A Tutorial on Fisher Information

Alexander Ly, Josine Verhagen, Raoul Grasman, and Eric-Jan Wagenmakers

University of Amsterdam

Abstract

The concept of Fisher information plays a crucial role in many statistical applications that are of key importance to mathematical psychologists. Here we explain Fisher information by illustrating its use across three different statistical paradigms: first, in the frequentist paradigm, Fisher information is used to determine the sample size with which we design an experiment; second, in the Bayesian paradigm, Fisher information is used to define a default parameter prior; finally, in the minimum description length paradigm, Fisher information is used to measure model complexity. Each application is illustrated with simple examples.

Keywords: Statistical inference; Jeffreys' prior; Model complexity; Model selection; Minimum description length

Mathematical psychologists develop and apply statistical models in order to describe human behavior and understand latent psychological processes. Examples include Stevens' law of psychophysics that describes the relation between the objective physical intensity of a stimulus and its subjectively experienced intensity (Stevens, 1957); Ratcliff's diffusion model

This work was supported by the starting grant "Bayes or Bust" awarded by the European Research Council. Correspondence concerning this article may be addressed to Alexander Ly, University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, the Netherlands. Email address: alexander.ly.nl@gmail.com.

of decision making that quantifies the different processes that drive behavior in speeded response time tasks (Ratcliff, 1978); multinomial processing tree models that decompose performance in memory tasks into the contribution of separate latent mechanisms (Batchelder & Riefer, 1980; Chechile, 1973), and so on and so forth.

When applying their models to data, mathematical psychologists may operate from within different statistical paradigms and focus on different substantive questions. For instance, working within the classical or frequentist paradigm a researcher may wish to decide upon the number of trials to be presented to a participant in order to estimate the participant's latent abilities. Working within the Bayesian paradigm a researchers may wish to know how to determine a suitable default prior on the model parameters. Working within the minimum description length (MDL) paradigm a researcher may wish to compare rival models and quantify their complexity. Despite the diversity of these paradigms and purposes, they are connected through the concept of Fisher information.

Fisher information plays a pivotal role throughout statistical modeling, but an accessible introduction for mathematical psychologists is lacking. The goal of this tutorial is to fill this gap and illustrate the use of Fisher information in the three statistical paradigms mentioned above: frequentist, Bayesian, and MDL. This work builds directly upon the *Journal of Mathematical Psychology* tutorial article by Myung (2003) on maximum likelihood estimation. The intended target group for this tutorial are graduate students and researchers with an affinity for cognitive modeling and mathematical statistics.

The outline of this article is as follows: To keep this tutorial self-contained, we start by establishing the standard notations for probabilistic modeling and introduce two running examples featuring the Bernoulli distribution and the normal distribution. The next section provides the definition of Fisher information and shows how it can be calculated in general. The next three sections exemplify the use of Fisher information for different purposes. Section 1 shows how Fisher information can be used to determine the desired sample size for an experiment; Section 2 shows how Fisher information can be used to define a default prior on model parameters, and Section 3 shows how Fisher information can be used to measure model complexity.

Basic Concepts

For concreteness, consider an experiment comprised of n trials of equal difficulty. The goal of the experiment is to determine a participant's ability to identify particular species of animals. On each trial, the participant is presented with a picture of an animal on a computer screen and is asked to identify the species. On each trial, the participant can either respond correctly or in error. We assume that if the participant does not recognize the species, the probability of guessing the correct answer is close to zero.

Notation Let n denote the planned number of trials in an experiment and let $\vec{X} = (X_1, \ldots, X_n)$ denote a *future* trial sequence that will be presented to a participant. More generally, we call $\vec{X}$ a random variable as we are uncertain about its outcomes. In a typical experimental set up, we design each trial X_i to be a replication of a prototypical trial, which we denote by X. The prototypical trial in our running example can take on two outcomes x, and we write $x = 1$ when the participant responded correctly, and $x = 0$ otherwise. Similarly, we denote $\vec{x} = (x_1, x_2, \ldots, x_n)$ for an outcome of the n-trial sequence $\vec{X}$. For example, after a participant completed the experiment with $n = 10$ trials, one possible outcome is $\vec{x} = (1, 0, 0, 0, 0, 0, 0, 1, 0, 0)$, implying that the participant responded correctly only on the first and the eighth trial. In order to model the chance with which this outcome occurs we construct a probabilistic model: the Bernoulli model.

Example 1 (The Bernoulli model)**.** *To model the behavior of the participant, we first construct a model for all possible outcomes of the prototypical trial X, assuming all such outcomes are caused by the participant's general "species recognition ability", which we denote by θ. This latent ability θ relates to the possible outcomes as follows:*

$$f(x \,|\, \theta) = P(X = x) = \theta^x (1 - \theta)^{1-x}, \text{ where } x = 0 \text{ or } x = 1. \tag{1}$$

We call this relation a probability density function (pdf).[1] *The pdf allows us to deduce the*

[1] Formally, Eq. (1) defines a probability mass function, as X takes on discrete values; for brevity we do not distinguish between the two here.

chances with which an outcome occurs when the latent ability is known. Hence, if $\theta = 1$ *the participant always correctly recognizes the animal species, as the chance of a correct response is 1. Conversely, if* $\theta = 0$ *the participant never correctly recognizes the animal species, as the chance of a correct response is 0. In reality, a participant's latent ability is likely to fall somewhere in between these two extremes,* $\theta \in (0, 1)$*. The relation Eq.* (1) *describes how a participant's true latent ability* θ *"causes" the behavioral outcomes* $\vec{x}$*. This particular relation is known as the Bernoulli model, denoted* $X \sim Ber(\theta)$*. As the full trial sequence* $\vec{X}$ *consists of n independent and identically distributed (i.i.d.) replications of* X *we can extrapolate a pdf for the vector-valued outcomes* $\vec{x}$ *by taking n products of Eq.* (1)*:*

$$f(\vec{x}\,|\,\theta) = \underbrace{\theta^{x_1}(1-\theta)^{1-x_1}}_{X_1 \sim Ber(\theta)} \cdot \ldots \cdot \underbrace{\theta^{x_n}(1-\theta)^{1-x_n}}_{X_n \sim Ber(\theta)} = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}. \tag{2}$$

Note that with $n = 10$ *we have* $2^{10} = 1024$ *possible outcomes* $\vec{x}$ *for the experiment* $\vec{X}$*.*

Typically, we do not report the raw outcome series $\vec{x}$ *but only the number of successful responses* y*. We can view* y *as an outcome of the summary statistic* $Y = \sum_{i=1}^{n} X_i$*. A statistic is by definition a function of possible outcomes and it therefore inherits the random mechanism of* $\vec{x}$*, in this case a participant's latent ability* θ*. The pdf that* Y *inherits from Eq.* (2) *is called a binomial distribution* $Y \sim Bin(\theta, n = 10)$*. For example, when* $\theta = 0.4$ *the chances of the outcomes are then given by*

$$P(Y = y) = f(y\,|\,n = 10, \theta = 0.4) = \frac{10!}{y!(10-y)!}(0.4)^y(0.6)^{10-y} \text{ for } y = 0, 1, \ldots, 10. \tag{3}$$

Note that Y *can only take on 11 values compared to the* $2^{10} = 1024$ *outcomes of* $\vec{X}$*, a compression rate of almost a hundred, due to the fact that* Y *ignores the order within* $\vec{x}$*, compare Eq.* (2) *with Eq.* (4)*. Fig. 1 shows the chances corresponding to the outcomes of* Y *with* $\theta = 0.4$*.*

Fig. 1 emphasizes the nature of probabilistic modeling as no possible outcome is entirely

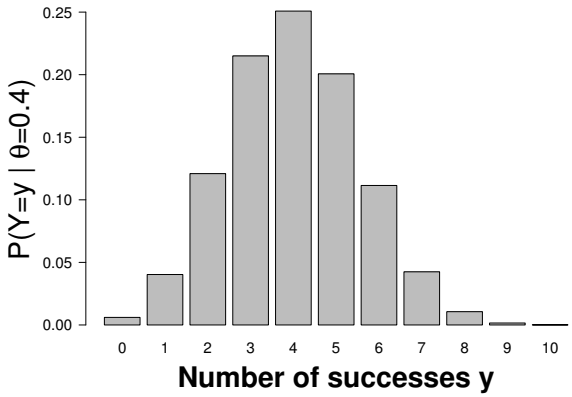

Figure 1. The chances of observing a particular number of successes in a binomial experiment with $n = 10$ and $\theta = 0.4$. The area under the bars sum to one. Note that, when a participant's latent ability is truly $\theta = 0.4$ an outcome of $Y = 0$ is still possible with chance $P(Y = 0 \,|\, \theta = 0.4) = 0.006$.

excluded. In particular, even when a participant's latent ability is truly $\theta = 0.4$ *there is a slim chance,* $P(Y = 0 \,|\, \theta = 0.4) = 0.006$*, that she will respond incorrectly on all trials. More generally, any number of trials* n *and latent ability* θ *between zero and one would yield outcomes of* Y *with the following probabilities:*

$$f(y \,|\, n, \theta) = \frac{n!}{y!(n-y)!}\theta^y(1-\theta)^{n-y} \text{ for } y = 0, 1, \ldots, n, \text{ for all } \theta \in (0, 1). \tag{4}$$

Equivalently, we say that the number of successes is modeled according to a binomial model, which we denote by $Y \sim Bin(\theta, n)$*. Moreover, this interpretation of Eq.* (4) *is in line with the idea that any outcome* y *is generated from the latent ability* θ*; schematically,* $\theta \overset{Bin(\theta,n)}{\Longrightarrow} y$*.* $\Diamond$

Likelihood In practice, we only observe data and do not know θ. To infer something about a participant's latent ability θ, we invert the data generating process by exchanging the roles of y and θ within a pdf. For example, when a participant responded correctly on seven out of ten trials, we plug this information into Eq. (4) and consider it as a function of possible latent abilities θ:

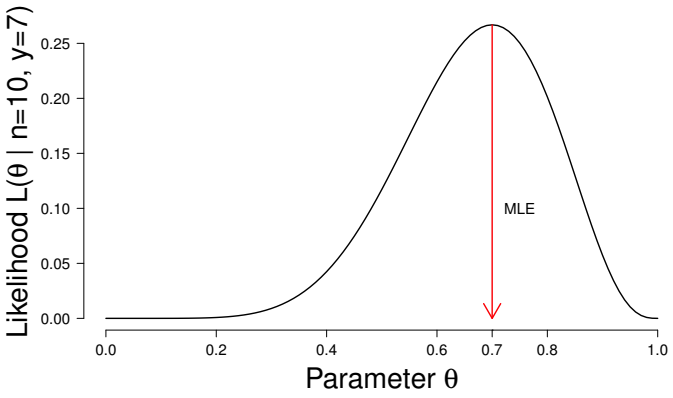

Figure 2. The likelihood function based on observing seven $y = 7$ out of ten $n = 10$ correct responses in a binomial experiment. The MLE equals $y/n = 0.7$. Note that the likelihood function is a continuous function of the parameter θ, whereas the observed data are discrete.

$$L(\theta \,|\, n = 10, y = 7) = \frac{10!}{7!3!}\theta^7(1-\theta)^3. \tag{5}$$

This function is shown in Fig. 2. To distinguish Eq. (5) from a pdf, we call it a likelihood function. Fisher used the likelihood function to develop his general purpose method of maximum likelihood estimation (Fisher, 1912; Fisher, 1922; Fisher, 1925; Myung, 2003). The idea is to use the modeled relation $f(\vec{x} \,|\, \theta)$ to "reverse engineer" the most likely value of θ that could have been responsible for the observed data. It can be easily shown that the maximum likelihood estimator (i.e., MLE) for the Bernoulli model is given by the sample mean $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Example 2 (The n-trial Normal Model)**.** ***Pre-experimental:*** *As a second example we consider the normal model, which forms the building block for common models in psychology (e.g., t-tests, ANOVA, and linear regression). In this model, the trial sequence $\vec{X}$ consists of n replications of a prototypical random variable X with a pdf given by*

$$f(x \,|\, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x-\mu^2}{2\sigma^2}\right), \tag{6}$$

where the parameter vector consists of the population mean and the population variance $\vec{\theta} = (\mu, \sigma^2)$. *To extrapolate the pdf for the n-trial experiment* $\vec{X}$ *we exploit the i.i.d. assumption by taking n products of Eq.* (6)*, which yields:*

$$f(\vec{x} \,|\, \vec{\theta}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right). \tag{7}$$

We then say that the possible outcomes follow a normal distribution, $X \sim \mathcal{N}(\mu, \sigma^2)$.[2]

Inference: *To invert the generating model we simply plug the observed data* $\vec{x}$ *the relation into Eq.* (7) *and the resulting likelihood function then consists of a two-dimensional surface spanned by both* μ *and* σ^2. *Maximizing the likelihood is equivalent to maximizing the natural logarithm of the likelihood function, i.e.,* $\log f(\vec{x} \,|\, \vec{\theta})$*, a function shown in Fig. 3 for fictitious data.*

For the normal model, the MLE consists of a pair $\hat{\mu}, \hat{\sigma}^2$*: the sample mean* $\hat{\mu} = \bar{X} = \frac{1}{n}X_i$ *and* $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. *See the online appendix for a derivation.* ◊

In Section 1 we will show how Fisher information is related to the performance of the aforementioned maximum likelihood estimators. First, however, we introduce the general definition of Fisher information and show how it can be computed for the two models discussed above.

Fisher Information

Definition of Fisher Information The (unit) Fisher information is a measure for the amount of information that is expected within the prototypical trial X about the parameter of interest θ. It is defined as the variance of the so-called score function, i.e., the derivative of the log-likelihood function with respect to the parameter,

[2] This is a slight abuse of notation; however, we believe the meaning is well understood, as we assume that the random variable $\vec{X} = X_1, \ldots, X_n$ consists of n i.i.d. copies of a prototypical trial X.

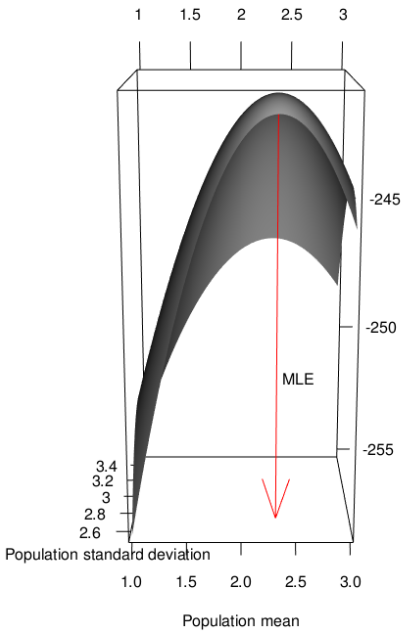

Figure 3. Log-likelihood function for 100 simulated normally distributed observations $\vec{x}$. The log-likelihood is a surface spanned by the two parameters μ and σ^2. On the x-axis is the population mean, and on the y-axis is the population standard deviation. The MLE $\binom{\hat{\mu}}{\hat{\sigma}} = \binom{2.32}{2.68}$ corresponds to the projection of the highest point on the log-likelihood surface onto the parameter space.

$$I(\theta) = \mathrm{Var}\Big(\underbrace{\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X\,|\,\theta)}_{\text{Score function}}\Big) \text{ or as } I(\theta) = -E\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X\,|\,\theta)\right), \tag{8}$$

under additional (mild) regularity conditions. To calculate $I(\theta)$ we keep θ fixed and take the expectation with respect to all possible outcomes x:

$$I(\theta) = -E\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X\,|\,\theta)\right) = -\int_{\mathcal{X}}\Big(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x\,|\,\theta)\Big) f(x\,|\,\theta)\,\mathrm{d}x, \tag{9}$$

where $\mathcal{X}$ denotes the outcome space of X. When X is discrete, we replace the integral by a summation, see Eq. (10) below.

Example 1 (Fisher Information in the Bernoulli Model, Continued)**.** *To calculate the information of a Bernoulli distribution, we take Eq.* (1) *and plug this into Eq.* (9) *which yields:*

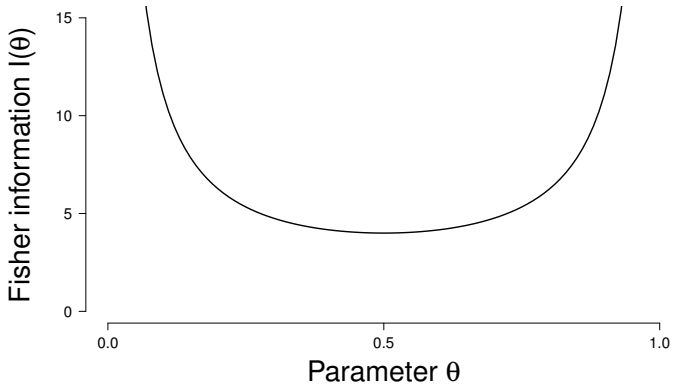

Figure 4. The Fisher information as a function of θ within the Bernoulli model. As θ reaches zero or one the expected information goes to infinity. When $\theta = 1$ the outcomes will always be 1, therefore clearly conveying this information within the data.

$$I(\theta) = -\sum_{x=0}^{1} \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log P(X = x)P(X = x) = -\Big(-\frac{1}{\theta^2}\theta - \frac{1}{(1-\theta)^2}(1-\theta)\Big) = \frac{1}{\theta(1-\theta)}. \quad (10)$$

Hence, when a participant's true latent ability is θ^* the expected information about her ability is then given by $I(\theta^*) = \frac{1}{\theta^*(1-\theta^*)}$ for a prototypical trial, for instance, $I(0.5) = 4$ when $\theta^* = 0.5$. Thus, the Fisher information contains two aspects of the model: (1) the sensitivity of the relationship $f(x \,|\, \theta)$ with respect to the parameter θ expressed by the score function at the true value θ^*, and (2) how this sensitivity at θ^* varies over (all possible) outcomes x a model can generate according to $f(x \,|\, \theta^*)$. This dependence on the true parameter value θ^* is shown in Fig. 4.

As the full trial sequence $\vec{X}$ consists of n i.i.d. replications of the prototypical X we formally refer to Eq. (8) *as the unit Fisher information. To extrapolate the information about θ within n trials we simply multiply the unit information by n, $I_n(\theta) = nI(\theta)$. For the* 10*-trial Bernoulli model, this means $I_n(\theta) = \frac{10}{\theta(1-\theta)}$. This connection can be formally shown by exploiting the i.i.d. assumption and by computing the integral Eq.* (8) *with the pdf Eq.* (2) *instead.*

Furthermore, we can also use Eq. (8) *to calculate the Fisher information about θ con-*

tained in the summary statistic Y for which we also get $I_Y(\theta) = \frac{10}{\theta(1-\theta)} = I_n(\theta)$. This means that no information is lost when we infer θ from the summary statistic Y rather than from the full trial sequence $\vec{X}$. ◊

Unit Fisher Information Within an Observed Sample The expected (unit) Fisher information as defined in Eq. (8) is an weighted average over all the possible outcomes of X. We might, however, be interested in the observed (unit) Fisher information within an observed sample $\vec{x}$ instead. We then replace the expectation in Eq. (9) by its empirical version

$$I_{\text{Obs}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(x_i \,|\, \theta). \tag{11}$$

For example, if we observed 7 successes out of 10 Bernoulli trials we have:

$$I_{\text{Obs}}(\theta) = \frac{1}{10}\sum_{i=1}^{10} \frac{x_i}{\theta^2} + \frac{n - x_i}{(1-\theta)^2} = \frac{1}{10}\left(\frac{7}{\theta^2} + \frac{3}{(1-\theta)^2}\right). \tag{12}$$

If the data are truly governed by $\theta^* = 0.7$, we would indeed expect to see $10\theta^* = 7$ successes and $10(1-\theta^*) = 3$ failures and the observed and expected Fisher information then coincide:

$$I_{\text{Obs}}(\theta^*) = \frac{1}{10}\left(\frac{10\theta^*}{(\theta^*)^2} + \frac{10(1-\theta^*)}{(1-\theta^*)^2}\right) = \frac{10}{10}\left(\frac{\theta^*}{(\theta^*)^2} + \frac{(1-\theta^*)}{(1-\theta^*)^2}\right) = \frac{1}{\theta^*(1-\theta^*)} = I(\theta^*). \tag{13}$$

On the other hand, if $\theta^* = 0.15$ the probability of seeing 7 out of 10 heads is about .01 and we then see that there is a big discrepancy between expected and observed Fisher information, $I(0.15) \approx 8$ versus $I_{\text{Obs}}(0.15) = 31.5$ respectively. This might imply that the hypothesis $\theta^* = 0.15$ was wrong to begin with.

In more realistic cases, we do not know θ^* and to calculate the observed Fisher information we replace θ^* by the MLE, which yields $I_{\text{Obs}}(\hat{\theta})$. For the examples we discuss this coincides with plugging in the MLE into the expected Fisher information, i.e., $I(\hat{\theta}) = I_{\text{Obs}}(\hat{\theta})$.[3]

[3] This equality holds for distributions that belong to the so-called exponential family, see Huzurbazar (1949).

Example 2 (Fisher Information in the Normal Model, Continued)**.** *When there are multiple parameters of interest, say* $\vec{\theta} = (\theta_1, \ldots, \theta_d)$*, the Fisher information turns into a* $d \times d$ *matrix. The* i, j*-th element of this matrix is then given by*

$$I(\vec{\theta})_{i,j} = Cov\left(\frac{\partial}{\partial \theta_i} \log f(X \mid \vec{\theta}), \frac{\partial}{\partial \theta_j} \log f(X \mid \vec{\theta})\right) \text{ or } I(\vec{\theta})_{i,j} = -E\left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X \mid \vec{\theta})\right), \tag{14}$$

under additional (mild) regularity conditions. For example, when X *is normally distributed with both* $\vec{\theta} = \binom{\mu}{\sigma^2}$ *unknown, the unit Fisher information matrix* $I(\vec{\theta})$ *turns into a* 2×2 *matrix, consisting of expectations of partial derivatives:*

$$I(\vec{\theta}) = -E\begin{pmatrix} \frac{\partial^2}{\partial \mu \partial \mu} \log f(x \mid \mu, \sigma^2) & \frac{\partial^2}{\partial \mu \partial \sigma^2} \log f(x \mid \mu, \sigma^2) \\ \frac{\partial^2}{\partial \sigma^2 \partial \mu} \log f(x \mid \mu, \sigma^2) & \frac{\partial^2}{\partial \sigma^2 \partial \sigma^2} \log f(x \mid \mu, \sigma^2) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}, \tag{15}$$

see the online appendix for a derivation. The off-diagonal elements are in general not zero, unless the pdfs are symmetric around its location μ*. When the data are modeled as* n *i.i.d. normal distributions, the expected Fisher information within* $\vec{X}$ *is then given by* $I_n(\vec{\theta}) = nI(\vec{\theta})$*. As in the Bernoulli model, the observed (unit) Fisher information can be calculated by plugging the MLE* $\hat{\sigma}$ *into its expected version Eq.* (15)*.* $\Diamond$

In this section we showed how to calculate the Fisher information, a measure for the expected amount of information within the trial sequence $\vec{X}$ about the parameter of interest θ. We now highlight the role of Fisher information in three different statistical paradigms: in the frequentist paradigm, Fisher information is used to determine the sample size with which we design an experiment; in the Bayesian paradigm, Fisher information is used to determine a default prior; and in the minimum description length paradigm, Fisher information is used to measure model complexity.

1. Using Fisher Information to Determine Sample Size

Introduction As mentioned above, a statistical model for a psychological process can be thought of as a particular mapping f between an individual's latent properties θ and that individual's possible behavior, i.e., the outcomes $\vec{x}$, schematically: $\theta \overset{f(\vec{x}\,|\,\theta)}{\Longrightarrow} \vec{x}$. Consequently, when we know an individual's latent ability θ, we then know the chances with which all possible outcomes $\vec{x}$ occur (see the discussion below Eq. (4)).

In practice, however, we only observe a single outcome $\vec{x}$ of size n and have to infer θ instead. We write θ^* for the true value of the parameter θ that uses the relationship $f(\vec{x}\,|\,\theta)$ to generate data $\vec{x}$. The goal of classical or frequentist point-estimation is to provide an educated guess of the true value θ^* by applying an estimator (i.e., a function), $T = t(\vec{X})$, to an anticipated outcome $\vec{x}$.

In this section we discuss how Fisher information can be used to determine the (asymptotically) least number of trials n that needs to be presented to a participant such that an estimator yields estimates at a certain level of accuracy. As n is decided upon pre-experimentally, we are uncertain of the outcomes of T. The randomness within T is inherited from the data that are assumed to be governed by $f(\vec{x}\,|\,\theta)$ and the participant's true latent ability θ^*. Furthermore, we call an observed outcome of T an estimate, which we denote as $t(\vec{x})$.

For instance, in Example 1 we infer a participant's latent ability θ using the mean estimator $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and write $\bar{x}$ for the sample mean, an outcome of $\bar{X}$. As Example 1 concerns an experiment we have to choose the number of trials n we plan to present to a participant.

Below we first show how n can be determined by the following scheme: (i) choose an estimator on which to base our inference, (ii) check whether this estimator is consistent, (iii) derive its sampling distribution, and (iv) derive n based on the sampling distribution for a certain target level of accuracy. We then apply this scheme to the normal model and show how it can be simplified by using the MLE in conjunction with the Fisher information.

Measuring the Performance of an Estimator: Consistency, Sampling Distributions, and Sample Size Determination **Step (i): Choose an estimator** When the

trial sequence $\vec{X}$ is assumed to be distributed according to the Bernoulli model, $X \sim \text{Ber}(\theta)$, as in the animal discrimination task, we typically infer θ using the mean estimator $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. **Step (ii): Assess consistency** The general belief is that the mean estimator is consistent, meaning that it is expected to extract more information about θ from an anticipated sample $\vec{x}$ as the number of trials n increases. This qualitative behavior can be deduced by studying the (asymptotic) mean and variance of the estimator, as the next example illustrates.

Example 1 (Consistent Estimation within the Bernoulli Model, Continued)**.** *Use θ^* to denote a participant's unknown true latent ability, which we plan to estimate with the mean estimator $\bar{X}$. To study the expected behavior of a generic estimator T on the anticipated data, we average its possible outcomes across the sample space, resulting in the mean of the estimator: $E(T)$. To do so for $T = \bar{X}$, we use the linearity of the expectation and the i.i.d. assumption to express $E(\bar{X})$ in terms of the mean of the prototypical trial X:*

$$E(\bar{X}) = E\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} E(X_i) \overset{i.i.d.}{=} \frac{n}{n}E(X) \tag{16}$$

$$= \sum_{x=0}^{1} x f(x \mid \theta^*) \overset{Eq.\ (1)}{=} 1\theta^* + 0(1-\theta^*) = \theta^*. \tag{17}$$

This implies that the mean estimator $E(\bar{X})$ is unbiased; however, to prove that $\bar{X}$ is consistent we also need to show that the anticipated outcomes of the estimator to concentrate near θ^* as n increases.

To this end we need to study the variance of T. In particular, to calculate $Var(\bar{X})$ we once again exploit the i.i.d. assumption and the fact that the variance of the prototypical X is given by $Var(X) = \theta^*(1-\theta^*)$:

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \overset{i.i.d.}{=} \frac{n\,Var(X)}{n^2} = \frac{\theta^*(1-\theta^*)}{n}. \tag{18}$$

Thus, the variance of $\bar{X}$ shrinks to zero as the number of trials n grows indefinitely. This

implies that the chance that $\bar{X}$ yield an estimate close to the true value goes to one and we therefore call $\bar{X}$ consistent.[4] ◊

Recall that when a participant's true latent ability is given by $\theta^* = 0.4$, there is a slim but non-zero chance that the participant never responds correctly in the animal discrimination task ($y = 0$; see Fig. 4). Such an extreme outcome will yield the estimate $\bar{x} = 0$, which is inaccurate as it is far from the true value $\theta^* = 0.4$. Hence, probabilistic models do not afford iron-clad guarantees: unrepresentative data will lead to inaccurate estimates.

However, Eq. (18) does justify the idea that an empirical scientist can obtain a more accurate measurement of a participant's latent ability θ by increasing n. In other words, consistency of $\bar{X}$, within the Bernoulli model, implies that the chance of $\bar{X}$ yielding an inaccurate estimate goes to zero, whenever the number of trials n is sufficiently large.

Step (iii): Derive sampling distribution Clearly, we cannot detain a participant and get her to sit through a million trials or more. Hence, to quantify "sufficiently large" we first obtain the sampling distribution of the estimator.

Example 1 (Sampling Distribution of the Sample Mean within the Bernoulli Model, Continued)**.** *Within the Bernoulli model we can consider $\bar{X}$ as a rescaling of the number of successes in n trials, since $\bar{X} = \frac{1}{n}Y$, where $Y = \sum_{i=1}^{n} X_i$. Consequently, by Eq.* (4) *we know that the sampling distribution of $\bar{X}$ is essentially a binomial with outcomes rescaled by the factor $\frac{1}{n}$. For instance, with $n = 10$ the estimator can take on the estimates (i.e., the outcomes) $0, 0.1, 0.2, \ldots, 0.9, 1.0$ which corresponds to $0, 1, 2, \ldots, 9, 10$ number of successes y.* ◊

This sampling distribution quantifies how both the true value of the parameter θ and the number of trials n affect the probable outcomes of the estimator $\bar{X}$, schematically: $(\theta, n) \overset{\text{Bin}(\theta,n)}{\Longrightarrow} \bar{x}$.[5] As experimenters we cannot control θ, but we can lower the chances of inaccurate estimates by increasing n. **Step (iv): Derive** n In the following example we

[4] Formally, we call an estimator T for θ consistent, if for every possible true value θ^*, the estimator converges in probability, i.e., $\lim_{n\to\infty} P\left(|t(\vec{x}) - \theta^*| \geq \epsilon\right) = 0$, where n refers to the number of trials involved in a sequence of outcomes $\vec{x}$.

[5] Note the direction, knowing the true value θ^* allows us to deduce the chances with which the estimates occur, and *not* the other way around. A sampling distribution does *not* give us the probabilities of θ given an observed estimate.

show how the sampling distribution can be used to pre-experimentally deduce the number of trials n for a pre-specified level of accuracy.

Example 1 (Determining the Number of Trials for the Sample Mean within the Bernoulli Model, Continued)**.** *We wish to determine the number of trials n for an experiment $\vec{X}$ with $X \sim Ber(\theta)$ such that the chance of obtaining an estimate that is more than 0.1 distance away from the true value is no larger than, say, $\alpha = 0.25$, regardless of the exact true value of θ. This means that we require n such that $P\left(\bar{X} \in [\theta^* \pm 0.1]\right) \geq 0.75$ for every possible true value θ^*. To translate this targeted level of accuracy into a sample size n we use properties of the sampling distribution of $\bar{X}$ as follows:*

From consistency we know that the sampling distribution will increasingly concentrated its estimates around the true value as n grows, since $Var(\bar{X}) = \frac{Var(X)}{n}$. Hence, when the single-trial variance $Var(X)$ is already low to begin with we require fewer trials to the target level of accuracy. For instance, when $\theta^* = 0.99$, almost all outcomes will be 1 and the resulting estimate $\bar{x}$ is also approximately 1 which is close to the true value.

On the other hand, when the single-trial variance is large, we require more trials to satisfy the aforementioned requirement. The goal is, thus, to tame the variance of the estimator and we should therefore set out to minimize the largest variation $Var(X) = \theta^*(1-\theta^*)$, which occurs at $\theta^* = 0.5$. By plugging in $\theta = 0.5$ in the sampling distribution Eq. (4) we deduce that we require the participant to sit through at least $n = 25$ trials for the specified requirement to hold.

In other words, suppose the true value is $\theta^* = 0.5$ and we were to replicate the experiment $\vec{X}$ say, $m = 100$ times yielding as many estimates $\bar{x}_1, \ldots, \bar{x}_m$. Further suppose that each replication experiment consists of $n = 25$ trials. Then we expect that 75 of the 100 replicated estimates will fall between 0.4 and 0.6, satisfying the requirement $P\left(\bar{X} \in [\theta^* \pm 0.1]\right) \geq 0.75$. ♢

Intermediate Conclusion In the above example we introduced concepts such as a "true value" θ^* for the parameter, consistency, and the sampling distribution of an estimator and showed how these concepts can be used in a pre-experimental analysis. In particular, we

showed that a specific estimator, $\bar{X}$, is consistent for θ, by studying the mean and the variance of this estimator. Furthermore, we showed how consistency of the mean estimator can be exploited to determine the number of trials n for an experiment at a pre-specified level of accuracy by controlling the largest variance at $\theta^* = 0.5$. Note that $\text{Var}(\bar{X}) \overset{Eq.\ (10)}{=} \frac{1}{nI(\theta)}$ and this worst-case scenario was therefore already hinted at in Fig. 4, as the data are then expected to be least informative. In the remainder of the section we further explore the connection of frequentist estimation theory and Fisher information by comparing two estimators within the normal model.

Evaluating the Performance of an Estimator Within More Complex Models **Step (i): Choose an estimator** In more complex models, there are often several viable estimators for a parameter of interest. For instance, if the data are modeled according to a normal distribution, i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$, we can then use both the mean $\bar{X}$ and median estimator M for μ. This is due to the fact that μ is not only the population mean, but also the population median. Furthermore, estimates of the mean and the median estimator will typically differ. From a design point of view, we prefer the estimator that requires the least number of samples to distill a certain amount of information about μ. **Step (ii)+(iii): Assessing consistency and deriving the sampling distributions** First we show that within the normal model both the $\bar{X}$ and M are consistent for μ by considering the means and variances of these estimators.

Example 2 (Mean and Variances of the Sample Mean and Sample Median within the Normal Model, Continued)**.** *When $X \sim \mathcal{N}(\mu, \sigma^2)$ the mean estimator $\bar{X}$ has mean $E(\bar{X}) = \mu^*$ and variance $Var(\bar{X}) = \frac{\sigma^{*2}}{n}$, whatever the true values of μ and σ^2. This can be shown as in Eq.* (16) *and Eq.* (18)*. Moreover, a sum of normals is also normal, from which it follows that the sampling distribution of the mean estimator is also normal, $\bar{X} \sim \mathcal{N}(\mu*, \frac{\sigma^{*2}}{n})$, this is a particular special case.*

More commonly, however, we cannot derive the sampling distribution of an estimator exactly, that is, for finite n and we then resort to so-called large-sample approximations. In particular, instead of deriving the sampling of the median estimator exactly we call upon

Corollary 21.5 from van der Vaart (1998), which tells us that the asymptotic sampling distribution of median estimator is also normal, that is, $M \rightsquigarrow \mathcal{N}\left(\mu^*, \frac{2\pi\sigma^{*2}}{4n}\right)$*, where the symbol* $\rightsquigarrow$ *conveys that this equality only holds for large enough* n*. Note that the variances of both estimators* $\bar{X}, M$ *tend to zero with the number of trials* n *increases.* $\Diamond$

Step (iv): Derive n Thus, both $\bar{X}$ and M are consistent estimators for μ, and we can, therefore, determine, for each of these estimators, the number of trials that are required to reach a targeted level of accuracy.

Example 2 (Determining the Number of Trials n within the Normal Model When the Variance is Known, Continued)**.** *We wish to determine the number of trials* n *for an experiment* $\vec{X}$ *with* $X \sim \mathcal{N}(\mu, 1)$ *such that the chance of obtaining an estimate that is less than* 0.2 *distance away from the true value* μ^* *is* $1 - \alpha = 95.8\%$*, regardless of the exact true value of* μ*.*

To this end we exploit the fact that the estimators have an (asymptotically) normal sampling distribution. More specifically, we know that a probability of 95.8% *of a normal distribution corresponds to two standard deviations of* $\bar{X}$ *away from the mean of the estimator, the true value* μ^**. Hence, we therefore translate the requirement* $P(\bar{X} \in [\mu^* \pm 0.2]) \leq 0.042$ *into* $Var(\bar{X}) \leq 0.01$*. By solving* $Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{1}{n} = 0.01$ *we see that we require at least* $n = 100$ *trials if we choose the mean estimator* $\bar{X}$ *to estimate* μ*.*

Similarly, as the asymptotic *distribution of the median estimator* M *is also normal, we use the same translation of the targeted accuracy* $1 - \alpha = 0.958$ *in terms of the quantiles of a normal distribution with an asymptotic variance as derived in the previous example. Solving* $Var(M) \rightsquigarrow \frac{2\pi}{4n} = 0.01$*, we deduce that we require at least* $n = 157$ *trials to estimate* μ *whenever we use* M*. Hence, we therefore prefer the mean estimator* $\bar{X}$ *over the median estimator* M*: the mean estimator* $\bar{X}$ *is able to extract more information about* μ *from* $n = 100$ *trials than the median estimator.* $\Diamond$

Note that we approximated the exact sampling distribution of the median estimator by its asymptotic version. This simplification is applied commonly to most estimators as the exact sampling distribution is typically intractable. In this example, we replaced the exact sampling distribution of M by a normal distribution that is fully specified by its mean and

variance, i.e., $M \rightsquigarrow \mathcal{N}\left(\mu^*, \frac{2\pi}{4n}\right)$. Subsequently, we translated the targeted accuracy of the estimator into the quantiles of this normal sampling distribution, $\mathcal{N}\left(\mu^*, \frac{2\pi}{4n}\right)$, to calculate the required sample size n.

Hence, this approximation of n can be applied to any estimator that has an asymptotic normal sampling distribution for which we have to identify the corresponding asymptotic mean and variance. In particular, this holds true for the MLE as shown below. Recall that $\bar{X}$ is also the MLE for μ whenever the data are modeled according to a normal distribution.

Computational Advantage of the MLE In general, the MLE $\hat{\theta}$ for θ within regular models[6] is: **Step (ii)** consistent and **Step (iii)** has an asymptotically normal sampling distribution with an asymptotic mean given by the true value of the parameter, thus, it is "aimed" at the right value and with an asymptotic variance given by the inverse Fisher information, i.e.,

$$\hat{\theta} \rightsquigarrow \mathcal{N}\left(\theta^*, \frac{I^{-1}(\theta^*)}{n}\right) \text{ or equivalently, } \hat{\theta} \rightsquigarrow \theta^* + \epsilon \text{ with } \epsilon \sim \mathcal{N}\left(0, \frac{I^{-1}(\theta^*)}{n}\right), \tag{19}$$

where $I^{-1}(\theta^*)$ is the inverse unit Fisher information at the true value θ^* of the parameter that governs the possible outcomes. The latter formulation of Eq. (19) can be thought of as a linear regression in which the MLE regresses to the true value of the parameter with an error determined by the inverse Fisher information. In particular, note that this error disappears when the sample size n grows indefinitely, which confirms the claim that the MLE is consistent.

Note that Eq. (19) is not that surprising for $\bar{X}$ as the central limit theorem states that the asymptotic sampling distribution of the mean estimator converges to a normal distribution $\bar{X} \rightsquigarrow \mathcal{N}\left(E(X), \frac{\mathrm{Var}(X)}{n}\right)$, where $E(X)$ and $\mathrm{Var}(X)$ refer to the population mean and the population variance respectively. The statement of the central limit theorem, however, is only restricted to mean estimators, while Eq. (19) also generalizes to other estimators as well:

[6]Regular models are models $f(x \mid \theta)$ that depend smoothly on the parameter θ, such that the Fisher information is non-zero and bounded for every θ not on the boundary, see van der Vaart (1998, pp. 61-64).

Example 2 (The MLE for σ^2 within the Normal Model, Continued)**.** *In many cases, we have to infer both the population mean and variance. If we choose the MLE for* σ^2*, that is* $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ *then Eq.* (19) *tells us that* $\hat{\sigma}^2 \rightsquigarrow \mathcal{N}(\sigma^{2*}, \frac{2\sigma^{4*}}{n})$*, regardless of the true value* σ^{2*} *of* σ^2*.*

In the particular case that the data are modeled according to a normal distribution, $X \sim \mathcal{N}(\mu, \sigma^2)$*, we can actually derive the sampling distributions of* $\hat{\sigma}^2$ *exactly and compare it to the approximation given in Eq.* (19)*. To do so, we use the fact that the sums of squares divided by* σ^2 *have a* χ^2*-distribution. As a result of this calculation we conclude that* $E\left(\hat{\sigma}^2\right) = \left(1 - \frac{1}{n}\right)\sigma^{2*}$ *and that* $Var\left(\hat{\sigma}^2\right) = \left(1 - \frac{1}{n}\right)\frac{2}{\sigma^{4*}n} = (1 - \frac{1}{n})\frac{I^{-1}(\sigma^{2*})}{n}$*, regardless of the true values of* μ *and* σ^2*.*

Note that the MLE is biased, i.e., $E\left(\hat{\sigma}^2\right) = \left(1 - \frac{1}{n}\right)\sigma^{2*}$ *and that this bias of* $\frac{1}{n}\sigma^{2*}$ *can be made arbitrarily small regardless of the true value of* σ^2 *by choosing* n *large enough. Similarly, by choosing* n *large enough we can get the variance of the MLE* $Var(\hat{\sigma}^2)$ *close to the inverse Fisher information. Hence, this confirms that the inverse Fisher information is indeed a quite accurate approximation of the exact variance of the MLE.* ◊

This last example allows us to view Eq. (19) as an extension of the central limit theorem to the MLE with an asymptotical mean and variance given by the true value of the parameter and the inverse Fisher information, respectively. In the example above Eq. (19) we also saw the merits of choosing the estimator with the lower asymptotic variance as this lower variance leads to smaller sample size n we require a participant to sit through.

Estimator Selection by Fisher Information Apart from being computationally convenient, the use of the MLE as an estimator for the parameter can also be theoretically justified by the Hájek-LeCam convolution theorem (Hájek, 1970; Inagaki, 1970).[7], which guarantees that the sample size n derived from the MLE is (asymptotic) minimal.

The Hájek-LeCam convolution theorem states that whenever we choose to model the relationship $f(x \mid \theta)$ between the data and the latent property θ such that the Fisher information exists for every possible true value θ^* (not on the boundary), then every regular

[7]See Ghosh (1985) for a beautiful review in which this theorem is related to the more popular, but also more restricted Cramér-Rao information lower bound derived by Rao (1945), Cramér (1946), and Fréchet (1943).

estimator[8], such as the median estimator M within the normal model and the MLE, can be decomposed into an independent sum of an arbitrary random variable W and a normally distributed random variable ϵ that has a variance given by the inverse Fisher information. That is, any regular estimator T can be written as:

$$T \rightsquigarrow W + \epsilon, \text{ where } \epsilon \sim \mathcal{N}\left(0, \frac{I^{-1}(\theta^*)}{n}\right). \tag{20}$$

when n is sufficiently large. Due to the independence of W and ϵ, this implies that the asymptotic variance of any regular estimator T is given by $\text{Var}(T) = \text{Var}(W) + \text{Var}(\epsilon) = \text{Var}(W) + \frac{I^{-1}(\theta^*)}{n}$. Furthermore, as the variance of W cannot be negative we know that the asymptotic variance of any regular estimator is therefore *at least* the inverse Fisher information at the true value of θ. Clearly, this is the case for the MLE, Eq. (19), as the true value θ^* takes on the role of W, which does not vary within the frequentists' framework, $\text{Var}(W) = 0$. Hence, of all regular estimators the MLE has the lowest asymptotical variance and the sample size determined based its sampling distribution is therefore the lowest. This means that the MLE is expected to be (asymptotically) efficient in extracting all the information about the parameter of interest from the anticipated data.

Summary of Section 1 We discussed how the MLE in conjunction with the Fisher information can be used to determine the number of trials n that are to be presented to a participant in order to estimate the latent properties θ from observed behavior $\vec{x}$ within a statistical model $f(\vec{x} \,|\, \vec{\theta})$ given a targeted level of accuracy $1 - \alpha$.

This method of sample size determination was justified by the Hájek-LeCam convolution theorem, which states that the resulting sample size is (asymptotically) minimal, if we choose to model the relationship $f(x \,|\, \theta)$ between the data and the latent property θ smoothly, that is, whenever the Fisher information is non-zero and finite for every possible true value θ^* that does not lie on the boundary of the parameter space.

[8] Regular estimators are estimators that have a limiting distributions that depends smoothly on the parameters, see Bickel, Klaassen, Ritov, and Wellner (1993) for a challenging, almost agonizing account. Furthermore, see Beran (1995) for a discussion of the convolution theorem and the bootstrap.

2. Using Fisher Information to Define a Default Prior According to Jeffreys' Rule

Introduction In this section we discuss the role Fisher information plays in Bayesian statistics. The Bayesian estimation procedure proceeds as follows: **Step 0** Propose a model that contains a parameter θ. For instance, one could choose the Bernoulli model in Example 1 which serves as the running example for the next two steps.

Step 1 Extend the model by assigning the parameter a distribution, denoted $g(\theta)$, prior to seeing the data. We can use $g(\theta)$ to incorporate previous findings about θ. Alternatively, we can use $g(\theta)$ to incorporate expert knowledge, and we can refer to $g(\theta)$ as the expert's prior beliefs. This implies that we consider θ as an additional random variable for which we can make probabilistic statements. **Step 2** Employ Bayes' theorem (see Eq. (21) below) to update knowledge about the parameter θ in light of the observed data. Bayes' theorem indicates how data should modify prior beliefs about θ to yield a posterior belief, measured by a posterior distribution $g(\theta \mid \vec{X} = \vec{x})$, schematically: $\theta \overset{g(\theta \mid \vec{X}=\vec{x})}{\Longleftarrow} \vec{x}$. If needed, the posterior distribution can be summarized by a point estimate and a credible interval, which reflects the uncertainty of the estimate.

In sum, the goal of coherently updating information using the observed data about the parameter of interest reduces to an application of Bayes' theorem. To this end, we first have to extend the model with a prior distribution $g(\theta)$ which might be unavailable when we venture into novel research. To nonetheless take advantage of the Bayesian machinery, one might be tempted to construct a "flat" or uniform prior distribution that assigns equal probability to each parameter value.

We start this section with the adoption of this flat prior to elaborate on the role of Bayes' theorem in **Step 2**. We then show the flaws of flat priors as they lead to conclusions that depend on the arbitrary way in which the problem is represented. To overcome this limitation we can use Jeffreys' rule to specify priors that are translation invariant Jeffreys (1946). The concept of Fisher information forms a key component of Jeffreys' rule.

From Prior to Posterior: The Principle of Insufficient Reason and Bayes' Theorem When there is a complete lack of knowledge about the parameter of interest it seems intuitively sensible to expand the model $f(\vec{x}\,|\,\theta)$ with a prior that assigns equal probability to every possible value of θ, a rule known as Laplace's (1749 – 1827) "principle of insufficient reason" (Laplace, 1986; Stigler, 1986). **Step 1** In particular, this suggests extending data that are modeled according to a Bernoulli distribution Eq. (2) with a uniform prior for the parameter of interest $\theta \sim U[0,1]$. **Step 2** Equipped with this prior, we can now draw conclusions about θ conditioned on observed data $\vec{x}$ using Bayes' theorem:

$$g(\theta\,|\,\vec{X}=\vec{x}) = \frac{f(\vec{x}\,|\,\theta)g(\theta)}{\int_\Theta f(\vec{x}\,|\,\theta)g(\theta)\,\mathrm{d}\theta}, \tag{21}$$

where $g(\theta)$ is the prior density function for the parameter θ. The left-hand side of Eq. (21) is known as the posterior density of θ, and Eq. (21) is often verbalized as:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}. \tag{22}$$

Note that the marginal likelihood does not involve the parameter θ, and that it is given by a single number that ensures that the area under the posterior distribution equals one, see the right panel in Fig. 5. It is common practice, therefore, to write $g(\theta\,|\,\vec{X}=\vec{x}) \propto f(\vec{x}\,|\,\theta)g(\theta)$, which says that the posterior is proportional to the likelihood times the prior. Furthermore, note that the posterior distribution is a combination of what we knew before we saw the data (i.e., the information in the prior distribution), and what we have learned from the data in terms of likelihood.

Example 1 (Uniform prior on θ, Continued)**.** *We assume that no prior research has been conducted for the animal discrimination task and set* $g(\theta) = 1$ *for every possible value of the ability parameter* θ *in* $[0,1]$ *in accordance with Laplace's principle. We then apply the following scheme*

$$\textit{Prior for } \theta \xrightarrow{\textit{Data}} \textit{Posterior for } \theta \tag{23}$$

to update our knowledge of θ*. Fig. 5 illustrates this updating scheme with data characterized by* $y = 7$ *successes out of* $n = 10$ *trials.*

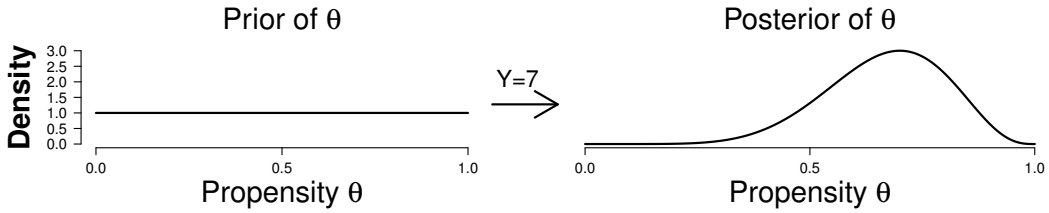

Figure 5. Bayesian updating based on observing $y = 7$ successes out of $n = 10$ Bernoulli trials. In the left panel, the flat prior distribution assigns equal probability to every possible value of θ. In the right panel, the posterior distribution is a compromise between the prior and the observed data (i.e., $y = 7$ and $n = 10$).

In this case, we have an analytically tractable posterior given by a beta distribution $g(\theta \,|\, y = 7) \sim \textit{Beta}(8, 11)$ *(e.g., Bayes & Price, 1763; Diaconis & Ylvisaker, 1979). If desired, the posterior distribution for a participant's latent animal discrimination ability* θ *can be summarized by a point estimate and a so-called credible interval. A commonly used point estimate is the posterior median, which is given by* 0.68 *and we can accompany it with a 95% credible interval that ranges from* 0.39 *to* 0.89 *to quantify our uncertainty about this estimate. Note the similarities between the posterior distribution in Fig. 5 and the likelihood function in Fig. 2; the maximum likelihood is in fact the mode of this posterior.* ♢

In more complicated data models, we might not be able to derive a posterior distribution analytically, as the integration involved in the marginal likelihood is typically hard to perform. However, due to the advent of computer-driven sampling methodology generally known as Markov chain Monte Carlo (MCMC: e.g., Gamerman & Lopes, 2006; Gilks, Richardson, & Spiegelhalter, 1996), we can directly sample sequences of values from the posterior distribution of interest, foregoing the need for closed-form analytic solutions (e.g., Lee & Wagenmakers, 2013). In the following, we omit the derivations of the posteriors and instead provided R code

on the first author's website.

The Principle of Insufficient Reason and its Dependence on the Parametrization

Step 1 Laplace's principle seems to yield priors that reflect complete ignorance about the parameter θ. In spite of its intuitive formulation, this is not true, as it depends on how a statistical problem is represented. To demonstrate this paradoxical result we exploit the similarity the Bernoulli model has to a coin flip experiment with a bent coin that has propensity $\theta = P(x = 1)$ to land heads instead of tails. Furthermore, we say that this propensity is due to the angle, denoted as ϕ, unbeknownst to the researcher, with which the coin is bent.

Moreover, suppose that the coin's physical relation between the angle ϕ and the propensity θ is given by the hypothetical function $h(\phi) = \frac{1}{2} + \frac{1}{2}\left(\frac{\phi}{\pi}\right)^3$.[9] Thus, the generative model can be modified in terms of ϕ by the following chain: $\phi \stackrel{h}{\Longrightarrow} \theta \stackrel{f(\vec{x}\,|\,\theta)}{\Longrightarrow} \vec{x}$. In the absence of prior knowledge we could therefore model our ignorance about the angles ϕ according to a uniform prior on $(-\pi, \pi)$, thereby extending the Bernoulli model with a prior density given by $g(\phi) = \frac{1}{2\pi}$ for every possible value ϕ of the angle. **Step 2** This uniform prior on the angles of ϕ then induces a prior on θ for which we can subsequently apply Bayes' theorem, as schematically described by Eq. (24):

$$\begin{array}{ccc} \text{Prior on } \phi & & \\ h \downarrow & & \\ \text{Prior on } \theta \text{ induced from } \phi & \stackrel{\text{Data}}{\longrightarrow} & \text{Posterior for } \theta \text{ induced from } \phi \end{array} \tag{24}$$

The results of this chain of induction is plotted in Fig. 6.

The resulting posterior can also be obtained by first updating the uniform prior on ϕ to a posterior on ϕ, which is subsequently transformed into a posterior for θ. Either way, the resulting posterior on θ will differ substantially from the previous result plotted in the right panel of Fig. 5. Hence, the principle that was set out to reflect ignorance depends greatly on how the problem is represented. In particular, a uniform prior on the coin's angle ϕ yields a

[9] This function was chosen purely for mathematical convenience.

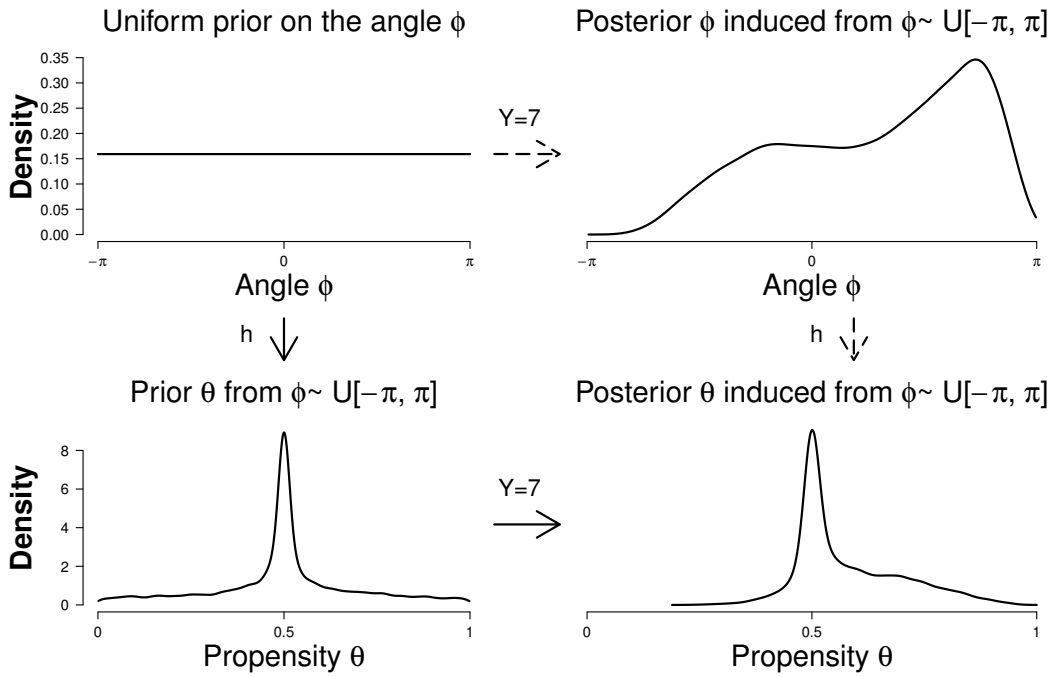

Figure 6. A uniform prior is not always uninformative: example of a bent coin landing heads in $y = 7$ out of $n = 10$ tosses. The top left panel is the starting point of our investigation with a uniform prior on the coin's angle ϕ. The corresponding bottom left panel shows that this prior distribution is highly informative for the propensity of the coin as it induces a lot of mass near $\theta = 0.5$. This is also apparent from the bottom right panel, which shows the posterior for θ with a posterior mean at 0.53 and a 95%-credible interval that ranges from 0.42 to 0.84. The same posterior on θ is obtained if we proceed via an alternative route in which we first updated the prior on ϕ to a posterior on ϕ (top left panel) and then use the transformation h to obtain the posterior on θ.

highly informative prior in terms of the coin's propensity θ. Similarly, a uniform prior on the propensity induces a prior on ϕ that assigns relatively much mass near the endpoints $-\pi$ and π.

It may not be clear, therefore, on what scale we should apply Laplace's principle of insufficient reason: one can defend a uniform prior on propensity or on angle, but neither representation seems privileged in any particular way.

Using Fisher Information to Construct a Translation Invariant Prior **Step 1** This dependence on how the problem is parameterized has been put forward as a major objection to Bayesian inference (e.g., Edwards, 1992; Fisher, 1930).

One way to meet this objection is to use a rule for constructing prior distributions that

are translation invariant, i.e., that will lead to the same conclusion (i.e., the same posterior distribution) regardless of how the problem is represented. The following rule proposes a prior that is proportional to the square root of the Fisher information:

$$g(\theta) = \frac{\sqrt{I(\theta)}}{\int \sqrt{I(\theta)}\,\mathrm{d}\theta} \propto \sqrt{I(\theta)}, \tag{25}$$

which is known as the prior obtained from Jeffreys' rule, or Jeffreys' prior (Jeffreys, 1946). We now apply this prior to the coin example both in terms of θ and ϕ to demonstrate that it is indeed translation invariant.

Example 1 (Invariance of the Jeffreys' Prior within the Bernoulli Model, Continued)**.** ***Step 2*** *To apply Jeffreys' rule, Eq.* (25)*, to the propensity θ for data distributed according to a Bernoulli distribution, recall that the unit information for θ is given by $I(\theta) = \frac{1}{\theta(1-\theta)}$, from which it follows that*

$$g(\theta) = \frac{1}{\pi}\theta^{-0.5}(1-\theta)^{-0.5}. \tag{26}$$

The proof that $\int \sqrt{I(\theta)}\,\mathrm{d}\theta = \pi$ can be found in the online appendix. Jeffreys' prior on the coin propensity θ is plotted in the bottom left panel of Fig. 7. We can then apply Bayes' theorem, using the data (i.e., $y = 7$ heads out of $n = 10$ tosses) to update this prior to a posterior on θ (bottom right panel) and subsequently use the backwards transform $k(\theta) = \phi$ to translate this result in terms of ϕ (top right panel).

A re-parametrization invariant rule implies that we could as well have started in terms of ϕ. The online appendix shows that the Fisher information for ϕ is given by $I(\phi) = \frac{9\phi^4}{\pi^6 - \phi^6}$, which yields the following prior

$$g(\phi) = \frac{1}{\pi}\frac{3\phi^2}{\sqrt{\pi^6 - \phi^6}}. \tag{27}$$

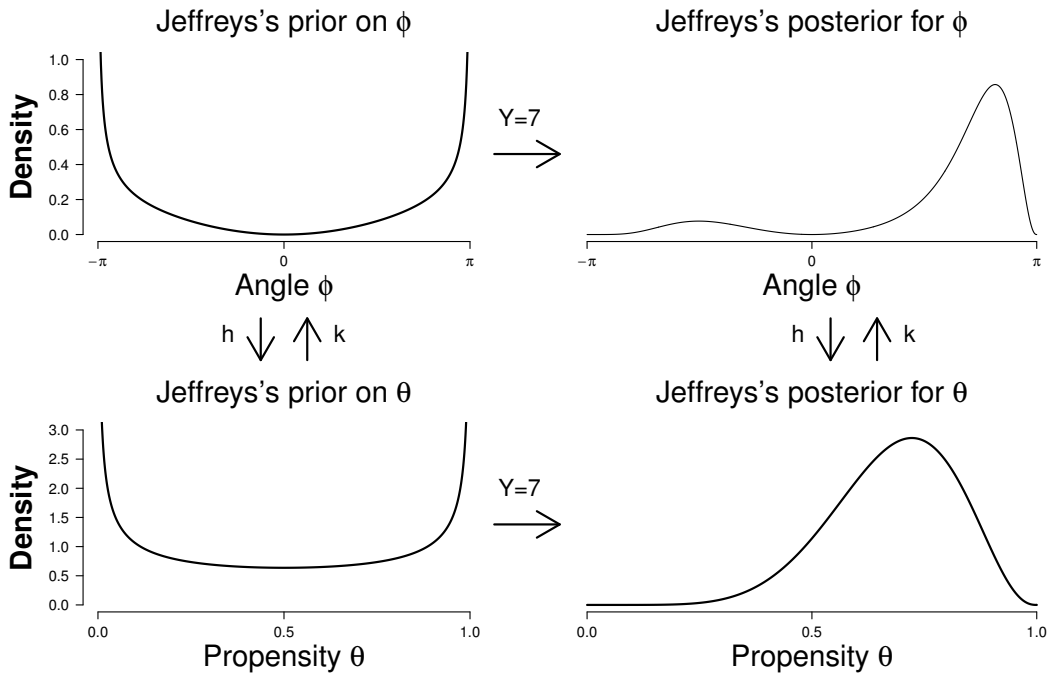

Figure 7. For priors constructed through Jeffreys' rule it does not matter whether the problem is represented in term of the coin's angles ϕ or its propensity θ. Thus, not only is the problem equivalent due to the transformations $\theta = h(\phi)$ and $\phi = k(\theta)$, the prior information is the same in both representations. This also holds for the posteriors. The posterior median for θ is given by 0.69 with a 95% credible interval of $(0.40, 0.91)$.

Note that the normalization constant for this prior is also π, a property we elaborate on in the next section. The top left panel of Fig. 7 shows the Jeffreys' prior in terms of ϕ and the top right panel shows the resulting posterior. This posterior is identical to the one obtained from the previously described updating procedure that starts by postulating Jeffreys' prior on θ instead of on ϕ. ◊

The above example shows how the Jeffreys' prior is constructed from a rule such that the same posterior knowledge is reached regardless of the scale for the prior. That is, one draws the same conclusions about θ regardless of whether we (1) use Jeffreys' rule to construct a prior on θ and update with the observed data, or (2) use Jeffreys' rule to construct a prior on ϕ, update to a posterior distribution on ϕ, and then transform that posterior in terms of θ. In contrast, a flat prior on ϕ induces a posterior distribution on θ that differs substantially from the one obtained by assigning a flat prior on θ (i.e, compare the bottom right panel of Fig. 6 to Fig. 5).

Summary of Section 2 In this section we showed how a model $f(\vec{x}\,|\,\theta)$ can be extended with default priors constructed according to Jeffreys' rule, a rule that uses Fisher information to define default priors that are translation invariant in the sense that they do not depend on the arbitrary scale on which the problem is defined.

One of the drawbacks of Jeffreys' rule is that it does not apply to multiple parameters simultaneously, for instance, when $\vec{\theta} = \binom{\mu}{\sigma^2}$ within the normal model. Jeffreys (1961) suggests to apply this rule to each parameter individually, which he justified due to the fact that the Fisher information has zeros in the off-diagonal entries, see Eq. (15) and the discussion about orthogonality below. Another drawback is that Jeffreys' rule may lead to improper priors (i.e., priors which are non-normalized). For instance, the Jeffreys' prior for μ within a normal model is uniform over the real line as μ does not depend on σ. Surprisingly, however, this prior does lead to proper normalized posteriors.

3. Using Fisher Information to Measure Model Complexity

Introduction In this section we show how Fisher information can be used to measure model complexity from the perspective of data compression and the principle of minimum description length (Grünwald, 2007; Myung, Forster, & Browne, 2000; Pitt, Myung, & Zhang, 2002).

Different theories are implemented as different statistical models, and model selection therefore helps to quantify the extent to which the data support one theory over another. However, the support in the data cannot be assessed by solely considering goodness-of-fit, as the ability to fit random data increases with model complexity, see (e.g., Roberts & Pashler, 2000).

For this reason, every method for model selection needs to take into account the trade-off between goodness-of-fit and parsimony: more complicated models necessarily lead to better fits but may in fact over-fit the data. Such overly complicated models capture idiosyncratic noise rather than general structure, resulting in poor model generalizability (Myung, Forster, & Browne, 2000; Wagenmakers & Waldorp, 2006). The goal of most model selection methods, therefore, is to acknowledge the trade-off between goodness-of-fit and parsimony in some

principled way, allowing the selection of models that best predict new data coming from the same source.

Most popular amongst the many model selection methods are the penalized maximum likelihood criteria, including the Akaike information criterion (AIC; Akaike, 1974; Burnham & Anderson, 2002), the Bayesian information criterion (BIC; Raftery, 1995; Schwarz, 1978), and the Fisher information approximation based on the minimum description length principle (FIA and MDL; Rissanen, 1996; Grünwald, 2007; Pitt et al., 2002). All these methods have a single component that quantifies goodness-of-fit (i.e., through maximum likelihood) and one or more penalty components that discount the goodness-of-fit by the degree to which it was accomplished using a complex model. The methods differ in the way that they quantify the penalty term for complexity. Specifically, AIC, BIC, and FIA are formalized as follows:

$$\text{AIC} = -2\log f(\vec{x}\,|\,\hat{\theta}) + 2d \tag{28}$$

$$\text{BIC} = -2\log f(\vec{x}\,|\,\hat{\theta}) + d\log(n) \tag{29}$$

$$\text{FIA} = \underbrace{-\log f(\vec{x}\,|\,\hat{\theta})}_{\text{Goodness-of-fit}} + \underbrace{\frac{d}{2}\log\frac{n}{2\pi}}_{\text{Dimensionality}} + \underbrace{\log\left(\int_{\Theta}\sqrt{\det I(\theta)}\,\mathrm{d}\theta\right)}_{\text{Geometric complexity}}, \tag{30}$$

where d refers to the number of free parameters, n refers to sample size, and $\hat{\theta}$ refers to the MLE. For all criteria, the model with the lowest criteria is best. Note that, in contrast to AIC and BIC, the FIA penalizes both for the number of models parameters and for their functional form. The latter is accomplished through the geometric complexity term, a term that is computed using the Fisher information (Myung, Balasubramanian, & Pitt, 2000).

The goal of this section is to visualize how Fisher information can be used to assess geometric complexity. These ideas are based on Kass (1989) and illustrated with a set of simple multinomial processing tree (MPT) models (e.g., Batchelder & Riefer, 1999; Wu, Myung, & Batchelder, 2010; Klauer & Kellen, 2011). For details about the philosophy of data compression as a method for model selection we refer the interested reader to the books by Grünwald, Myung, and Pitt (2005) and Grünwald (2007).

Running Example with Three Outcomes To demonstrate the role of Fisher information for model selection we use the following source-memory task: during an initial study phase, participants are presented with two list of words on a computer screen. List L is projected on the left-hand side and list R is projected on the right-hand side. In a later test phase, the participant is presented with two words, side by side, that can stem from either list, ll, lr, rl, rr. The participant is asked to categorize these pairs as follows:

1. Both words come from list L, i.e., ll,
2. The words are mixed M, i.e., lr or rl,
3. Both words come from list R, i.e., rr.

As before we assume that the participant is presented with n test pairs of equal difficulty, yielding the trial sequence $\vec{X} = (X_1, \ldots, X_n)$ consisting of n i.i.d. copies of a prototypical trial X that has three outcomes. Below we propose three process models for X and show how MDL model selection using Fisher information can be used to measure the complexity of each model.

General Model for a Random Variable with Three Outcomes As a starting point, we discuss a general model for random variables X with three outcomes. This model, shown in the top left panel of Fig. 8, assumes that the participant categorizes words as pairs in a cascaded fashion: with probability a, the participant concludes that the words are mixed. With probability $1 - a$, both words are perceived to stem from the same list, and the participant subsequently decides whether the two words come from list L (with probability b) or list R (with probability $1 - b$).

The model parameters naturally depend on the nature of the stimulus; for mixed pairs, for instance, parameter a will be higher than for pure pairs. For the purposes of this tutorial we wish to keep the models as simple as possible, and hence we assume that the presented word pair stimulus subject to modeling is always "rr".

In particular, when the participant's parameters are given by $a = \frac{1}{3}, b = \frac{1}{2}$, then the model predicts that $P(X = L) = P(X = M) = P(X = R) = \frac{1}{3}$ meaning that the participant

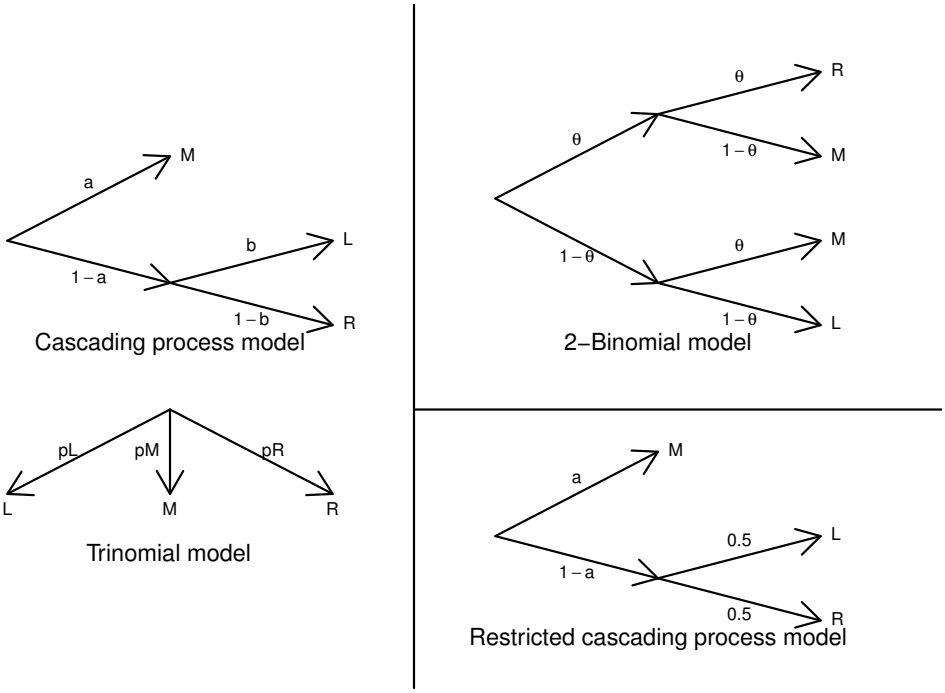

Figure 8. Three MPT models that describe a participant's choices (i.e., L, M, or R) in the source-memory task described in the main text. The left panel shows two equivalent models: the cascading process model and the trinomial model. The top right panel shows the 2-binomial model and the bottom right panel shows a restricted version of the cascading process model with b fixed at $b = 0.5$. Note: in order to limit the number of model parameter, we assume that the presented word pair stimulus subject to modeling is always "rr".

is to respond L, M, R with the same chances, and we therefore expect to see a data pattern that consists of an equal number of responses L, M, and R. In general, we can model the expected data patterns directly by re-parametrizing the cascading process model in terms of $pL = P(X = L) = (1-a)b, pM = P(X = M) = a$, and $pR = P(X = R) = (1-a)(1-b)$. This model is generally referred to as a (one-trial) trinomial model, $X \sim \text{Tri}(pL, pM, pR)$, which in fact has only two free parameters, pL, pM, as $pR = 1 - pL - pM$. This latter equation allows us to depict the data patterns that this model accommodates as a surface in three dimensions, see Fig. 9. As the components of each point $\vec{p} = (pL, pM, pR)$ on the surface sum to one, we may refer to each such point as an expected data pattern of the cascading process model or a pdf[10] of the trinomial model.

A more natural representation of the trinomial model, which helps relate Fisher infor-

[10]This is clearly a probability mass function, since we are dealing with discrete outcomes. As the ideas elaborated here extend to pdfs as well, we do not distinguish the two here.

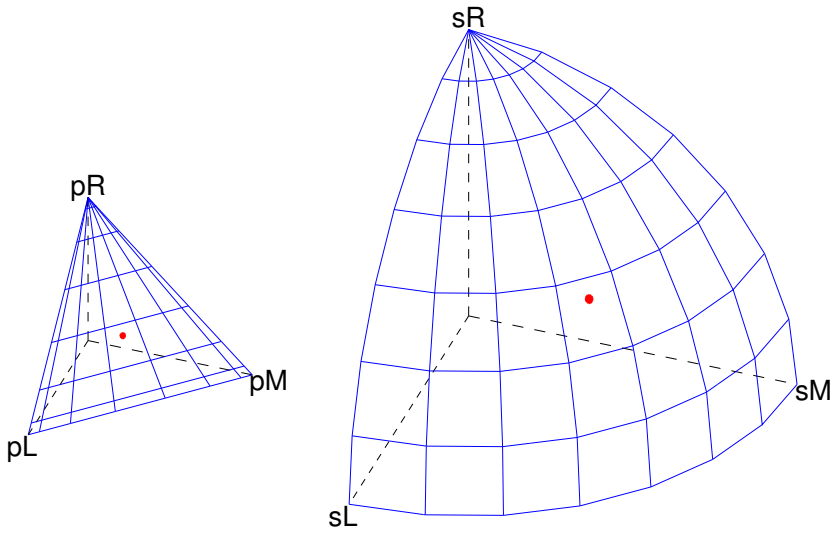

Figure 9. Expected data patterns from the trinomial model (left panel) and the transformed trinomial model (right panel). In the left panel, for instance, the left hand corner $\vec{p} = (1, 0, 0)$ represents the pdf that assigns all its mass to the L response and none to the other outcomes. The dot in the middle of the surface represents the pdf $\vec{p} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ that assigns equal probability to every outcome. Points outside the surface cannot be pdfs as they do not sum to one.

mation to model complexity, can be obtained by transforming the probabilities pL, pR, pM to two times the square roots of each probability: $sL = 2\sqrt{pL}, sM = 2\sqrt{pM}, sR = 2\sqrt{pR}$; the resulting surface is shown in the right panel of Fig. 9. Instead of forming a triangular surface, the transformed trinomial model now forms the positive octant of a sphere with radius two. In this representation all possible data patterns of the trinomial model are two units away from the origin, which can be easily verified by the Pythagorean theorem:

$$\|\vec{s}\| = \sqrt{sL^2 + sM^2 + sR^2} = \sqrt{4(pL + pM + pR)} = 2, \tag{31}$$

where we have written $\|\vec{s}\|$ for the length of an arbitrary point $\vec{s}$ on the sphere. In conclusion, the trinomial model is a general model for X and we will use this fact to visualize the predictions from the more restricted 2-binomial model defined below.

The Single-List Memorization Strategy An alternative theory for the underlying cognitive process, shown in the top right panel of Fig. 8, assumes that the participant assesses the words individually but only with respect to a single list, say list R. In other words, with probability θ the participant decides that the first word comes from the right list. This decision process is repeated for the second word, after which the complete word pair can be categorized as L, M, or R.

This model has one parameter, θ, yielding expected data patterns given by $pL = (1-\theta)^2, pM = 2\theta(1-\theta)$ and $pR = \theta^2$. These probabilities correspond to a 2-binomial model which we define by $X \sim \text{Bin}(\theta, 2)$ where a single word coming from list R is seen as a success with probability θ. For the representation on the sphere, this yields:

$$\theta \mapsto \vec{s}(\theta) = \begin{pmatrix} sL(\theta) \\ sM(\theta) \\ sR(\theta) \end{pmatrix} = \begin{pmatrix} 2\sqrt{pL} \\ 2\sqrt{pM} \\ 2\sqrt{pR} \end{pmatrix} = \begin{pmatrix} 2(1-\theta) \\ 2\sqrt{2\theta(1-\theta)} \\ 2\theta \end{pmatrix} \tag{32}$$

As a 2-binomial distribution also has three outcomes, we can represent the corresponding probabilities within the trinomial model, see Fig. 10. Note that because the 2-binomial model has only one free parameter, its data patterns form a line instead of a surface.

Expected Data Patterns and Model Specificity The 2-binomial model can be simplified even further. For instance, we might entertain the hypothesis that the participant's responses are governed by a specific value for the free parameter, such as $\theta_0 = 0.5$. From this hypothesis, we expect to see $30\,L, 60\,M$, and $30\,R$ responses in $n = 120$ trials, hence, the data pattern $\vec{p} = (0.25, 0.50, 0.25)$.

Real data, however, will typically deviate around the expected data pattern even if the hypothesis $\theta_0 = 0.5$ holds true exactly. To see this, Fig. 11 shows data patterns of 200 synthetic participants each completing 120 trials of the source-memory task with $\theta_0 = 0.5$ according to the 2-binomial model.

Fig. 11 confirms the fact that even though the *expected* data patterns from the 2-

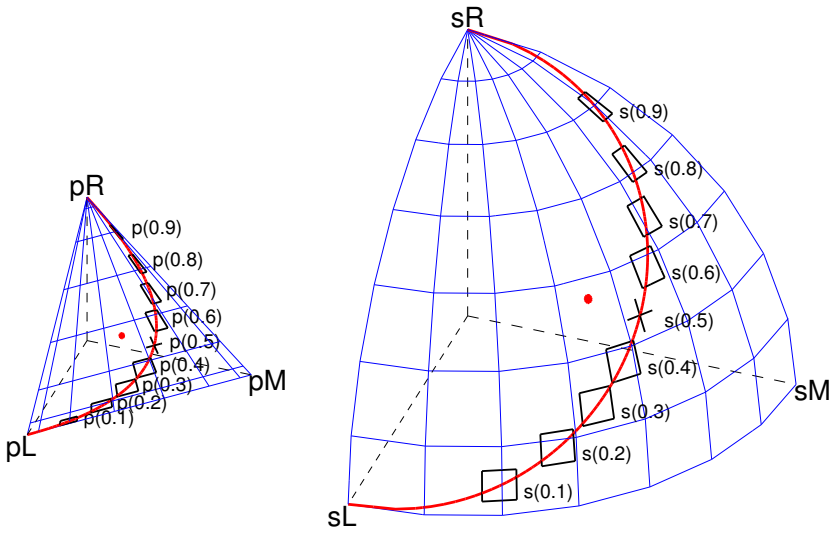

Figure 10. Expected data patterns from the 2-binomial model are represented by the curves within the trinomial model (left panel) and the transformed trinomial model (right panel). In the 2-binomial model $\theta = 0$ corresponds to a point on the surface floor, as we then expect only L responses. Data patterns corresponding to parameter values $0.1, 0.2, \ldots, 0.8, 0.9$ are represented by rectangular symbols, except for the parameter value $\theta = 0.5$ which is represented by a cross. See Eq. (32) for the definition of $\vec{s}(\theta)$.

binomial model are confined to the line, random data patterns generated from such a model are not.

To illustrate how specific the 2-binomial model is in generating data, Fig. 12 shows the results from a similar simulation using other values of the θ parameter: $\theta = 0.1, 0.2, \ldots, 0.9$.

Each dot in in Fig. 12 represents the data pattern of a synthetic participant completing 120 trials and as the number of trials increases indefinitely we will see that the observed data pattern will coincide with the expected data pattern. This confirms the relation between the size of the expected data pattern, i.e., the curve in this case, and model specificity.

To contrast the specificity of the 2-binomial model to that of the trinomial model, note that the expected data patterns for the trinomial model are not restricted to the curve. In effect, the predictions of the trinomial model cover the whole sphere. This also implies that the trinomial model is less specific in its predictions, making it harder to falsify. This is why the trinomial model is said to be more complex than the 2-binomial model.

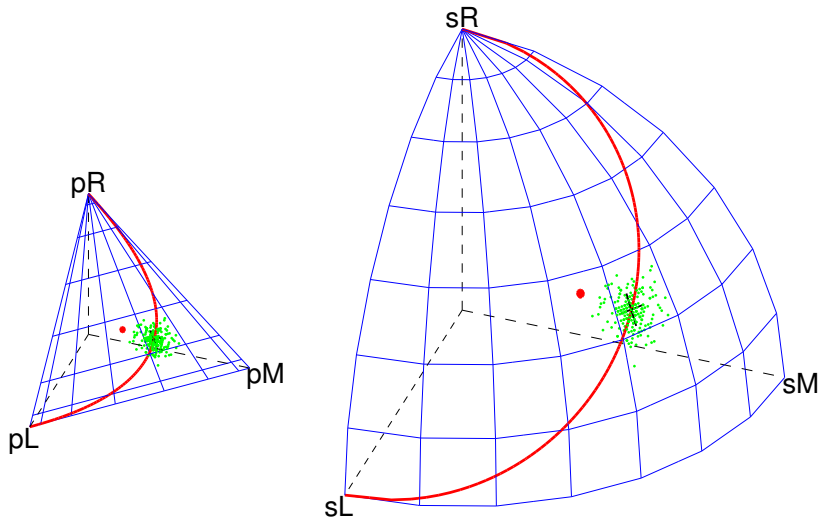

Figure 11. The curve represents the expected data patterns from the 2-binomial model and the cross represents the expected data pattern from to the restricted 2-binomial model with $\theta_0 = 0.5$. Each dot represents the data of a synthetic participant completing $n = 120$ trials.

The Trade-off between Model Specificity and Goodness-of-Fit What the trinomial model lacks in specificity is compensated by its ability to produce good fits for all data patterns with three outcomes. In contrast, when the data are simulated from the equiprobable pdf $X \sim \text{Tri}(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ we cannot yield good fits within the 2-binomial model as this pdf is not included on the line, see Fig. 10 - 12.

Thus, as the surface plot suggest, the 2-binomial model can never yield better fits than the trinomial model, as the former is nested within the latter. In other words, the maximum likelihood for the trinomial model is as least as large as the maximum likelihood for the 2-binomial model: $\log f(\vec{x} \,|\, \hat{\theta} \text{ within trinomial}) \geq \log f(\vec{x} \,|\, \hat{\theta} \text{ within 2-binomial})$. For this reason, model selection information criteria such as AIC and BIC do not consider only goodness-of-fit measures but also penalize for the number of free parameters. Within the MDL philosophy Eq. (30) such a penalty is incomplete because it ignores differences in model complexity due to the functional relationship between the parameters and the data, a relationship that can be measured using Fisher information.

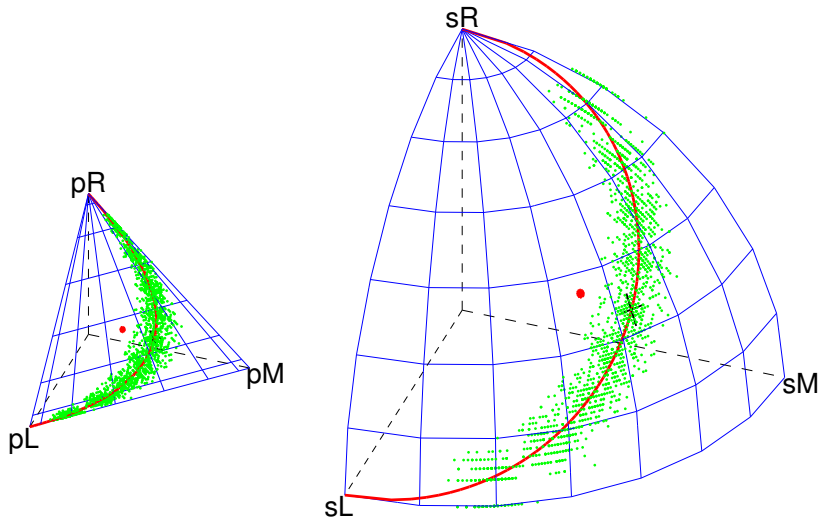

Figure 12. The curve represents the expected data patterns from the 2-binomial model and the cross represents the expected data pattern from to the restricted 2-binomial model with $\theta_0 = 0.5$. Each dot represents the data of a synthetic participant completing $n = 120$ trials. Data were simulated for each $\theta = 0.1, 0.2, \ldots, 0.9$.

Using Fisher Information to Measure the Size of a Model In the previous section we related the complexity of a model to its expected data patterns. An intuitive measure for model complexity is therefore its size. In particular, we can use the length of the curve in Fig. 10 to measure the complexity of the 2-binomial model. Before we show that this length can be calculated from the Fisher information, recall that a single trial of a 2-binomial model consists of two independent Bernoulli flips. We therefore conclude that the unit information for the 2-binomial model is given by $I(\theta \,|\, \text{2-binomial}) = \frac{2}{\theta(1-\theta)}$.

To calculate the length of the curve, we use the fact that we can linearly approximate, i.e., Taylor approximate, the parametrization from $\theta \mapsto \vec{s}(\theta)$. This strategy consists of two steps: (a) linear approximations and (b) summing the linear approximations over the domain of θ.

Step a. Linear Approximation In Fig. 13 we illustrate the first step by a linear approximation of the curve by considering three approximation points, namely: $0.05, 0.40, 0.75$.[11]

[11] For technical reasons we avoid $\theta = 0$ for now.

For each of these approximation points we calculate $\vec{s}(\theta)$ and the key idea is to approximate

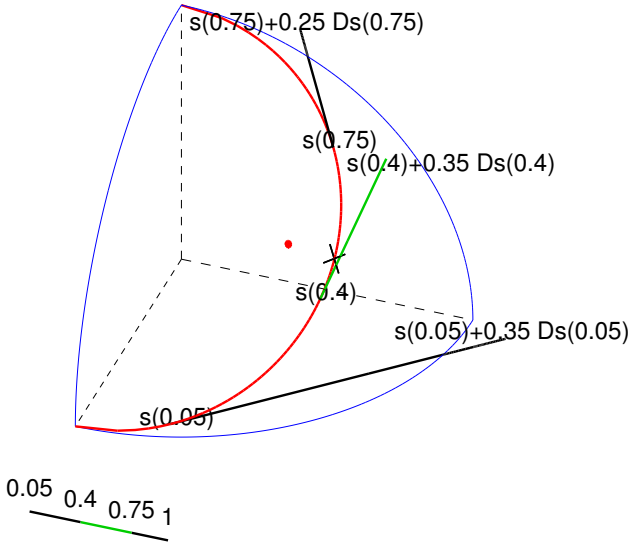

Figure 13. The size of a model can be calculated by linear approximations. The smooth curve represents the expected data patterns under the 2-binomial model. The three line segments below the bottom left corner show the partitioning of the subdomain $(0.05, 1)$ for the corresponding tangent vectors along the curve. The length of the curve can be approximated by summing the length of the three lines tangent to the curve. For clarity, we removed the grid lines that were drawn in Fig. 9, but we retained both the reference point corresponding to the data pattern $\text{Tri}(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and the plus sign for the expected data pattern for $\theta = 0.5$ within the 2-binomial model.

the "gap" between, say, $\vec{s}(0.05)$ and $\vec{s}(0.4)$ by a linear extrapolation at $\vec{s}(0.05)$. Hence, for $\theta_0 = 0.05$ this yields

$$\vec{s}(\theta_0 + h) \approx \vec{s}(\theta_0) + h\mathrm{D}\vec{s}(\theta_0), \tag{33}$$

where $D\vec{s}(\theta_0)$ denotes the 3-dimensional tangent vector at θ_0 and we used $h = 0.40 - 0.05 = 0.35$ at $\theta_0 = 0.05$ in Fig. 13. This procedure is repeated for the other points $\theta = 0.40$ and $\theta = 0.75$ and we approximate the arc length by summing the length of the tangent vectors at each point. The accuracy of this approximation of the arc length increases with the number of approximation points. For this reason we desire an expression for the tangent vector at each arbitrarily chosen point θ, which can be obtained by taking the derivative with respect to θ for each component (i.e., possible outcomes L, M, R):

$$\mathrm{D}\vec{s}(\theta) = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}\theta} sL \\ \frac{\mathrm{d}}{\mathrm{d}\theta} sM \\ \frac{\mathrm{d}}{\mathrm{d}\theta} sR \end{pmatrix} = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}\theta} \log(pL)\sqrt{pL} \\ \frac{\mathrm{d}}{\mathrm{d}\theta} \log(pM)\sqrt{pM} \\ \frac{\mathrm{d}}{\mathrm{d}\theta} \log(pR)\sqrt{pR} \end{pmatrix} = \begin{pmatrix} -2 \\ \frac{2-4\theta}{\sqrt{2\theta(1-\theta)}} \\ 2 \end{pmatrix} \tag{34}$$

where the second equality is due to the general relation between $\vec{s}(\theta)$ and $\vec{p}(\theta)$. The length of this tangent vector can be calculated by an application of the Pythagorean theorem, which yields:

$$\|\mathrm{D}\vec{s}(\theta)\| = \sqrt{(-2)^2 + \left(\frac{2-4\theta}{\sqrt{2\theta(1-\theta)}}\right)^2 + 2^2} = \sqrt{\frac{2}{\theta(1-\theta)}} = \sqrt{I(\theta \,|\, \text{2-binomial})}, \tag{35}$$

As θ was chosen arbitrarily we conclude that the tangent length at each $\theta \in (0,1)$ within the 2-binomial model is given by $\sqrt{I(\theta \,|\, \text{2-binomial})}$, the square root of the Fisher information.

Step b. Accumulating the Tangent Lengths To establish the size of a model, we now have to sum the tangent lengths over all approximation points. This sum becomes more accurate as we increase the number of approximation points, eventuating into an integral. For the 2-binomial this yields

$$V_{\text{2Bin}} = \int_0^1 \sqrt{I(\theta \,|\, \text{2-binomial})}\,\mathrm{d}\theta = \sqrt{2}\int_0^1 \frac{1}{\sqrt{\theta(1-\theta)}}\,\mathrm{d}\theta = \sqrt{2}\pi, \tag{36}$$

see the online appendix for the full calculation. Hence, the length of the curve is given by $V_{\text{2Bin}} = \int_\Theta \sqrt{I(\theta \,|\, \text{2-binomial})}\,\mathrm{d}\theta = \sqrt{2}\pi$, and we can use it as a measure for the specificity of the 2-binomial model, and, hence, its model complexity.

Fisher Information as a Measure of Complexity for Models with Vector-valued Parameters The ideas developed above can be easily generalized to models with vector-valued parameters such as the trinomial model with free parameters pL, pM or, equivalently,

the cascading process model with free parameters a, b. As there are two parameters to vary over we then have two tangent vectors which form a tangent surface. To compute the size of the model we then have to calculate the surface area of the Fisher information matrix, which is given by its determinant (for details see the online appendix). Hence,

$$V = \int \sqrt{\det\left(I(\vec{\theta})\right)}\, \mathrm{d}\vec{\theta}, \tag{37}$$

where we have written $\vec{\theta}$ for a vector-valued parameter and wrote V for the volume of the model (Pitt et al., 2002). Fig. 14 shows the tangent surface at the equiprobable pdf $\vec{p}_e = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$: on the left parameterized in terms of pL, pM and on the right parameterized in terms of a, b.

The two models are equivalent as we can re-parameterize pL, pM to a, b and vice versa. This implies that the two models accommodate the same expected data patterns, hence, the two models are equal in size. In effect, we are free to choose the parametrization that allows us to calculate the size of the trinomial model conveniently as described by Eq. (37). What qualifies as convenient can be derived from the Fisher information matrix, which is given by:

$$I(pL, pM) = \frac{1}{1 - pL - pM} \begin{pmatrix} 1 - pM & 1 \\ 1 & 1 - pL \end{pmatrix}, \tag{38}$$

in terms of the parameters pL, pM, yielding $\det\left(I(pL, pM)\right) = \frac{pLpM - pL - pM}{(1-pL-pM)^2}$. Consequently, the volume of the trinomial model is then given by

$$V_{\text{Tri}} = \int_0^1 \left(\int_0^{1-pL} \sqrt{\frac{pL \cdot pM - pL - pM}{(1 - pL - pM)^2}}\, \mathrm{d}pM \right) \mathrm{d}pL. \tag{39}$$

This integral is hard to compute as the inner integral depends on the outer integral, because the probabilities pL, pM and pR have to sum up to one. We prefer the Fisher information in the parametrization a, b as it leads to zero off-diagonal terms:

$$I(a,b) = \begin{pmatrix} \frac{1}{a(1-a)} & 0 \\ 0 & \frac{1-a}{b(1-b)} \end{pmatrix}. \tag{40}$$

Whenever the Fisher information matrix parametrized in terms of a, b has zero-off diagonals we then say that a, b are orthogonal parameters, see Fig. 14. Orthogonal parameters allow us to decouple the summing up to one restriction $pL+pR+pM = 1$ into two separate conditions $a \in (0,1)$ and $b \in (0,1)$, compare the integration bounds of Eq. (39) to those of Eq. (41). The volume of the cascading process model (thus, also of the trinomial model) is then given by

$$V_{\text{Tri}} = \int_0^1 \int_0^1 \frac{1}{\sqrt{ab(1-b)}}\,\mathrm{d}a\,\mathrm{d}b = \int_0^1 \frac{1}{\sqrt{a}}\,\mathrm{d}a \int_0^1 \frac{1}{\sqrt{b(1-b)}}\,\mathrm{d}b \overset{Eq.\ (63)}{=} \pi \int_0^1 \frac{1}{\sqrt{a}}\,\mathrm{d}a = 2\pi, \tag{41}$$

which equals an eighth of the surface area of a sphere of radius two, $\frac{1}{8}4\pi 2^2$, as one would expect. The size of the trinomial model can therefore be expressed as 2π.

An Application of Fisher Information to Model Selection To illustrate the merits of model selection using FIA over AIC and BIC, we introduce another model and compare it against the 2-binomial model. The new model is a version of the cascading process model where the parameter b is fixed to 0.5. Hence we refer to this model as the restricted cascading process model. According to the new model, participants only discriminate mixed from pure pairs of words and then randomly answer L or R. Fig. 15 shows the corresponding expected data patterns.

Note that the 2-binomial model and the restricted cascading process model are not nested and –because both models have only a single parameter– AIC and BIC discriminate the two models based on goodness-of-fit alone. Hence, when the observed data patterns lie near the right corner $P(X = M) = 1$, AIC and BIC will prefer the restricted cascading process model over the 2-binomial model. Conversely, when the observed data patterns lie near the bottom left corner $P(X = L) = 1$, AIC and BIC will prefer the 2-binomial model

over the restricted cascading process model.

Thus, AIC and BIC are unable to discriminate the 2-binomial model from the restricted cascading process model when the goodness-of-fit terms are close to each other, i.e., for data patterns near the point $\vec{p} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, where the two models intersect. Hence, whenever the likelihood and dimensions of the two models are similar, model selection by FIA then boils down to the assessment of the geometric complexity term, Eq. (30).

As noted above, the 2-binomial model has a volume term that equals $V_{\text{2Bin}} = \sqrt{2}\pi$, whereas the restricted cascading process model has a volume term given by $V_{\text{Restrict}} = \int_0^1 \frac{1}{\sqrt{a(1-a)}}\, \mathrm{d}a = \pi$ (for details see the online appendix). Thus, the restricted cascading process model is slightly less complex than the 2-binomial model, e.g., $\log(V_{\text{Restrict}}) < \log(V_{\text{2Bin}})$, and FIA will automatically prefer the simpler model whenever the goodness-of-fit is non-diagnostic (i.e., when the data are near the plus sign in Fig. 15).

Summary of Section 3 In this section we showed how Fisher information can be used to quantify model complexity by measuring the size of model in the space of expected data patterns. This sophisticated conceptualization of model complexity takes into account the functional form of model parameters to strike an appropriate balance between parsimony and goodness-of-fit.

Concluding Comments

Fisher information is a central statistical concept that is of considerable relevance for mathematical psychologists. We illustrated the use of Fisher information in three different statistical paradigms: in the frequentist paradigm, Fisher information can be used to determine the sample size required to estimate parameters at a target level of accuracy; in the Bayesian paradigm, Fisher information can be used to specify a default, translation-invariant prior distribution; finally, in the paradigm of information theory, data compression, and minimum description length, Fisher information can be used to measure model complexity.

Our goal was to use concrete examples to provide more insight about Fisher information, something that may benefit psychologists who propose, develop, and compare mathematical

models for psychological processes. Our goal was not to provide a comprehensive or complete treatment of all the uses of Fisher information throughout statistics. In our opinion, such a treatment would require a book (e.g., Frieden, 2004) rather than a tutorial article.

Other usages of Fisher information are in the detection of model misspecification, (Golden, 1995; Golden, 2000; Waldorp, Huizenga, & Grasman, 2005; Waldorp, 2009; Waldorp, Christoffels, & van de Ven, 2011; White, 1982) and in the reconciliation of frequentist and Bayesian estimation methods through the Bernstein–von Mises theorem (van der Vaart, 1998; Bickel & Kleijn, 2012). In sum, Fisher information is a key concept in statistical modeling. We hope to have provided an accessible and concrete tutorial article that explains the concept and some of its uses for applications that are of particular interest to mathematical psychologists.

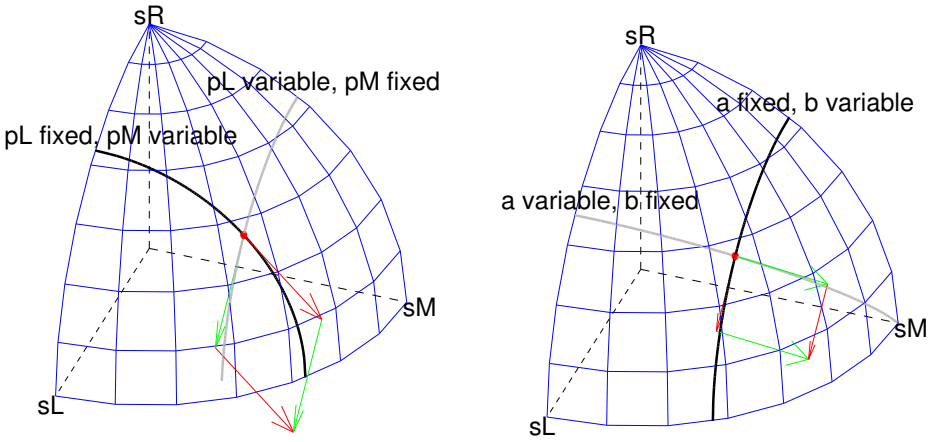

Figure 14. The size of the model is invariant under re-parametrization and we can therefore choose a parameterization which simplify the calculations. Recall that the size of an one-dimensional model, the arc length, is approximated by tangent lines, see Fig. 13. The size of a two-dimensional model on the other hand can be thought of an surface area that needs to be approximated by a tangent surface. This tangent surface can be decomposed into two tangent lines that can be derived by letting each parameter vary while the other stays fixed. The arrows in the left panel span up the tangent surface of the trinomial model with parameters pL, pM, pR at the dot that is represented by $pL = pM = pR = \frac{1}{3}$. The arrow that is directed to the left is tangent to the curve when pL is allowed to vary freely between zero and $1-pM$, while $pM = \frac{1}{3}$ is set fixed. Similarly, the arrow that is directed to the right is tangent to the curve when pM is allowed to vary freely between zero and $1 - pR$, while $pR = \frac{1}{3}$ is set fixed. Analogously, the arrows in the right panel span up the tangent surface of the cascading process models with parameters a, b at the dot that is represented by $a = \frac{1}{3}, b = \frac{1}{2}$. The arrow that is directed to the left is tangent to the curve when a is allowed to vary freely between zero and one, while $b = \frac{1}{2}$ is set fixed. Similarly, the arrow that is directed to the right is tangent to the curve when b is allowed to vary freely between zero and one, while $a = \frac{1}{2}$ is set fixed. Note that the area of this tangent surface is easier to calculate due to the fact that the arrows are orthogonal to each other, which allows us simplify the calculations see Eq. (41).

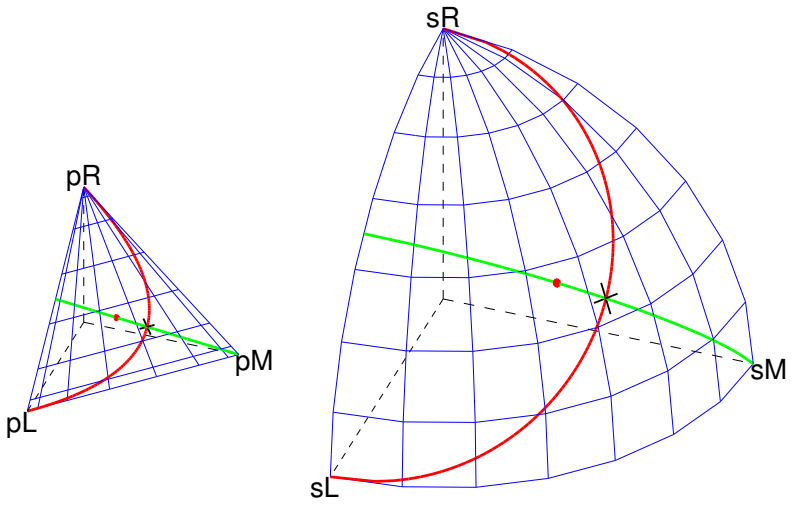

Figure 15. Expected data patterns from the 2-binomial model versus the restricted cascading process model. The latter model is represented by the line that traverses the surface starting from the bottom right corner. Note that it accommodates both the dot, i.e., the data pattern for which $pL = pR = pM = \frac{1}{3}$, and the cross that corresponds to $\theta = 0.5$ within the 2-binomial model. As before, the 2-binomial model is represented by the line that curves upwards starting from the bottom left corner.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review*, *87*, 375–397.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86.

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. by the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions (1683-1775)*, 370–418.

Beran, R. (1995). The role of Hájek's convolution theorem in statistical theory. *Kybernetika*, *31*(3), 221–237.

Bickel, P. J., Klaassen, C. A., Ritov, Y., & Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore.

Bickel, P. J., & Kleijn, B. J. K. (2012). The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, *40*(1), 206–237.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information–theoretic approach (2nd ed.)*. New York: Springer Verlag.

Chechile, R. A. (1973). *The relative storage and retrieval losses in short–term memory as a function of the similarity and amount of information processing in the interpolated task*. Unpublished doctoral dissertation, University of Pittsburgh.

Cramér, H. (1946). Methods of mathematical statistics. *Princeton University Press*, *23*.

DasGupta, A. (2011). *Probability for statistics and machine learning*. Springer.

Diaconis, P., & Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, *7*(2), 269–281.

Edwards, A. W. F. (1992). *Likelihood*. Baltimore, MD: The Johns Hopkins University Press.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, *41*, 155–160.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*, 309–368.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 22, pp. 700–725).

Fisher, R. A. (1930). Inverse probability. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 26, pp. 528–535).

Fréchet, M. (1943). Sur l'extension de certaines evaluations statistiques au cas de petits echantillons. *Revue de l'Institut International de Statistique*, 182–205.

Frieden, B. R. (2004). *Science from Fisher information: A unification.* Cambridge University Press.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference.* Boca Raton, FL: Chapman & Hall/CRC.

Ghosh, J. (1985). Efficiency of estimates–part I. *Sankhyā: The Indian Journal of Statistics, Series A*, 310–325.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* Boca Raton (FL): Chapman & Hall/CRC.

Golden, R. M. (1995). Making correct statistical inferences using the wrong probability model. *Journal of Mathematical Psychology*, *39*, 3-20.

Golden, R. M. (2000). Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology*, *44*(1), 153–170.

Grünwald, P. (2007). *The minimum description length principle.* Cambridge, MA: MIT Press.

Grünwald, P., Myung, I. J., & Pitt, M. A. (Eds.). (2005). *Advances in minimum description length: Theory and applications.* Cambridge, MA: MIT Press.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *14*(4), 323–330.

Huzurbazar, V. S. (1949). On a property of distributions admitting sufficient statistics. *Biometrika*, *36*(1-2), 71–74.

Inagaki, N. (1970). On the limiting distribution of a sequence of estimators with uniformity property. *Annals of the Institute of Statistical Mathematics*, *22*(1), 1–13.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *186*(1007), 453–461.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statistical Science*, 188–219.

Klauer, K. C., & Kellen, D. (2011). The flexibility of models of recognition memory: An analysis by the minimum-description length principle. *Journal of Mathematical Psychology*, *55*(6), 430–450.

Laplace, P. S. (1986). Memoir on the probability of the causes of events. *Statistical Science*, *1*(3), 364–378.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian modeling for cognitive science: A practical course.* Cambridge University Press.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100.

Myung, I. J., Balasubramanian, V., & Pitt, M. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*(21), 11170–11175.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection [Special issue]. *Journal of Mathematical Psychology*, *44*(1–2).

Pitt, M., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Cambridge: Blackwells.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, *37*(3), 81–91.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*, 40–47.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing in psychology. *Psychological Review*, *107*, 358–367.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Stigler, S. (1986). Laplace's 1774 memoir on inverse probability. *Statistical Science*, *1*(3), 359–363.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Wagenmakers, E.-J., & Waldorp, L. (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, *50*(2).

Waldorp, L. (2009). Robust and unbiased variance of GLM coefficients for misspecified autocorrelation and hemodynamic response models in fMRI. *International Journal of Biomedical Imaging*, *2009*, 723912.

Waldorp, L., Christoffels, I., & van de Ven, V. (2011). Effective connectivity of fMRI data using ancestral graph theory: Dealing with missing regions. *NeuroImage*, *54*(4), 2695 – 2705.

Waldorp, L., Huizenga, H., & Grasman, R. (2005). The Wald test and Cramér–Rao bound for misspecified models in electromagnetic source analysis. *IEEE Transactions on Signal Processing*, *53*(9), 3427-3435.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25.

Wu, H., Myung, I. J., & Batchelder, W. H. (2010). Minimum description length model selection of multinomial processing tree models. *Psychonomic Bulletin & Review*, *17*, 275–286.

Appendix A

Calculating the MLE and Fisher information for the normal distribution

For the normal distribution we have the following log-likelihood functions.

$$\log f(\vec{x}\,|\,\vec{\theta}) = -\frac{n}{2}\log(2\pi v) - \frac{1}{2v}\sum_{i=1}^{n}(X_i-\mu)^2, \tag{42}$$

where we have written $v = \sigma^2$ to avoid confusion of differentiation with respect to the standard deviation. To derive the maximum likelihood functions we are required to solve the likelihood equations. Hence, find that parameter value such that the following partial derivatives are zero:

$$\frac{\partial}{\partial\mu}\log f(\vec{X}\,|\,\hat{\mu}_n, v) = \frac{1}{2v}\sum_{i=1}^{n}(X_i-\mu) \tag{43}$$

$$\frac{\partial}{\partial v}\log f(\vec{X}\,|\,\hat{\mu}_n, v) = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(X_i-\mu)^2 \tag{44}$$

To calculate the Fisher information we need to calculate the second derivates, the Hessian matrix, and integrate these with respect to $\vec{X}$:

$$\frac{\partial^2}{\partial\mu\partial\mu}\log f(\vec{X}\,|\,\mu, v) = -\frac{1}{v}, \tag{45}$$

$$\frac{\partial^2}{\partial\mu\partial v}\log f(\vec{X}\,|\,\mu, v) = -\frac{1}{2v^2}\sum_{i=1}^{n}(X_i-\mu), \tag{46}$$

$$\frac{\partial^2}{\partial v\partial v}\log f(\vec{X}\,|\,\mu, v) = \frac{n}{2v^2} - \frac{1}{v^3}\sum_{i=1}^{n}(X_i-\mu)^2. \tag{47}$$

We use these equations to show:

Theorem 3 (Maximum likelihood estimators for normal distributions)**.** *If $\vec{X} = (X_1, \ldots, X_n)$ are i.i.d. with $X_1 \sim \mathcal{N}(\mu, \sigma^2)$, then the MLEs are given by $\hat{\mu}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \hat{\sigma}_n^2 = S_n =$*

$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. *Furthermore, the expected Fisher information matrix is then given by*

$$I_n(\vec{\theta}) = n \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}, \tag{48}$$

where the matrix on the right-hand side represents the unit information of a normally distributed random variable $\vec{X}$. ◊

Proof. For the MLE of μ, we set (43) to zero, which yields $\sum_{i=1}^{n} X_i) = n\mu$. Hence, $\hat{\mu}_n$ is given by the sample mean. We use this knowledge to derive the MLE for the variance v. Setting (44) to zero yields $n = \frac{1}{v}\sum_{i=1}^{n}(X_i - \mu)^2$ after we multiplied both sides by $2v$ and plugging in the MLE for μ yields S_n.

To compute the Fisher information matrix, we have to integrate the negatives of (45), (46) and (47) with respect to $\vec{X}$, see (9). First note that (45) does not depend on $\vec{X}$ so it is straightforward to see that $I(\mu) = \frac{1}{v}$. Furthermore, as (46) defines an uneven function about μ, it is as much positive as negative on each side of μ, it will integrate to zero. By i.i.d. we know that the integration on the right-hand side of (47) times -1 yields $\frac{nv}{v^3} - \frac{n}{2v^2} = n\frac{1}{2v^2}$. □

Appendix B

Calculating the mean and variance of $\hat{\sigma}_n^2$ for the normal distribution

Theorem 4. *If* $\vec{X} = (X_1, \ldots, X_n)$ *are i.i.d. with* $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ *with an MLE* $\hat{\sigma}_n^2$ *and the standard sample variance* $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, *we have:*

$$E(\hat{\sigma}_n^2) = \left(1 - \frac{1}{n}\right) E(S_{n-1}^2) \text{ and } E(S_{n-1}^2) = \sigma^2 \tag{49}$$

$$Var(\hat{\sigma}_n^2) = \left(1 - \frac{1}{n}\right)^2 Var(S_{n-1}^2) < Var(S_{n-1}^2) = \frac{2\sigma^4}{n-1} \tag{50}$$

Proof. First note that as n is known $1 - \frac{1}{n}$ can be considered as a constant and that $\hat{\sigma}_n^2 = \frac{n-1}{n}S_{n-1}^2$. This explains the factors on the left-hand side of both results. Furthermore, the result (49) follows directly from the linearity of the expectation and the fact that the sample

variance S^2_{n-1} is unbiased, $E(S^2_{n-1}) = \sigma^2$. By Basu's theorem, (Ch. 18 DasGupta, 2011) we can show that $\frac{n-1}{\sigma^2}S^2_{n-1} \sim \chi^2(n-1)$ is distributed according to a chi-squared distribution with $n-1$ degrees of freedom and thus a variance of $2(n-1)$. Hence,

$$2(n-1) = \mathrm{Var}(\frac{n-1}{\sigma^2}S^2_{n-1}) = \frac{n-1}{\sigma^2}\mathrm{Var}(S^2_{n-1}), \tag{51}$$

and the result (50) then follows directly from multiplying (51) with $\frac{\sigma^2}{n-1}$ on both sides. □

Appendix C
Angle coin

Let $P(X=1) = \theta$ and ϕ is the angle of a bent coin with respect to the horizontal axis.

$$h : \Phi \to \Theta : \phi \mapsto \frac{1}{2} + \frac{1}{2}\left(\frac{\phi}{\pi}\right)^3 \text{ with } \frac{\mathrm{d}\theta}{\mathrm{d}\phi} = \frac{3}{2\pi^3}\phi^2 \tag{52}$$

and an inverse function

$$k = h^{-1} : \Theta \to \Phi : \theta \mapsto \begin{cases} -\pi(1-2\theta)^{\frac{1}{3}} & \text{with } \frac{\mathrm{d}\phi}{\mathrm{d}\theta} = \frac{2\pi}{3(1-2\theta)^{\frac{2}{3}}} \quad \text{when } 0 \le \theta < 0.5 \\ 0 & \text{with } \frac{\mathrm{d}\phi}{\mathrm{d}\theta} = \frac{2\pi}{3(1-2\theta)^{\frac{2}{3}}} \quad \text{when } \theta = 0.5 \\ \pi(2\theta-1)^{\frac{1}{3}} & \text{with } \frac{\mathrm{d}\phi}{\mathrm{d}\theta} = \frac{2\pi}{3(2\theta-1)^{\frac{2}{3}}} \quad \text{when } 0.5 < \theta \le 1 \end{cases} \tag{53}$$

For $I(\phi)$, write

$$\log p(x \,|\, \phi) = x \log\left(\frac{1}{2} + \frac{1}{2}\left(\frac{\phi}{\pi}\right)^3\right) + (1-x)\log\left(\frac{1}{2} - \frac{1}{2}\left(\frac{\phi}{\pi}\right)^3\right) \tag{54}$$

Furthermore,

$$\frac{\mathrm{d}}{\mathrm{d}\phi}\log p(x\,|\,\phi) = \frac{3}{\pi^3}\left(x\frac{\phi^2}{1+\left(\frac{\phi}{\pi}\right)^3} - (1-x)\frac{\phi^2}{1-\left(\frac{\phi}{\pi}\right)^3}\right) \tag{55}$$

Moreover, (this is where it gets exciting)

$$\frac{\mathrm{d}^2}{\mathrm{d}\phi^2}\log p(x\,|\,\phi) = \frac{3}{\pi^3}\left(x\frac{2\phi-\frac{\phi^4}{\pi^3}}{(1+\left(\frac{\phi}{\pi}\right)^3} - (1-x)\frac{2\phi+\frac{\phi^4}{\pi^3}}{(1-\left(\frac{\phi}{\pi}\right)^3}\right) \tag{56}$$

Hence,

$$I(\phi) = \frac{9\phi^4}{\pi^6-\phi^6} \tag{57}$$

Or, based on the fact that the Fisher information in basis θ is given by $I(\theta) = \frac{1}{\theta(1-\theta)}$:

$$I(\phi) = \left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2 I(\theta(\phi)) = \left(\frac{3\phi^2}{2\pi^3}\right)^2 \frac{1}{(\frac{1}{2}+\frac{1}{2}\left(\frac{\phi}{\pi}\right)^3)(\frac{1}{2}-\frac{1}{2}\left(\frac{\phi}{\pi}\right)^3)} \tag{58}$$

To calculate the normalizing constant for Jeffreys' prior we have to integrate $\int_{-\pi}^{\pi}\sqrt{I(\phi)}\mathrm{d}\phi$.

$$\int_{-\pi}^{\pi}\sqrt{I(\phi)}\mathrm{d}\phi = \int_{-\pi}^{\pi}\sqrt{\frac{9\phi^4}{\pi^6-\phi^6}}\mathrm{d}\phi \tag{59}$$

$$= \int_{-\pi}^{\pi}\frac{3\phi^2}{\sqrt{\pi^6-\phi^6}}\frac{1}{\frac{\pi^3}{\sqrt{\pi^6}}}\mathrm{d}\phi \tag{60}$$

$$= \int_{-\pi}^{\pi}\frac{1}{\sqrt{1-\left(\frac{\phi}{\pi}\right)^6}}\frac{3\phi^2}{\pi^3}\mathrm{d}\phi. \tag{61}$$

If we substitute $s = \left(\frac{\phi}{\pi}\right)^3$ we then have $ds = \frac{3\phi^2}{\pi^3}\mathrm{d}\phi$. Furthermore, when $\phi = \pi$ this means that $s = 1$ and we get $s = -1$ as a lower bound. Hence,

$$\int_{-\pi}^{\pi} \sqrt{I(\phi)} \mathrm{d}\phi = \int_{-1}^{1} \frac{1}{\sqrt{1-(s)^2}} \mathrm{d}s, \tag{62}$$

$$= \sin^{-1}(s)\Big|_{s=-1}^{s=1} = \frac{\pi}{2} - \frac{-\pi}{2} = \pi, \tag{63}$$

where we have used a standard calculus result to derive the identity with $\sin^{-1}$. Similarly, the normalizing constant for the Jeffreys prior in θ basis is given by

$$\int_0^1 \sqrt{I(\theta)}\, \mathrm{d}\theta = \int_0^1 \frac{1}{\sqrt{\theta(1-\theta)}}\, \mathrm{d}\theta. \tag{64}$$

Substituting $\theta = \sin^2(z)$ we have $\mathrm{d}\theta = 2\sin(z)\cos(z)$ yields the result. Note that Jeffreys' prior on θ is in fact Beta$(0.5, 0.5)$ and since a beta prior leads to a beta posterior we know that the posterior for θ is then given by:

$$\frac{1}{B(7.5, 3.5)} \theta^{6.5} (1-\theta)^{2.5}. \tag{65}$$

Similarly, we then have this yields the following

$$\frac{1}{B(7.5, 3.5)} \left(\frac{1}{2} + \frac{1}{2} \left(\frac{\phi}{\pi} \right)^3 \right)^{6.5} \left(\frac{1}{2} - \frac{1}{2} \left(\frac{\phi}{\pi} \right)^3 \right)^{2.5} \frac{3\phi^2}{2\pi^3}, \tag{66}$$

for ϕ.

Appendix D

Computations for Section 3

Fisher information for the trinomial model By definition we can derive the Fisher information by the second partial derivatives of the log-likelihood functions. The first partial derivatives for each single outcome are given by:

$$\begin{array}{llll} \frac{\partial}{\partial pL}\log P(X=L) = & \frac{1}{pL} & \frac{\partial}{\partial pM}\log P(X=L) = & 0 \\ \frac{\partial}{\partial pL}\log P(X=M) = & 0 & \frac{\partial}{\partial pM}\log P(X=M) = & \frac{1}{pM} \\ \frac{\partial}{\partial pL}\log P(X=R) = & \frac{-1}{1-pL-P_M} & \frac{\partial}{\partial pM}\log P(X=R) = & \frac{-1}{1-pL-P_M} \end{array} \tag{67}$$

The non-zero unmixed second partial derivatives for each single outcome are then given by

$$\begin{array}{llll} \frac{\partial^2}{\partial pL^2}\log P(X=L) = & \frac{-1}{pL^2} & \frac{\partial^2}{\partial pM^2}\log P(X=M) = & \frac{-1}{pM^2} \\ \frac{\partial^2}{\partial pL^2}\log P(X=R) = & \frac{-1}{(1-pL-P_M)^2} & \frac{\partial^2}{\partial pM^2}\log P(X=R) = & \frac{-1}{(1-pL-P_M)^2} \end{array} \tag{68}$$

Hence, taking the expectation over minus the left vector of (68) yields

$$I_{1,1} = I_{pL,pL} = \frac{1}{pL^2}pL + \frac{1}{(1-pL-pM)^2}(1-pL-pM) = \frac{1-pM}{1-pL-pM} \tag{69}$$

Hence, taking the expectation over minus the right vector of (68) yields

$$I_{2,2} = I_{pM,pM} = \frac{1}{pM^2}pM + \frac{1}{(1-pL-pM)^2}(1-pL-pM) = \frac{1-pL}{1-pL-pM} \tag{70}$$

The only non-zero mixed second partial derivatives with respect to both arguments pL, pM is given by

$$\frac{\partial^2}{\partial pL \partial pM}\log P(X=R) = \frac{-1}{(1-pL-pM)^2} \tag{71}$$

which yields

$$I_{2,1} = I_{1,2} = I_{pL,pM} = \frac{1}{(1-pL-P_M)^2}(1-pL-pM) \tag{72}$$

Fisher information for the cascading process model By definition we can derive the Fisher information by the second partial derivatives of the log-likelihood functions. The first partial derivatives for each single outcome are given by:

$$\begin{array}{llll} \frac{\partial}{\partial a}\log P(X=L) = & \frac{-1}{1-a} & \frac{\partial}{\partial b}\log P(X=L) = & \frac{1}{b} \\ \frac{\partial}{\partial a}\log P(X=M) = & \frac{1}{a} & \frac{\partial}{\partial b}\log P(X=M) = & 0 \\ \frac{\partial}{\partial a}\log P(X=R) = & \frac{-1}{1-a} & \frac{\partial}{\partial b}\log P(X=R) = & \frac{-1}{1-b} \end{array} \tag{73}$$

The unmixed second partial derivatives with respect to both a, b for each single outcome are then given by

$$\begin{array}{llll} \frac{\partial^2}{\partial a^2}\log P(X=L) = & \frac{-1}{(1-a)^2} & \frac{\partial^2}{\partial b^2}\log P(X=L) = & \frac{-1}{b^2} \\ \frac{\partial^2}{\partial a^2}\log P(X=M) = & \frac{-1}{a^2} & \frac{\partial^2}{\partial b^2}\log P(X=M) = & 0 \\ \frac{\partial^2}{\partial a^2}\log P(X=R) = & \frac{-1}{(1-a)^2} & \frac{\partial^2}{\partial b^2}\log P(X=R) = & \frac{-1}{(1-b)^2} \end{array} \tag{74}$$

Hence, taking the expectation over minus the left vector of (74) yields

$$I_{1,1} = I_{a,a} = \frac{1}{(1-a)^2}(1-a)b + \frac{1}{a^2}a + \frac{1}{(1-a)^2}(1-a)(1-b) = \frac{1}{a(1-a)} \tag{75}$$

Hence, taking the expectation over minus the right vector of (74) yields

$$I_{2,2} = I_{b,b} = \frac{1}{(b)^2}(1-a)b + \frac{1}{(1-b)^2}(1-a)(1-b) = \frac{1-a}{1-b} \tag{76}$$

All mixed second partial derivatives are zero. Hence, $I_{1,2} = I_{2,1} = 0$