13 Extrasensory perception

More than a hundred years ago, some of the world's most renowned scientists were curious whether or not people can look into the future (precognition), move objects with their mind (telekinesis), or transmit messages by thought (telepathy). These so-called extrasensory abilities were studied by members of the *Society for Psychical Research*, a society that included intellectual heavyweights such as William James, Carl Jung, and Alfred Wallace. Even as late as 1950, the great Alan Turing argued that his famous test for artificial intelligence should be carried out in a telepathy-proof room.

How times have changed. Scientific work on extrasensory perception, or ESP, is now conducted only by a few self-pronounced academic mavericks, on a Quixotic mission to demonstrate to the world that the phenomenon is real. In 2011, the debate on the existence of ESP was re-ignited when reputable social psychologist Dr Daryl Bem published nine ESP experiments with over 1000 subjects (Bem, 2011). On the basis of these data, Bem argued that people are able to look into the future. In Bem's first experiment, for example, subjects had to guess whether a picture was going to appear on the left or the right side of the computer screen. The location of the picture was random, and this means that, on average, subjects cannot do better than a chance rate of 50% correct—unless, of course, people can look into the future. Bem (2011) found that people guessed the upcoming location of the pictures with above-chance accuracy of 53.1%. Interestingly, this effect occurred only for erotic pictures, and was absent for neutral pictures, romantic but not erotic pictures, negative pictures, and positive pictures. It was also found that the effect was largest for extravert women.

Do the Bem studies show that people can look into the future? Hardly. The Bem studies have been criticized on several grounds (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Here we use Bayesian parameter estimation and model

Box 13.1 Turing on telepathy

"I assume that the reader is familiar with the idea of extra-sensory perception, and the meaning of the four items of it, *viz.* telepathy, clairvoyance, precognition and psycho-kinesis. These disturbing phenomena seem to deny all our usual scientific ideas. How we should like to discredit them! Unfortunately the statistical evidence, at least for telepathy, is overwhelming." (Turing, 1950, p. 453).

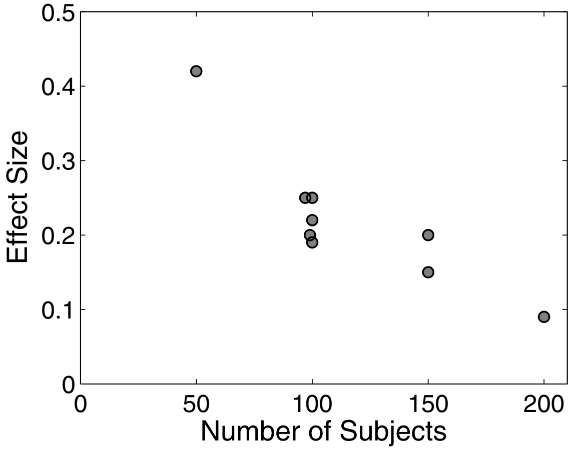


Fig. 13.1 The relationship between the number of subjects and the effect size for the experiments reported by Bem (2011).

selection methods to explore one of the criticisms, namely that the original Bem results had been obtained by optional stopping. We also analyze data from a replication experiment and assess the evidence for stable individual differences in ability, and for the effect of extraversion.

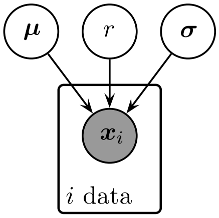
13.1 Evidence for optional stopping

When researchers report p-values they often do not realize that they have to specify a sampling plan in advance of data collection. When you state that you are going to test 100 subjects, you are not allowed to take sneak peeks at the data and stop whenever the result is significant (e.g., $p < 0.05$); nor are you allowed to test more than 100 subjects in case the outcome of your test is ambiguous (e.g., $p = 0.09$). The reason for this requirement is that researchers who take sneak peeks at their data can achieve any desired p-value, no matter how low, even if the null hypothesis is exactly true. For this reason, the optional stopping procedure is also known as "sampling to a foregone conclusion."

For a single study it can be very difficult to determine whether or not the results were due to optional stopping. Whenever a researcher reports multiple studies, however, a diagnostic tool is to plot the number of subjects or observations against the effect size (Hyman, 1985). Figure 13.1 shows this relation for the experiments reported by Bem.[1]

The negative relation between the number of subjects and effect size suggests that the results are contaminated by optional stopping. When the effect size is

[1] We thank Ray Hyman for attending us to this regularity.



$$\mu_1, \mu_2 \sim \text{Gaussian}(0, 0.001)$$
$$\sigma_1, \sigma_2 \sim \text{InvSqrtGamma}(0.001, 0.001)$$
$$r \sim \text{Uniform}(-1, 1)$$
$$\boldsymbol{x}_i \sim \text{MvGaussian}\left((\mu_1, \mu_2) , \begin{bmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1} \right)$$

Fig. 13.2 Graphical model for inferring a correlation coefficient.

large, the researcher can afford to stop early. When the effect size is small, more subjects need to be tested before the result reaches significance.

How strong is the evidence that there is a negative association between sample size and effect size? Within the Bayesian framework, two natural ways to address this question come from the parameter estimation and model selection perspectives on inference. Parameter estimation involves inferring the posterior distribution of the correlation coefficient, as covered in Section 5.1. Model selection involves comparing, for example, the hypothesis that the correlation is zero to the hypothesis that the correlation is some other value. This can be done by applying the Savage–Dickey density ratio test, as covered in Section 7.6.

A graphical model for inferring the correlation coefficient is shown again in Figure 13.2. The script `Correlation_1.txt` implements the graphical model in WinBUGS:

```
# Pearson Correlation
model{
  # Data
  for (i in 1:n){
    x[i,1:2] ~ dmnorm(mu[],TI[,])
  }
  # Priors
  mu[1] ~ dnorm(0,.001)
  mu[2] ~ dnorm(0,.001)
  lambda[1] ~ dgamma(.001,.001)
  lambda[2] ~ dgamma(.001,.001)
  r ~ dunif(-1,1)
  # Reparameterization
  sigma[1] <- 1/sqrt(lambda[1])
  sigma[2] <- 1/sqrt(lambda[2])
  T[1,1] <- 1/lambda[1]
  T[1,2] <- r*sigma[1]*sigma[2]
  T[2,1] <- r*sigma[1]*sigma[2]
  T[2,2] <- 1/lambda[2]
  TI[1:2,1:2] <- inverse(T[1:2,1:2])
}
```

The graphical model is used to infer the posterior distribution of the correlation coefficient, assuming the prior distribution is $r \sim \text{Uniform}\big(-1, 1\big)$. In other words, all values of the correlation coefficient are deemed equally likely a priori. In hypothesis testing or model selection terms, this corresponds to the alternative hypothesis

Box 13.2 Flexibility of Bayesian data collection

"The rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience." (Edwards et al., 1963, p. 193) . . . "if you set out to collect data until your posterior probability for a hypothesis which is unknown to you is true has been reduced to .01, then 99 times out of 100 you will never make it, no matter how many data you, or your children after you, may collect." (Edwards et al., 1963, p. 239)

$\mathcal{H}_1$. Under the null hypothesis $\mathcal{H}_0$, the assumption is that there is no correlation. Thus, using the Savage-Dickey density ratio method, the Bayes factor is simply the height of the prior divided by the height of the posterior, evaluated at the point of test $r = 0$.

The code `OptionalStopping.m` or `OptionalStopping.R` applies the graphical model to the data from Figure 13.1, plots the posterior distribution, and applies the Savage–Dickey method.

The results are shown in Figure 13.3, with the left panel showing the data again for convenience. The right panel shows the prior (horizontal dotted line) and posterior (solid line) distribution for the correlation coefficient. The expected value of the posterior is about -0.77, and the mode is near the frequentist value of -0.87 shown by the broken vertical line.

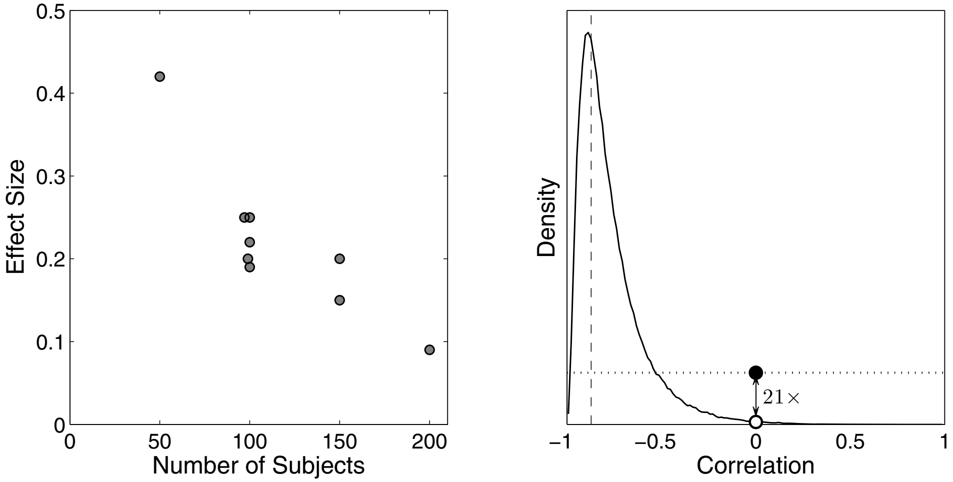


Fig. 13.3 Raw data (left panel), and Bayesian analysis (right panel) for the correlation between sample size and effect size in the Bem (2011) experiments.

Box 13.3 Boring Bayes

"What is the principal distinction between Bayesian and classical statistics? It is that Bayesian statistics is fundamentally boring. There is so little to do: just specify the model and the prior, and turn the Bayesian handle. There is no room for clever tricks or an alphabetic cornucopia of definitions and optimality criteria. I have heard people who should know better use this 'dullness' as an argument against Bayesianism. One might as well complain that Newton's dynamics, being based on three simple laws of motion and one of gravitation, is a poor substitute for the richness of Ptolemy's epicyclic system." (Dawid, 2000, p. 326)

The right panel of Figure 13.3 also shows the density of the prior and posterior at the point of test $r = 0$ by black and white circles, respectively. The density of the prior is about 21 times greater than that of the posterior at this point, so the Bayes factor is about 21 in favor of the alternative hypothesis that the correlation is not zero.

Exercises

Exercise 13.1.1 What does the Bayesian analysis tell you about the association between sample size and effect size in the Bem (2011) studies?

Exercise 13.1.2 Section 5.2 considered extending the correlation model in Figure 13.2 to incorporate uncertainty about the measures being related. Could that extension be usefully applied here?

Exercise 13.1.3 A classical p-value test on the Pearson product-moment correlation coefficient yields $r = -0.87$, 95% CI $= [-0.97, -0.49]$, $p = 0.002$. What conclusions would you draw from this analysis, and how do they compare to the conclusions you drew from the Bayesian analysis?

Exercise 13.1.4 We do not need to compute the Savage–Dickey density ratio on the original scale. For example, there are good arguments first to transform the posterior samples using the Fisher z-transform, so that $z = \text{arctanh}\,(r)$. Try using this transformation. What difference do you observe?

13.2 Evidence for differences in ability

Wagenmakers, Wetzels, Borsboom, van der Maas, and Kievit (2012) conducted a replication of Bem's (2011) original experiment, which differed in a few details. Because Wagenmakers et al. (2012) wanted to maximize the probability of finding an effect, they tested only women, and included only neutral and erotic pictures.

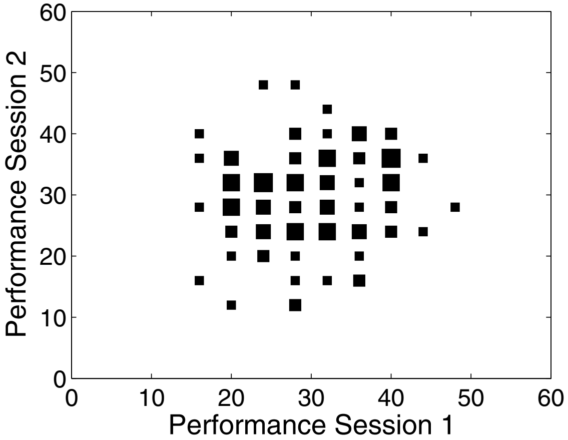


Fig. 13.4 Performance of 100 subjects in two sessions, each with 60 trials, in correctly identifying the hidden location of erotic pictures. The size of each point indicates the number of subjects with that combination of correct predictions.

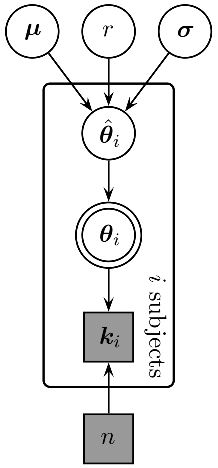
Another difference was that they included two consecutive sessions, reasoning that "If participants have ESP, this trait should be related from session 1 to session 2. In other words, individual differences in ESP express themselves statistically as a positive correlation between performance on erotic pictures for session 1 and session 2."

Figure 13.4 shows the ability of all 100 subjects, on two sessions of 60 trials, to identify the hidden locations of erotic pictures. The visual impression is that there is no systematic association between performance on session 1 and session 2.

As before, it makes sense to make inferences about the correlation coefficient, and test alternative hypotheses. Since the hypothesis being tested is specifically about the possibility of a positive correlation, the alternative hypothesis $\mathcal{H}_1$ now states that the correlation is positive, and so uses the prior distribution $r \sim \text{Uniform}\big(0, 1\big)$.

Another difference in this example is that it is clear how to model the uncertainty in psychological variables that generate the behavioral measures. Performance on the two sessions is simply a count of the number of correct responses for each person, assumed to be generated by an underlying ability. Thus, if the ith person has k_{i1} correct answers in the first session out of $n = 60$ trials, this performance is related to an underlying rate of correctly responding θ_{i1} on the first session by $k_{i1} \sim \text{Binomial}\big(\theta_{i1}, n\big)$. In this way, by providing a complete account of the probabilistic process by which the observed data are generated, the inherent uncertainty in the underlying abilities for people and sessions is naturally taken into account.

Figure 13.5 shows a graphical model for inferring the correlation coefficient between performance on the first and second session, and for modeling the subjects' behavior in making correct decisions from underlying abilities. The script `Ability.txt` implements the graphical model in WinBUGS:



$$\begin{aligned}
\mu_1, \mu_2 &\sim \text{Gaussian}(0, 0.001) \\
\sigma_1, \sigma_2 &\sim \text{InvSqrtGamma}(0.001, 0.001) \\
r &\sim \text{Uniform}(0, 1) \\
\hat{\boldsymbol{\theta}}_i &\sim \text{MvGaussian}\left((\mu_1, \mu_2)\,, \begin{bmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1}\right) \\
\theta_{ij} &\leftarrow \Phi(\hat{\theta}_{ij}) \\
k_{ij} &\sim \text{Binomial}(\theta_{ij}, n)
\end{aligned}$$

Fig. 13.5 Graphical model for inferring the correlation coefficient between performance across subjects on the first and second session of the ESP replication experiment.

```
# Ability Correlation for ESP Replication
model{
  # Data
  for (i in 1:nsubjs){
    thetap[i,1:2] ~ dmnorm(mu[],TI[,])
    for (j in 1:2){
      theta[i,j] <- phi(thetap[i,j])
      k[i,j] ~ dbin(theta[i,j],ntrials)
    }
  }
  # Priors
  mu[1] ~ dnorm(0,.001)
  mu[2] ~ dnorm(0,.001)
  lambda[1] ~ dgamma(.001,.001)
  lambda[2] ~ dgamma(.001,.001)
  r ~ dunif(0,1)
  # Reparameterization
  sigma[1] <- 1/sqrt(lambda[1])
  sigma[2] <- 1/sqrt(lambda[2])
  T[1,1] <- 1/lambda[1]
  T[1,2] <- r*sigma[1]*sigma[2]
  T[2,1] <- r*sigma[1]*sigma[2]
  T[2,2] <- 1/lambda[2]
  TI[1:2,1:2] <- inverse(T[1:2,1:2])
}
```

The code `Ability.m` or `Ability.R` applies the graphical model to the data from Figure 13.4, plots the posterior distribution, and applies the Savage–Dickey method.

The results are shown in Figure 13.6. The left panel shows the inferred abilities, for each subject on each session. The circles show the expected value of the abilities for each subject, and the lines connect this expectation to a sample of points from the joint posterior. The right panel shows the prior (horizontal dotted line) and

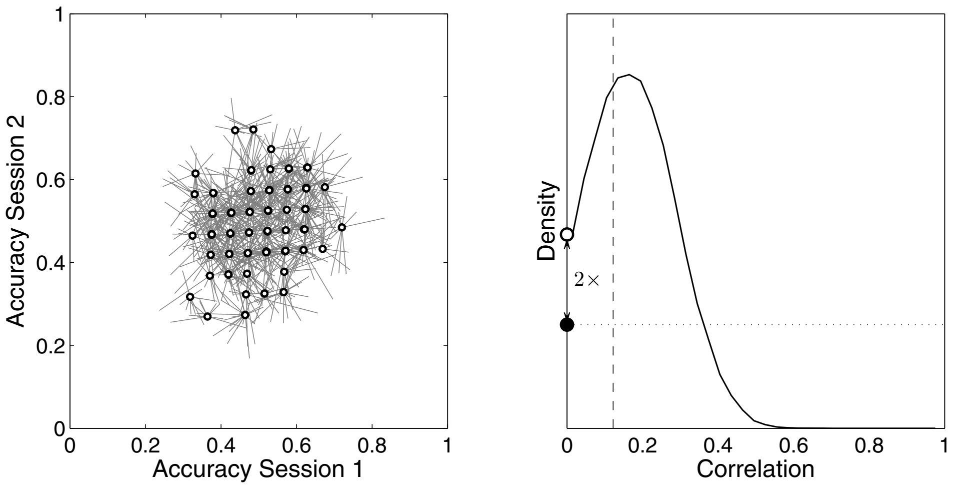


Fig. 13.6 Inferred abilities (left-hand panel) and correlation analysis (right-hand panel) for the relationship between sample size and effect size in ESP replication experiments.

posterior (solid line) distribution for the correlation coefficient. The expected value of the posterior is about 0.18. The mode is near the frequentist value of 0.12 shown by the broken vertical line, but does not correspond as closely as Figure 13.3.

The right panel of Figure 13.6 also shows the density of the prior and posterior at the point of test $r = 0$ by black and white circles, respectively. The density of the posterior is about 2 times greater than that of the prior at this point, so the Bayes factor is about 2 in favor of the null hypothesis that the correlation is zero. This is an evidence level that Jeffreys (1961) called "not worth more than a bare mention."

Exercises

Exercise 13.2.1 Suppose that the alternative hypothesis does not assume a positive correlation between the abilities of subjects over the two sessions, but instead allows for any correlation, so that the prior is $r \sim \text{Uniform}\big(-1, 1\big)$. Intuitively, what is the value of the Bayes factor in this case?

Exercise 13.2.2 A classical analysis yields $r = 0.12$, 95% CI $= [-0.08, 0.31]$, $p = 0.23$. This non-significant p-value, however, fails to indicate whether the data are ambiguous or whether there is evidence in favor of $\mathcal{H}_0$. How does the Bayes factor resolve this ambiguity?

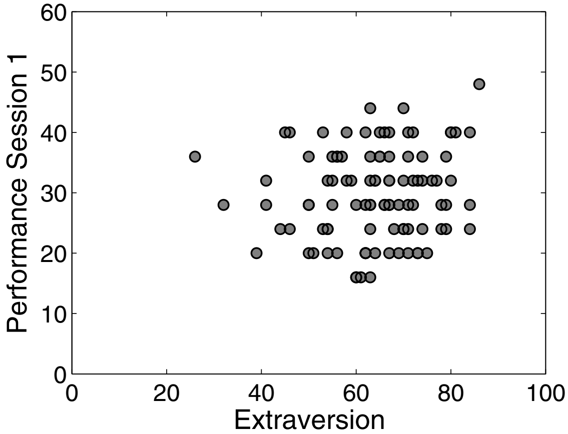


Fig. 13.7 The extraversion score, and performance on the first session, for 100 subjects in the ESP replication experiment

13.3 Evidence for the impact of extraversion

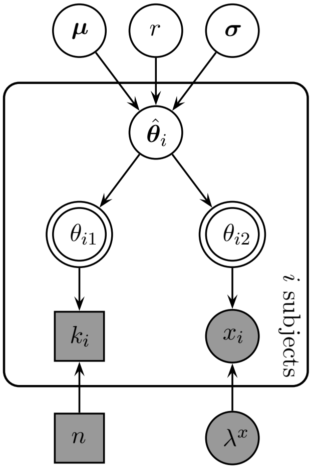
Wagenmakers et al. (2012), following a suggestion of Bem (2011), also considered the possibility of there being a positive correlation between performance and extraversion across subjects. Figure 13.7 shows the data for the extraversion score of each subject, and their performance on the first session. The visual impression is that there is no strong correlation.

The performance on the first session can be modeled as before, but modeling the inherent uncertainty in extraversion requires additional assumptions. One approach, also considered in Section 5.2, is to treat each extraversion score as the mean of a Gaussian distribution with some standard deviation. The assumed value of the standard deviation then corresponds to the assumed precision of the psychometric instrument used to generate the observed score.

A graphical model that represents this approach is shown in Figure 13.8. For the ith subject, the counts of correct predictions k_i is modeled as before, with $k_i \sim \text{Binomial}\big(\theta_{i1}, n\big)$. Their extraversion score x_i is modeled as $x_i \sim \text{Gaussian}\big(\theta_{i2}, \lambda^x\big)$, where $\theta_{i2} = 100\Phi\big(\hat{\theta}_{i2}\big)$ is the underlying true extraversion on a 0–100 scale, and λ^x is the precision of the Gaussian. Note that the prior on the correlation coefficient has reverted to $r \sim \text{Uniform}\big(-1, 1\big)$.

The script `Extraversion.txt` implements the graphical model in WinBUGS:

```
# Extraversion Correlation for ESP Replication
model{
  # Data
  for (i in 1:nsubjs){
    thetap[i,1:2] ~ dmnorm(mu[],TI[,])
    theta[i,1] <- phi(thetap[i,1])
    k[i] ~ dbin(theta[i,1],ntrials)
```



$$\begin{aligned}
\mu_1, \mu_2 &\sim \text{Gaussian}(0, 0.001) \\
\sigma_1, \sigma_2 &\sim \text{InvSqrtGamma}(0.001, 0.001) \\
r &\sim \text{Uniform}(-1, 1) \\
\hat{\boldsymbol{\theta}}_i &\sim \text{MvGaussian}\left((\mu_1, \mu_2), \begin{bmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1}\right) \\
\theta_{i1} &\leftarrow \Phi(\hat{\theta}_{i1}) \\
\theta_{i2} &\leftarrow 100\Phi(\hat{\theta}_{i2}) \\
k_i &\sim \text{Binomial}(\theta_{i1}, n) \\
x_i &\sim \text{Gaussian}(\theta_{i2}, \lambda^x)
\end{aligned}$$

Fig. 13.8 Graphical model for inferring the correlation coefficient between performance across subjects on the first and second blocks of the ESP replication experiment.

```
    theta[i,2] <- 100*phi(thetap[i,2])
    x[i] ~ dnorm(theta[i,2],lambdax)
  }
  # Priors
  mu[1] ~ dnorm(0,.001)
  mu[2] ~ dnorm(0,.001)
  lambda[1] ~ dgamma(.001,.001)
  lambda[2] ~ dgamma(.001,.001)
  r ~ dunif(-1,1)
  # Reparameterization
  sigma[1] <- 1/sqrt(lambda[1])
  sigma[2] <- 1/sqrt(lambda[2])
  T[1,1] <- 1/lambda[1]
  T[1,2] <- r*sigma[1]*sigma[2]
  T[2,1] <- r*sigma[1]*sigma[2]
  T[2,2] <- 1/lambda[2]
  TI[1:2,1:2] <- inverse(T[1:2,1:2])
}
```

The code `Extraversion.m` or `Extraversion.R` applies the graphical model to the data from Figure 13.7, plots the posterior distribution, and applies the Savage-Dickey method. Note that the code makes the assumption about the extraversion test precision on the standard deviation scale, which seems an easier one to express the uncertainty of measurement, and converts it to a precision to supply to the graphical model.

The results when $\lambda^x = 1/9$—that is, when the standard deviation is 3 for the extraversion measure—are shown in Figure 13.9. The left panel shows the inferred ability on the first session, and underlying level of extraversion, for each subject. The circles show the expected values, and the lines connect this expectation to a sample of points from the joint posterior. The right panel shows the prior (horizontal

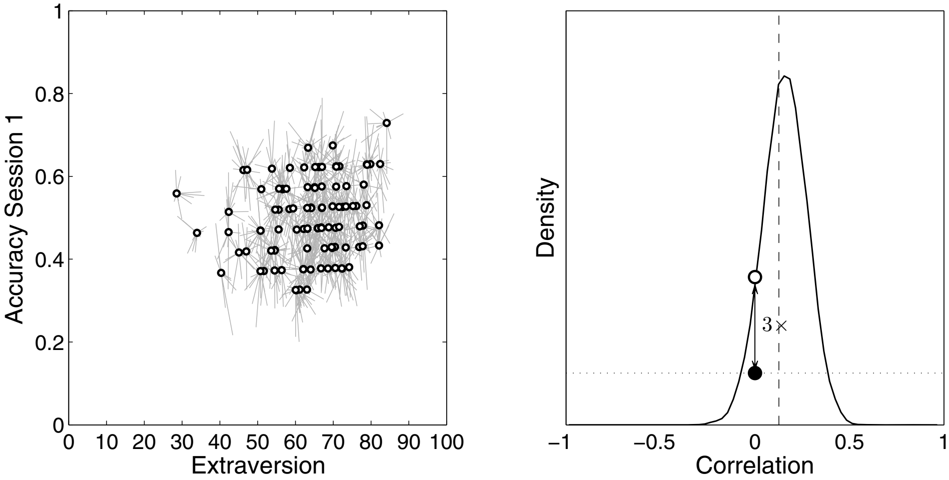


Fig. 13.9 Inferred extraversion measures and abilities (left-hand panel), and correlation analysis (right-hand panel) for the relationship between extraversion and performance in the ESP replication experiment.

dotted line) and posterior (solid line) distribution for the correlation coefficient. The expected value of the posterior is about 0.16. The mode is near the frequentist value of 0.12 shown by the broken vertical line.

The right panel of Figure 13.9 also shows the density of the prior and posterior at the point of test $r = 0$ by black and white circles, respectively. The density of the posterior is about 3 times greater than that of the prior at this point, so the Bayes factor is about 3 in favor of the null hypothesis that the correlation is zero.

Exercises

Exercise 13.3.1 What do you conclude about whether or not the correlation is zero, based on the Bayes factor?

Exercise 13.3.2 Try more extreme assumptions about the accuracy with which extraversion is measured, by setting $\lambda^x = 1$ and $\lambda^x = 1/100$. How does the Bayes factor change in response to this change in available information?