data. Sometimes this won’t matter, but it is a nuisance if we want to compare the total error across samples of different sizes. An easy solution is to divide by the number of scores ($N$), to give us the “average” squared error, known as the mean squared error. When the model is the mean, the mean squared error has a special name, the variance. We can take the previous equation and add to it by dividing by the number of scores $N$. The symbol for variance in the population is $\sigma^2$. Our total squared error was 12, and that was based on nine scores, so what is the mean squared error, or variance?

\[
\text{mean squared error} = \frac{SS}{N} = \frac{\sum_{i=1}^{n} (\text{outcome}_i - \text{model}_i)^2}{N} \tag{4.18}
\]

\[
\text{variance} (\sigma^2) = \frac{SS}{N} = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}
\]

Using my diePad I divided 12 by 9. ‘It will be 1.33,’ I said proudly.

‘Yes. Well done. There is one problem with the variance as a measure.’
‘Frippin’ hell, Milton, is there ever not a problem? First the total error is no good, then the total squared error is no good, now the average squared error is no good. Is anything ever going to be good?’

Purrhaps “problem” is too strong a word. The mean squared error is fine, but because we squared each error in the calculation it gives us a measure in units squared. For Alice’s RAS scores, we would say that the average error of the mean was 1.33 RAS units squared. There is nothing wrong with saying that, but it is sometimes useful to convert back to the original scale of measurement, so we can talk about the error with reference to the RAS scale. To do this, we would take the square root of the variance (which converts the average error back to the original units of measurement). This measure is known as the standard deviation and the symbol for it in the population is $\sigma$.

$$\sigma = \sqrt{\text{variance}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$ (4.19)

$$= \sqrt{1.33}$$

$$= 1.15$$

4.3.2 Estimating the fit of the mean from a sample

This was making some kind of sense to me, and that realization was exciting. I wished Alice was here to see how well I was doing. I felt sure that if she was having doubts about us she’d forget them if she could see me taking in this stuff. Milton had explained that we fit models to data to summarize the general pattern. Like, what’s the pattern in Alice’s relationship scores? The most basic way to do that is to just look at the middle of the distribution: what is the most common score (the mode), the middle score (the median) or the balancing point of the data (the mean). The mean is a model because it doesn’t have to be a value actually observed in the data. It also made sense that if you have a summary then you want to know how good that summary is. With the mean you can do that by looking at the average error between the data and the mean.

I was feeling pretty pleased with myself, but Milton was about to throw me a curve ball of monumental proportions. He rose up, arched his back in an enormous stretch as if preparing himself for something big. Then he dropped the ‘degrees of freedom’ bomb. More powerful than any nuclear explosion, this was a bomb designed to ruthlessly and precisely destroy all brain matter with which it came into contact. My brain was only centimetres from the epicentre. Batticks!

‘Of course,’ Milton began, ‘I have simplified things a smidge.’

‘What do you mean by “a smidge”?’

‘I mean, a lot,’ replied Milton.

‘A lot?’

‘An enormous amount, really. Everything I have said about testing how well the mean fits the model is true if you have the population data (i.e., all of the data in which you are interested), but in most situations we have only a sample, and we estimate the population mean and the “fit” from that sample.’

‘So, what’s the problem?’

‘The problem is that if we use the equations that I just mentioned we will underestimate the variance and standard deviation of the population.’
‘Why?’
Milton jumped onto my diePad again, his paws moving in a whirlwind around the screen. It was amazing that a cat had such dexterity.

‘Look at this drawing. It shows the histogram of a population of scores on the RAS. You remember what a histogram is?’

‘Yes, you have each score listed along the bottom, and the height of the bar tells you how often a score was observed in the data.’

‘Excrathlety, so you can see in the population that most scores cluster around the centre, and scores at the extremes have low bars indicating that these scores do not occur very often in the population. Also, notice that the population uses up the whole of the scale: scores range from the minimum value of 7 on the RAS to the maximum value of 35. If we were to take a sample of scores from this population, it is much more likely that we would select a score from the middle (because there are lots of them in the population) and much less likely that we would sample a score from the extremes (because there are fewer of them). That is not to say that we will not get scores from the extremes in the sample, but it is less likely than getting one from the centre. The resulting sample will, on average, contain lots of scores from the middle of the population and relatively fewer at the extremes. The result is that the sample distribution is typically narrower than the population, as you can see in the drawing. What that means is that in the sample scores are closer to the middle than they are in the population. Now what does the variance measure?’

‘It measures the fit of the mean, or the average error between the mean and the observed scores.’

‘Yes, and we define the error as the distance between the mean and the score. So, the variance measures the average distance of scores from the mean. In a narrow distribution the scores are closer to the centre than in a wide distribution.’

‘Which means that the variance, the average distance from the middle, will be smaller too in a narrow distribution. So because the sample distribution will be narrower than the population, its variance will be smaller than the population.’

‘Brilliant!’ Milton patted my arm with his paw. ‘I will make a statistician of you yet! The variance in the sample will underestimate the variance in the population because the sample is likely to be narrower than the population from which it comes.’

‘If the sample underestimates the population variance, then doesn’t that mean that it’s a bit lame?’

‘It means that it is a biased estimator.’

‘Is that statisticians’ code for “a bit lame”?’

‘Purrhaps. A biased estimator is a statistic taken from a random sample that does not equal its corresponding population parameter, whereas an unbiased estimator is a statistic from a sample that does equal the corresponding population parameter.’

‘Surely a biased estimate is a bad thing. I mean, if my pocket watch told me the wrong time I’d buy a new one.’

‘Would you? What about if your watch always displayed the time exactly 5 minutes behind the actual time? It underestimated the time by exactly 5 minutes. Always.’

‘Well, I suppose if I knew the watch was always 5 minutes slow I could add 5 minutes to the time it displays.’

‘— and save some money. Exactly, and samples are like this slow watch: they provide a biased estimate of the population variance, but it is always biased in exactly the same way. This means
Figure 4.11 Why do samples underestimate the population variance?
that we can correct it to make it unbiased. Remember that the variance is the average deviation from the mean, and to get this average we divided by the number of scores ($N$)? Well, if we are using the sample to estimate the population variance then we divide by $N - 1$, rather than by $N$.

‘What? That’s the only difference?’

‘Yes, a small but important difference.’

‘We just replace $N$ with $N - 1$ in that equation? Seriously? That’s it?’

$$\text{mean squared error} = \frac{SS}{n - 1} = \frac{\sum_{i=1}^{n} (\text{outcome}_i - \text{model}_i)^2}{n - 1}$$

$$\text{variance} (s^2) = \frac{SS}{n - 1} = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n - 1} \quad (4.20)$$

‘Basically, yes. However, because we are now working with a sample, we would also replace the $\mu$, which is the symbol for the population mean, with $M$ or $\bar{X}$, the symbol for the sample mean, and we use $n$ instead of a capital $N$ to show that we are using the number of scores in the sample, rather than in the population. The variance itself we give the symbol $s^2$ to show that it is based on a sample, rather than using $\sigma^2$. For Alice’s RAS scores, if we assume that the nine scores we have are a sample of her scores for the population of all the weeks you have been together, then we could estimate the variance of this population by dividing the sum of squares by 8, rather than 9. This gives us a variance of 1.50, which is slightly higher than before, which illustrates that this correction increases the estimate to counteract the fact that samples typically underestimate the population variance.’

‘And the standard deviation of the population is estimated from the sample by taking the square root of this variance estimate?’

‘Purrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrr
time. It is the same here, we do not need to know the maths, we need only to know that using \( n - 1 \) works, and I can show you that it does.’

It was a boast that Milton followed up on.

‘Imagine we have a population of just 5 people, and they gave us RAS scores of 26, 23, 17, 22, and 20.’ Milton again started playing frantically with my diePad until he had produced a really complex diagram. The mean of this population is 21.6, and the variance is about 9. Now imagine that we took 10 random samples of 3 scores from this population; because there are only 5 scores in the population, some of these samples will be the same and others different. In each sample, I will compute the variance using both \( N \) (in each case this will be 3) and \( n - 1 \) (in each case 2). When calculating the variance using \( n \), the samples give estimates below the actual population value of 9 most of the time. If we use \( n - 1 \) then the estimates are, in general, closer to the actual value of 9. The crucial thing is what happens if we average across the estimates for the 10 samples. The average estimate when we calculate the variance using \( n \) is 6.18. In other words, on average, we have underestimated the true value of the population. However, when we calculate the variance using \( n - 1 \) the average estimate of the variance is 9.26, which is closer to the true value. This example shows that on average estimating the population variance from a sample by dividing by \( n - 1 \) gives us an unbiased estimate of the true value.’

I looked at Milton’s drawing. It seemed to me that for some samples using \( n - 1 \) made the variance estimate worse. ‘Look at the last sample,’ I said. ‘The variance estimate is 20.33 when you use \( n - 1 \), but only 13.56 when you use \( N \), so it’s closer to the true value of 9.04 when you use \( N \).’

The cat stared at me. ‘The point is that on average the estimate is better, not necessarily for a particular sample. This is important because in the real world we would not know what the population variance is (and if we did we would not need to estimate it), so we have no way of knowing how close the estimate in our sample is to the true value, but by using \( n - 1 \) we know that on average it will be better than using \( N \).’

That was a brain frip, but I had to admit that using \( n - 1 \) did seem to work, even though I didn’t understand why.

‘\( n - 1 \) is known as the degrees of freedom for the sample, which is a tricky concept, so concentrate. In statistical terms the degrees of freedom are the number of scores that are free to vary.’ Once more Milton attacked my diePad, feverishly creating another visual masterpiece. ‘Imagine we have a population. Of course, we do not actually have the data from the population, so we take a sample of four scores. These scores can be any value – they are free to vary. If we are interested only in the variance of those four scores, then we can calculate it using the sample size \( n \). However, what if we want to use the sample to estimate the variance in the population? To do this we need to know what the value of the mean is in the population, but we do not know it so we assume that it is the same value as in the sample. This assumption is fairly reasonable because, as I mentioned...’
Figure 4.12 Why we use $n - 1$ when estimating the population variance
before, the mean is relatively stable in different samples from the same population. In assuming that the sample and population means are the same, we are fixing a parameter. Imagine that the mean in the sample was 10, and therefore we fix the population mean to be 10 also. How many scores in the sample of four could be randomly selected from the population?'

'Surely it's four because they were all randomly selected?'

'Really? Let us see. Say that the first three scores we sample from the population are 8, 11 and 12. What is the mean in the sample?'

'It's \((8 + 11 + 12)/3\), so 31/3.'

'Yes, it is 10.33. But we have fixed the mean to be 10, so the final score we sample cannot be any score from the population – it has to be a score that brings the sample mean to that fixed value of 10. If the final score that we sample is a 6, then we end up with a mean of \((8 + 11 + 12 + 6)/4 = 9.25\), which is no good because we have fixed the mean to be 10, not 9.25. The only value that will bring the sample mean to the fixed value of 10 is 9: \((8 + 11 + 12 + 9)/4 = 10\). So, in a sample of four, the first three scores can be truly randomly sampled from the population, they are free to vary, but the last score is not: it has to be a score that brings the mean to 10. The same is true if we sampled five people from this distribution: the first four scores can take on any value from the population, but the final score has to be a value that makes the mean, in this case, 10. In general, if we hold one

4.3 THE ‘FIT’ OF THE MEAN: VARIANCE
Fitting models (central tendency)

Parameter constant then the degrees of freedom must be one less than the number of scores used to calculate that parameter. This fact explains why when we use a sample to estimate the mean squared error (or indeed the standard deviation) of a population, we divide the sums of squares by \( n - 1 \) rather than \( N \): because only \( n - 1 \) scores are truly free to vary, there are \( n - 1 \) degrees of freedom.

‘That made no sense whatsoever.’

‘Worry not: nobody understands degrees of freedom.’

4.3.3 Outliers and variance

I was starting to feel stupid again: that whole degrees of freedom nonsense had melted my brain and I felt like I’d got too cocky in thinking I understood everything. Perhaps Milton was putting me back in my place, or perhaps this stuff really was important. I needed to get the conversation back into territory that I understood. I remembered that we’d ignored Alice’s extreme score when we calculated the variance. I asked Milton why.

‘I did that to keep the maths simple, but you remind me of a very good point. As a bit of practice, compute the variance with that score included and see how it affects the variance.’

‘Should I use \( N \) or \( n - 1 \)’?

‘As I said, the vast majority of times we use the sample to estimate the value in the population, so use \( n - 1 \).’

**CHECK YOUR BRAIN:**

Estimate the population variance of Alice’s RAS scores from the sample including the score of 11.

So began the monumental task of working out the variance. Milton jumped off of the table and went on a little tour of Oceam’s café. He wandered under tables looking for unsuspecting diners to lavish affection upon him. Some looked horrified at the presence of a cat, others simply accepted that he was there and stroked and fusssed him. I felt sure that they would have been less enthusiastic if they had known they were tickling a ‘dead’ physics professor. He returned some time later, by which time I had finished the calculation.

‘The variance with the outlier is 37.43,’ I said.

‘Good, very good. You took a while, but nevertheless correct – well done. Can you remember what the variance was before we added the outlier?’

‘It was 1.5, wasn’t it?’

‘Clawrect.’

‘Wow, that’s an insane difference.’

‘Indeed it is. This goes to show you that, like the mean itself, extreme scores influence the variance. In fact the variance is more biased than the mean – the extreme score will not only have a large deviance from the mean, but that large deviance then gets squared when the variance is computed, which makes it disproportionately big.’
When I’d woken up yesterday morning I was looking forward to a nice brunch with Alice, but the weekend was broken: I woke up with a broken heart, I broke into someone’s office, then this morning I’d sneaked into a high-security research centre and stolen a cat who was now teaching me statistics inbetween licking a cup of lactose-free latte. I’d had my spooks quota for one week. I was feeling out of my depth on every level. We’d been here ages, I was tired, I needed to go home, check in with Nick and actually do something that was going to tell me what the frip was happening. I told Milton I was going home.

‘That’s very kind of you, I would love to come home with you, rest under your warm radiator and eat all of your food. Thank you for offering.’

‘But I …’

‘Before we go, just a few final things,’ he said as he padded across my lap to the other side of the bench.

### 4.4.1 The standard deviation as an indicator of dispersion

Before I could protest, Milton had launched into another monologue. I wondered if his plan was to distract me for so long that I’d forget about Alice.

‘I have described the sum of squares, variance and standard deviation to you. I talked about them measuring the “fit” of the mean to the data. Another way to think of “fit” is as the dispersion or spread of data around the mean. I have hinted at this already. Remember that the sum of squares, variance and standard deviation all represent the same thing, but one is a total, one is an average, and one is the average converted back to the original units of measurement. A small standard deviation (relative to the value of the mean) indicates that data points are close to the mean. A large standard deviation (relative to the mean) indicates that the data points are distant from the mean. A standard deviation of 0 would mean that all of the scores were the same.

‘I will draw you two pictures showing some RAS scores for two people over a five-week period. Both people had the same RAS score on average (\(M = 20.2\)), but their scores are very different. The first person had a standard deviation of 1.6 (very small compared to the mean). RAS scores across the five weeks were consistently close to the mean for this person. Put another way, the scores are not spread too widely around the mean. Contrast this with the second person whose scores had a standard deviation of 12 (fairly high compared to the mean). The second person’s scores are spread more widely around the mean than the first: some weeks he rated his relationship very highly but other weeks he was very unsatisfied.’

‘I can see that the mean “fits” the scores better for the first person than the second,’ I said trying to sound enthusiastic.

‘Yes, and this is reflected in the low standard deviation. The variance and standard deviation also tell us about the shape of the distribution of scores. If the mean “fits” the data well then most of the scores will cluster close to the mean and the resulting standard deviation will be small relative to the mean. When the mean is a worse “fit” of the data, the scores will cluster more widely around the mean and the standard deviation will be larger.'
I'll draw you another picture. These two distributions of RAS scores have the same mean ($M = 21$) but different standard deviations. The one on the left has a large standard deviation relative to the mean ($SD = 6$), and this results in a flatter, more spread-out, distribution. The one on the right has a standard deviation half the size ($SD = 3$). Notice that this gives the distribution a narrower, pointier shape: scores close to the mean are very frequent but scores further from the mean rapidly become increasingly infrequent. The key message is that as the standard deviation gets larger, the distribution gets fatter.
4.4.2 The range and interquartile range

Milton paused to scratch his ear, and I thought our lesson was over. Instead, and completely oblivious to the desire that his relentless statistics lecture was giving me to pray for sweet death, he carried on.

‘You know, there are simpler ways to look at dispersion. A very simple way is to take the largest score and subtract from it the smallest score. This is known as the **range** of scores. If we order Alice’s RAS scores we get 11, 28, 29, 30, 30, 30, 31, 31, 32. The highest score is 32 and the lowest is 11; therefore, the range is 32 – 11 = 21. What happens if we compute the range, but excluding Alice’s low score of 11?

**Check Your Brain:** Compute the range but excluding the score of 11.

The highest score was still 32, and the lowest was now 28. It was easy enough to work out that the range would be 32 – 28, or 4.

‘Without the extreme score the range drops dramatically from 21 to 4 – less than a quarter the size. This is a problem with the range: outliers can radically influence it. One solution is to calculate the range excluding values at the extremes of the distribution. It is common to cut off the top and bottom 25% of scores and calculate the range of the middle 50% of scores – known as the **interquartile range**. Let us do this with Alice’s RAS scores. First we need to calculate quartiles, which are the three values that split the sorted data into four equal parts. This is not as hard as it first seems if you understood the median, because it involves computing three medians.’

Milton popped out a claw and started carving a diagram into the table again. ‘First calculate the median of the whole data set, which is also called the second quartile, because it splits the data into two equal parts. We already know that the median for these data is 30. Use this value to split the data into two equal halves. For the RAS scores, this means the lower half of scores contains 11, 28, 29, 29 and 30; the higher half of scores contains 30, 30, 31, 31, and 32. We now need to compute the median of these two halves. For the lower scores, the median will be in position \((n + 1)/2 = (5 + 1)/2 = 6/2 = 3\) in the list, which is the number 29. This is called the lower quartile \((Q_1)\) or first quartile. We do the same for the upper half of scores: the median of this half is again in position \((n + 1)/2 = (5 + 1)/2 = 3\) in the list, which is the score of 31. This is known as the upper quartile \((Q_3)\) or first quartile. The interquartile range is the lower quartile subtracted from the upper one.’

\[
\text{IQR} = Q_3 - Q_1 \tag{4.22}
\]

‘So for the RAS scores it would be 31 – 29, which is 2.’

‘Supurrb. However, there is a complication.’

‘I knew there would be ...’

‘You see, as a rule of thumb the median value is not included in the two halves when they are split. If we had removed the median then our lower half of scores should have been 11, 28, 29, 29 (and 30 is ignored) and the upper half likewise ignores the two 30s leaving us with 31, 31, and 32.’

‘But then the two halves aren’t equal.’
Indeed. That is why I included the median, because I could divide the data more sensibly. It is acceptable to include the median if it makes things easier; you might say it is two different ways to skin a human.' He paused expecting me to laugh, which I didn't. 'How is the IQR affected if we remove the extreme score?'

I thought about this question for a moment. I had worked out before that with 11 removed the median was still 30. Splitting the data in half at 30, I'd get 28, 29, 30 and 30, 31, 31, 32, which is weirds because one half is bigger than the other. I could not think what to do and, not for the first time, I felt a wave of inadequacy come over me.

Sensing my despair Milton prompted me: 'Remember that I said that usually you would exclude the median from the halves, perhaps that would help.'

It did help, as Milton well knew. Without including the value of 30 I had a lower half of 28, 29, 29 and an upper half of 31, 31, and 32. The middle scores were easy enough to find: they would be in position \((n + 1)/2 = (3 + 1)/2 = 4/2 = 2\). In other words, the upper and lower quartiles were 31 and 29, the same as before, and the interquartile range would still be 2.

'The interquartile range doesn't change if you exclude the extreme score,' I proclaimed.

'Purrfectly true. This is an advantage of the interquartile range: it is not affected by extreme scores at either end of the distribution because it focuses on the middle 50% of scores. Of course, the downside is that you ignore half of your data. It is worth pointing out here that quartiles are special cases of things called quantiles, which are values that split a data set into equal portions. Quartiles are quantiles that split the data into four equal parts, but you can have other quantiles such as percentiles (points that split the data into 100 equal parts), noniles (points that split the data into nine equal parts) and so on. You look awful.'

I felt awful, too. I explained to Milton that, much as I appreciated his time, all of this information had got me nowhere. I felt it was a waste of a day, a day that I could have spent finding out...
whether Nick had got any leads from our fans, or any clues on memoryBank, a day where maybe I could have called the WGA, or done something that would get me closer to Alice.

The cat looked twitchy at the mention of the WGA, but he positioned himself on my lap and looked up at me with his big eyes. If you forgot who he was, he was pretty cute. His tone softened, and he spoke gently. ‘I understand. I am sure it is not easy to lose someone who you love.’ I wondered whether he had been in love, or whether his work was his love. ‘Sometimes, you need to play the long game: a quick fix is not always best. Please try to trust me. I think we have learnt more than you realize: Alice’s relationship scores are generally high. When we ignored her extreme score

There are several ways to quantify how well the mean ‘fits’ the data. The deviation or error is the distance of each score from the mean.

The sum of squared errors is the total amount of error in the mean. The errors/deviances are squared before adding them up.

The variance is the average distance of scores from the mean. It is the sum of squares divided by the number of scores. It tells us about how widely dispersed scores are around the mean, and is also a measure of how well the model ‘fits’ the observed data. If you have all of the population data it can be calculated as:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

If you want to estimate the population variance from a sample of data, then use this formula instead:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n - 1}$$

The standard deviation is the square root of the variance. It is the variance converted back to the original units of measurement of the scores used to compute it. For this reason it is the most commonly cited measure of the ‘fit’ of the mean, or the dispersion of scores around the mean.

Large standard deviations relative to the mean suggest data are widely spread around the mean (it is a poor ‘fit’); whereas small standard deviations suggest data are closely packed around the mean (it is a good ‘fit’).

The range is the distance between the highest and lowest score.
The interquartile range is the range of the middle 50% of the scores.
When describing data with a measure of central tendency use $M$ (capital M in italics) to represent the sample mean or mathematical notation ($\bar{x}$). Be consistent: if you choose $M$ to represent the mean then use it throughout your write-up. If you report a summary of the data such as the mean, you should also report the appropriate measure of ‘fit’: in the case of the mean that would be the spread of scores. Therefore, it’s common that when scientists report the mean, they also report the variance or standard deviation. The standard deviation is usually denoted by $SD$, but it is also common to simply place it in parentheses as long as you indicate that you’re doing so in the text. Here are some examples:

- Zach has 2 ‘emos’ on memoryBank. On average, a sample of other users ($N = 11$), had considerably more, $M = 95$, $SD = 56.79$.
- Alice’s average RAS score, $\bar{x} = 30$, $SD = 1.22$, was higher than the national average.
- On average ($SD$ in parentheses), people spend 3.45 (0.25) hours a day on their Proteus and drink 1.33 (0.25) lactose-free lattes a day.

Note that in the first example, I used $N$ to denote the size of the sample. Although a capital $N$ usually denotes the population size, it is common to see a capital $N$ representing the entire sample and a lower-case $n$ representing a subsample (e.g., the number of cases within a particular group).

When we report medians, a common abbreviation is $Mdn$, and again it is useful to report the range or interquartile range as well ($IQR$ is commonly used for the interquartile range). Therefore, we could report:

- Alice’s RAS scores, $Mdn = 30$, $IQR = 2$, were higher than the national average.
- Alice’s RAS scores, $Mdn = 30$, range = 21, were higher than the national average.

Alice’s Lab Notes 4.1 Reporting descriptive statistics

her mean satisfaction was 30, and the standard deviation was very small, 1.22, which suggests that this is a good “fit” of the population of all of Alice’s satisfaction scores. It seems that she is very satisfied with her relationship with you. It is understandable when you look at the lengths you are going to in order to find her. Now, she had a very low score on the last week, but even with this score included her average satisfaction was 28.1 out of a maximum of 32. There is, perhaps, a question to ask about why her satisfaction levels were so low in that last week, but the fact remains that over the last 10 weeks, on the whole she is very satisfied.’

I felt a little overwhelmed; I wasn’t sure whether it was hearing that Alice liked being with me, or just the stress of the last few days letting itself out, but I hugged the cat, and took a deep breath to control myself. ‘Steady on!’ he said as he extricated himself from my embrace.
KEY TERMS

- Arithmetic mean
- Average
- Biased estimator
- Bimodal
- Central tendency
- Degrees of freedom
- Deviance
- Fit
- Interquartile range
- Lower quartile
- Mean
- Mean squared error
- Median
- Mode
- Multimodal
- Noniles
- Outlier
- Percentiles
- Population mean
- Quantiles
- Quartiles
- Range
- Residual
- Sample mean
- Standard deviation
- Sum of squared errors
- Unbiased estimator
- Upper quartile
- Variance

JIG:SAW’S PUZZLES

1. What does the variance measure?
2. Why does the variance in a sample underestimate the variance in the population?
3. What does a small standard deviation relative to the mean tell us about our data?
4. Milton recruited a group of nine cats and recorded how many lactose-free lattes they drank in a week: 7, 9, 16, 20, 21, 28, 26, 32, 45. Calculate the mean, median and mode of these data.
5. It seems that people are spending more and more time on their electronic devices. Zach asked a group of 20 people how long (in minutes) they spend on their Proteus each day: 65, 125, 34, 90, 45, 25, 10, 22, 22, 24, 30, 50, 60, 65, 34, 90, 100, 15, 20, 35. Calculate the sum of squares, variance and standard deviation of these data.
Would you say that the mean in puzzle 5 ‘fits’ the data well? Explain your answer.

While Zach was worrying about whether Alice had left him, he ruminated about how successful couples often seem to divorce. Alice is a brilliant scientist and he a brilliant musician, so perhaps their relationship is doomed. To see if his observation might be true he got The Head to check the (approximate) length in days of some celebrity marriages from before the revolution: 240 (J-Lo and Cris Judd), 144 (Charlie Sheen and Donna Peele), 143 (Pamela Anderson and Kid Rock), 72 (Kim Kardashian, if you can call her a celebrity, and Chris Humphries), 30 (Drew Barrymore and Jeremy Thomas), 26 (Axl Rose and Erin Everly), 2 (Britney Spears and Jason Alexander), 150 (Drew Barrymore again, but this time with Tom Green), 14 (Eddie Murphy and Tracy Edmonds), 150 (Renee Zellweger and Kenny Chesney), 1657 (Jennifer Aniston and Brad Pitt). Compute the mean, median, standard deviation, range and interquartile range for these lengths of celebrity marriages.

Repeat puzzle 7 but excluding Jennifer Aniston and Brad Pitt’s marriage. How does this affect the mean, median, range, interquartile range and standard deviation? What do the differences in values between puzzles 7 and 8 tell us about the influence of unusual scores on these measures?

Zach asked Nick to get 15 of their fans on memoryBank to rate his new song, ‘The Gene Mixer’, on a scale ranging from 0 (the worst thing you’ve ever recorded) to 10 (the best thing you’ve ever recorded). The ratings were: 3, 5, 7, 8, 2, 4, 10, 8, 5, 7, 9, 10, 6. Calculate the mean, standard deviation, median, range and interquartile range for these ratings of the song.

Is the mean in puzzle 9 a good ‘fit’ to the data? Explain your answer.

IN THE NEXT CHAPTER ZACH DISCOVERS …

The cost of striving for perfection
Types of graphs
How to present data
How to avoid chartjunk
Never to show a pie chart to a man who has attacked you with bulldog clips