3

Parameter Estimation

CONTENTS

To evaluate and apply the Rasch model in practical research, we need to be able to estimate its model parameters based on empirical data. Therefore, this chapter presents different approaches for estimating the person and item parameters of the Rasch model from observed data. Understanding this chapter requires a basic understanding of both maximum likelihood and Bayesian estimation. Brief introductions to these topics can be found in Appendix B.1. Just like Chapter 2, this chapter is intentionally thorough.

In this chapter, we will present four approaches to estimating the parameters of the Rasch model from observed test data. All of these approaches can be used to estimate both the item and the person parameters, but do so in different ways. Two of the approaches—joint maximum likelihood (Section 3.1) and Bayesian inference (Section 3.4)—estimate the person and item parameters simultaneously. The other two approaches—conditional maximum likelihood (Section 3.2) and marginal maximum likelihood (Section 3.3)—estimate them separately. Section 3.5 introduces the concepts of item and test information. From a practical perspective, the information is related to the uncertainty of the estimation. In Section 3.6 we discuss the sample size requirements for estimating the parameters of the Rasch model.

All of the approaches presented here rely on the likelihood function. As explained in Appendix B.1, the likelihood is the probability of the observed data, expressed as a function of the unknown model parameters. The likelihood contribution of person p's response to item i is

$$L_{u_{pi}}(\theta_p, \beta_i) = \Pr(U_{pi} = u_{pi} \mid \theta_p, \beta_i) = \frac{\exp\{u_{pi} \cdot (\theta_p - \beta_i)\}}{1 + \exp(\theta_p - \beta_i)}$$

We can compute the likelihood of person p's response to all of the test items

$i = 1, \ldots, I$ by computing the product of the likelihoods of all of p's responses (see Section 2.4.2 for details), i.e.,

$$L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta}) = \prod_{i=1}^{I} \frac{\exp\{u_{pi} \cdot (\theta_p - \beta_i)\}}{1 + \exp(\theta_p - \beta_i)}$$
$$= \frac{\exp(r_p \cdot \theta_p - \sum_{i=1}^{I} u_{pi} \cdot \beta_i)}{\prod_{i=1}^{I} [1 + \exp(\theta_p - \beta_i)]}. \tag{3.1}$$

This is the staring point for all of the estimation approaches presented here. However, the way the person parameters are dealt with depends on the approach.

3.1 Joint Maximum Likelihood Estimation

In joint maximum likelihood (often abbreviated JML) estimation, we find the person and item parameters that maximize the joint likelihood in Equation (3.1). It makes sense to select the parameters that maximize the joint likelihood, because these parameters are the most likely to have generated the observed data.

In psychometrics (as well as in statistic in general), we typically try to collect samples that are as large as possible, because they provide the most information about the population we are interested in. Suppose that we want to estimate the average income of college graduates. It would be unwise to base our estimates on a sample containing just five people, since we know the average of small samples can vary substantially. A random sample of college graduates will sometimes contain a millionaire or a pauper, even though neither will occur very often in a larger sample. These extreme values can have a large effect on the sample mean. As a consequence, we will not have much confidence in the accuracy of the sample mean for a small sample.

If we used a sample of size 100, the sample mean would be much less affected by the occasional outlier. Thus, we would have much more confidence in the accuracy of the sample mean for a sample of size 100 than a sample of size 5. In statistical terms, this means that the variance of the sample mean estimator is smaller for a sample of size 100 than a sample of size five. Were we to use a sample of size 1000, the variance of the sample mean estimator would be even smaller than for the sample of size 100. The larger the sample we use, the closer to zero the variance will be. In fact, for an infinitely large sample, the variance will essentially be zero. Many estimators have this property, which is known as *consistency*.

While straightforward, joint maximum likelihood is rarely used, because

it does not provide consistent estimators for the item parameters of a given test, even if the number of persons goes to infinity.[1]

In R, the joint likelihood can be numerically maximized using the `tam.jml` and `tam.jml2` functions in the `TAM` package (Robitzsch, Kiefer, & Wu, 2017) and the `glm` function in the `stats` package (R Core Team, 2020). However, since the joint maximum likelihood estimation is not consistent, it is not recommended and we do not provide example code.

3.2 Conditional Maximum Likelihood Estimation

One solution to this problem is to estimate the person and item parameters using a two-stage approach. In the first stage, we "hide" the person parameters, allowing us to estimate just the item parameters. In the second stage, the person parameters are estimated *conditional* on the item parameters estimated in the first stage.

As we have already noted, the likelihood of person p's responses to all of the test items $i = 1, \ldots, I$ is

$$L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta}) = \frac{\exp(r_p \cdot \theta_p - \sum_{i=1}^{I} u_{pi} \cdot \beta_i)}{\prod_{i=1}^{I}[1 + \exp(\theta_p - \beta_i)]}.$$

This includes person p's test score, or row sum, r_p. Recall from Section 2.4.1 that a test taker's score is a sufficient statistic for their person parameter θ_p. This fact can be used to factorize the joint likelihood into two parts

$$L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta}) = h(\boldsymbol{u}_p \mid r_p, \theta_p, \boldsymbol{\beta}) \cdot g(r_p \mid \theta_p, \boldsymbol{\beta}).$$

This works, because any joint probability can be split into the product of a conditional and a marginal probability. In this case, the joint probability is $\Pr(\boldsymbol{u_p}, r_p \mid \theta_p, \boldsymbol{\beta})$, the conditional probability is $h(\boldsymbol{u_p} \mid r_p, \theta_p, \boldsymbol{\beta})$ and the marginal probability is $g(r_p \mid \theta_p, \boldsymbol{\beta})$. The joint probability is

$$\Pr(\boldsymbol{u_p}, r_p \mid \theta_p, \boldsymbol{\beta}) = \Pr(\boldsymbol{u_p} \mid \theta_p, \boldsymbol{\beta}) = L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta}),$$

because the information in the score r_p is already contained in p's response pattern $\boldsymbol{u}_p$. As a consequence, including the score will not change the probability.

For us, the most interesting factor is the conditional likelihood $h(\boldsymbol{u}_p \mid$

[1] It should be noted that in general any maximum likelihood estimator is consistent (see an introductory text on mathematical statistics, like Casella and Berger (2002), for details). The lack of consistency of joint maximum likelihood is only for a given test with fixed length (Molenaar, 1995). If we could also sample infinitely many test items, joint maximum likelihood would be consistent—but this is not a realistic scenario.

$r_p, \theta_p, \boldsymbol{\beta})$. It turns out that the person parameter cancels out in the conditional likelihood, so it only depends on p's test responses u_p, their marginal sums r_p and the item parameters $\boldsymbol{\beta}$. This allows us to estimate the item parameters independently from the person parameters.

We can compute the conditional likelihood by rearranging the factorization of the likelihood of person p's responses. This gives

$$h(\boldsymbol{u}_p \mid r_p, \theta_p, \boldsymbol{\beta}) = \frac{L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta})}{g(r_p \mid \theta_p, \boldsymbol{\beta})}.$$

We already know $L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta})$, so we only need to work out $g(r_p \mid \theta_p, \boldsymbol{\beta})$, the probability of observing a particular score r_p given p's ability and the difficulties of the items. The probability of a score r_p is the probability of observing a response pattern whose sum is r_p. We can compute this probability by summing up the individual probabilities of all of the response patterns with r_p ones and $I - r_p$ zeros. Let Γ_p be the set of such response patterns. Then,

$$\begin{aligned} g(r_p \mid \theta_p, \boldsymbol{\beta}) &= \sum_{\boldsymbol{u}_p \in \Gamma_p} \Pr(\boldsymbol{u_p} \mid \theta_p, \boldsymbol{\beta}) \\ &= \sum_{\boldsymbol{u}_p \in \Gamma_p} \frac{\exp(r_p \cdot \theta_p - \sum_{i=1}^{I} u_{pi} \cdot \beta_i)}{\prod_{i=1}^{I}[1 + \exp(\theta_p - \beta_i)]}, \end{aligned}$$

after substituting the result from Equation (2.3) in Section 2.4.2.

We can simplify this expression using the fact that $\exp(x + y) = \exp(x) \cdot \exp(y)$ (see Appendix A). This allows us to factor the numerator into $\exp(r_p \cdot \theta_p) \cdot \exp(-\sum_{i=1}^{I} u_{pi}\beta_i)$, which gives

$$g(r_p \mid \theta_p, \boldsymbol{\beta}) = \sum_{\boldsymbol{u}_p \in \Gamma_p} \frac{\exp(r_p \cdot \theta_p) \cdot \exp(-\sum_{i=1}^{I} u_{pi} \cdot \beta_i)}{\prod_{i=1}^{I}[1 + \exp(\theta_p - \beta_i)]}.$$

Moreover, since $\exp(r_p \cdot \theta_p)$ and $\prod_{i=1}^{I}[1 + \exp(\theta_p - \beta_i)]$ do not contain $\boldsymbol{u}_p$, we can pull them out of the sum, yielding

$$g(r_p \mid \theta_p, \boldsymbol{\beta}) = \frac{\exp(r_p \cdot \theta_p)}{\prod_{i=1}^{I}[1 + \exp(\theta_p - \beta_i)]} \times \sum_{\boldsymbol{u}_p \in \Gamma_p} \exp\left(-\sum_{i=1}^{I} u_{pi} \cdot \beta_i\right).$$

We illustrate how the sum over response patterns can be computed by example. Suppose that test taker p correctly answered three of the items comprising a five item test correctly. Then, Γ_p is the set of response patterns with three correct items and two incorrect items. A few of these response patterns are enumerated in Table 3.1.

Let's look at the value of $\exp(-\sum_{i=1}^{I} u_{pi} \cdot \beta_i)$ for the first of these response

Item				
1	2	3	4	5
1	1	1	0	0
1	1	0	1	0
1	1	0	0	1
1	0	1	1	0
1	0	1	0	1
⋮	⋮	⋮	⋮	⋮

TABLE 3.1: Exemplary response patterns having three out of five items correct.

patterns, $\boldsymbol{u}_p = (1, 1, 1, 0, 0)$. This is

$$\begin{aligned}\exp(-\sum_{i=1}^{I} u_{pi} \cdot \beta_i) &= \exp\{-(1 \cdot \beta_1 + 1 \cdot \beta_2 + 1 \cdot \beta_3 + 0 \cdot \beta_4 + 0 \cdot \beta_5)\} \\ &= \exp(-\beta_1 - \beta_2 - \beta_3) \\ &= e^{-\beta_1} \cdot e^{-\beta_2} \cdot e^{-\beta_3}.\end{aligned}$$

If we define $\varepsilon_i = e^{-\beta_i}$, then we can simplify $e^{-\beta_1} \cdot e^{-\beta_2} \cdot e^{-\beta_3}$ to $\varepsilon_1 \cdot \varepsilon_2 \cdot \varepsilon_3$. Similarly, the value for the second response pattern, $\boldsymbol{u}_p = (1, 1, 0, 1, 0)$, is

$$\exp\left(-\sum_{i=1}^{I} u_{pi} \cdot \beta_i\right) = e^{-\beta_1} \cdot e^{-\beta_2} \cdot e^{-\beta_4} = \varepsilon_1 \cdot \varepsilon_2 \cdot \varepsilon_4,$$

and so on. Thus, the sum of

$$\begin{aligned}\sum_{\boldsymbol{u}_p \in \Gamma_p} \exp\left(-\sum_{i=1}^{I} u_{pi} \cdot \beta_i\right) = \varepsilon_1 \cdot \varepsilon_2 \cdot \varepsilon_3 + \varepsilon_1 \cdot \varepsilon_2 \cdot \varepsilon_4 + \varepsilon_1 \cdot \varepsilon_2 \cdot \varepsilon_5 \\ + \varepsilon_1 \cdot \varepsilon_3 \cdot \varepsilon_4 + \varepsilon_1 \cdot \varepsilon_3 \cdot \varepsilon_5 + \ldots\end{aligned}$$

This function is known in mathematics as an *elementary symmetric function*. The elementary symmetric functions are sums of products of a set of atoms. The number of terms in each product is the *order* of the elementary symmetric function. In our example, the atoms are $\varepsilon_1, \ldots, \varepsilon_5$. The order of the elementary symmetric function is three, because each product has three terms. Note that the order of the elementary symmetric function is equal to r_p. This is not by accident. Were we to repeat this exercise with $r_p = 2$ instead of three, we would get the elementary symmetric function of order two,

$$\begin{aligned}\gamma_2(\boldsymbol{\beta}) = \varepsilon_1 \cdot \varepsilon_2 + \varepsilon_1 \cdot \varepsilon_3 + \varepsilon_1 \cdot \varepsilon_4 + \varepsilon_1 \cdot \varepsilon_5 + \varepsilon_2 \cdot \varepsilon_3 \\ + \varepsilon_2 \cdot \varepsilon_4 + \varepsilon_2 \cdot \varepsilon_5 + \varepsilon_3 \cdot \varepsilon_4 + \varepsilon_3 \cdot \varepsilon_5 + \varepsilon_4 \cdot \varepsilon_5.\end{aligned}$$

More generally, for a test with I items, we have I atoms $\varepsilon_1, \ldots, \varepsilon_I$, where $\varepsilon_i = e^{-\beta_i}$, resulting in the general notation

$$\sum_{\boldsymbol{u}_p \in \Gamma_p} \exp\left(-\sum_{i=1}^{I} u_{pi} \cdot \beta_i\right) = \gamma_{r_p}(\boldsymbol{\beta}).$$

This allows us to simplify the expression of $g(r_p \mid \theta_p, \boldsymbol{\beta})$ to

$$g(r_p \mid \theta_p, \boldsymbol{\beta}) = \frac{\exp(r_p \cdot \theta_p) \cdot \gamma_{r_p}(\boldsymbol{\beta})}{\prod_{i=1}^{I}[1 + \exp(\theta_p - \beta_i)]}.$$

Substituting for $g(r_p \mid \theta_p, \boldsymbol{\beta})$ in the definition of $h(\boldsymbol{u}_p \mid r_p, \theta_p, \boldsymbol{\beta})$ yields

$$\begin{aligned} h(\boldsymbol{u_p} \mid r_p, \theta_p, \boldsymbol{\beta}) &= \frac{L_{\boldsymbol{u_p}}(\theta_p, \boldsymbol{\beta})}{g(r_p \mid \theta_p, \boldsymbol{\beta})} \\ &= \frac{\dfrac{\exp(r_p \cdot \theta_p - \sum_{i=1}^{I} u_{pi} \cdot \beta_i)}{\prod_{i=1}^{I}[1 + \exp(\theta_p - \beta_i)]}}{\dfrac{\exp(r_p \cdot \theta_p) \cdot \gamma_{r_p}(\boldsymbol{\beta})}{\prod_{i=1}^{I}[1 + \exp(\theta_p - \beta_i)]}} \\ &= \frac{\exp(-\sum_{i=1}^{I} u_{pi} \cdot \beta_i)}{\gamma_{r_p}(\boldsymbol{\beta})} \end{aligned}$$

after dividing out the common $\exp(r_p \cdot \theta_p)$ and $\prod_{i=1}^{I}[1 + \exp(\theta_p - \beta_i)]$ terms. As promised, the fact that the marginal sum r_p is a sufficient statistic for θ_p allowed us to derive a conditional likelihood that does not depend on the person parameter θ_p. Thus, we can write $h(\boldsymbol{u}_p \mid r_p, \theta_p, \boldsymbol{\beta})$ as $h(\boldsymbol{u}_p \mid r_p, \boldsymbol{\beta})$, which no longer conditions on θ_p.

We can emphasize that $h(\boldsymbol{u}_p \mid r_p, \boldsymbol{\beta})$ is a likelihood by writing it as $L_{\boldsymbol{u}_p}(r_p, \boldsymbol{\beta})$. We can use the individual likelihoods of the people to work out the conditional likelihood for the entire data matrix. As discussed in Section 2.4.2, responses are assumed to be independent in the Rasch model. Thus,

$$\begin{aligned} L_{\boldsymbol{u}}(\boldsymbol{r}, \boldsymbol{\beta}) &= \prod_{p=1}^{P} \frac{\exp(-\sum_{i=1}^{I} u_{pi} \cdot \beta_i)}{\gamma_{r_p}(\boldsymbol{\beta})} \\ &= \frac{\exp(-\sum_{p=1}^{P} \sum_{i=1}^{I} u_{pi} \cdot \beta_i)}{\prod_{p=1}^{P} \gamma_{r_p}(\boldsymbol{\beta})}, \end{aligned}$$

after pulling the product into the exponent. Switching the order of the sums we get that

$$L_{\boldsymbol{u}}(\boldsymbol{r}, \boldsymbol{\beta}) = \frac{\exp(-\sum_{i=1}^{I} \sum_{p=1}^{P} u_{pi} \cdot \beta_i)}{\prod_{p=1}^{P} \gamma_{r_p}(\boldsymbol{\beta})}.$$

The sums $\sum_{p=1}^{P} u_{pi}$ are simply the column sums c_i. This allows to simplify the conditional likelihood as

$$L_{\boldsymbol{u}}(\boldsymbol{r}, \boldsymbol{\beta}) = \frac{\exp(-\sum_{i=1}^{I} c_i \cdot \beta_i)}{\prod_{p=1}^{P} \gamma_{r_p}(\boldsymbol{\beta})}. \tag{3.2}$$

The item parameters are estimated by finding the value of $\boldsymbol{\beta}$ that maximizes the conditional likelihood. Owing to the complicated form of the conditional likelihood, we must apply numerical methods to determine the maximum likelihood estimate for $\boldsymbol{\beta}$. In R, we can use the `RM()` function in the `eRm` package (Mair, Hatzinger, & Maier, 2016). An example demonstrating the `RM()` function can be found in Chapter 6.

The lack of a unique origin for the latent scale (see Section 2.4.5 for details) forces us to use a linear constraint to identify the model. In practice, we usually use one of two possible constraints. The first constraint sets the value of the first item parameter to zero. In this case, we can interpret the β_i as the difficulty of item i relative to the difficulty of the first item. The second constraint sets the sum of the β_i to be 0. In this case, we can interpret the value of β_i as the difficulty of item i relative to the average difficulty of all of the items. The scaling factor of the latent scale is defined implicitly by setting the slope of all items to be one.

Once the item parameters have been estimated, they can be substituted into the the joint likelihood in order to estimate the person parameters. In doing so, the uncertainty from estimating the item parameters is typically ignored. This can lead to confidence intervals for the person parameters that are too short (Tsutakawa & Johnson, 1990), so we should be careful not to rely heavily on their exact widths.

One difficulty with using conditional maximum likelihood to estimate the person parameters is its inability to estimate the ability of test takers who correctly answered all or none of the test items. Conceptually, this makes sense. A person able to answer every test item correctly most likely has an ability that is at least as large as the most difficult item. Unfortunately, we never presented the test taker an item difficult enough to pin down exactly what their ability is. Therefore, it is reasonable not to provide an estimate for that test taker (Mair & Hatzinger, 2007).

When an estimate is required, it is possible to extrapolate values for these test takers using estimates from all of the other test takers (Mair et al., 2016). This will produce estimates, but these estimates will be affected by the quality of our extrapolation. The `person.parameter` function in `eRm` will estimate the person parameter from an item parameter fit either way.

3.3 Marginal Maximum Likelihood Estimation

Conditional maximum likelihood estimation replaces each test taker's person parameter with their corresponding sum score. This allows both the person and item parameters to be consistently estimated. Another approach to "get rid of" the person parameters when estimating the item parameters is marginal maximum likelihood estimation. In this approach, the person parameters are "averaged out" of the joint likelihood.

Marginal maximum likelihood estimation requires a marginal, or population, distribution for the person parameters. This distribution expresses the relative probability of each possible ability parameter. As De Ayala (2009) points out, marginal maximum likelihood estimation treats the abilities as random effects in the sense of mixed (or multilevel) models, whereas joint maximum likelihood estimation treats them as fixed effects.

A common choice of population distribution for the person parameters is the normal distribution. This makes sense when we can reasonably assume a symmetric distribution where the majority of people have person parameters near the mean, with just a few people having very high or very low values. Intelligence is a good example of this, since most people have average intelligence and just a few people have very high or very low intelligence.

When the population does not follow a normal distribution, assuming normality can be misleading. For example, when the population is heavily skewed, the normal distribution will do a poor job of capturing the ability distribution of the test takers. In this case, assuming normality can distort estimates of both the person and item parameters (Zwinderman & Van den Wollenberg, 1990).

Let f denote the population distribution, so that $f(\theta_p)$ is the density of p's person parameter. The density $f(\theta_p)$ will be large when θ_p is common and small when θ_p is rare. To compute the marginal density of $\boldsymbol{u}_p$, we first compute the joint density of $\boldsymbol{u}_p$ and θ_p and then we marginalize θ_p from this joint density. By definition, the joint density is

$$\begin{aligned} f(\boldsymbol{u}_p, \theta_p \mid \boldsymbol{\beta}) &= \Pr(\boldsymbol{u}_p \mid \theta_p, \boldsymbol{\beta}) f(\theta_p) \\ &= L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta}) f(\theta_p), \end{aligned}$$

since $L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta}) = \Pr(\boldsymbol{u}_p \mid \theta_p, \boldsymbol{\beta})$.

We marginalize out θ_p by integrating $f(\theta_p, \boldsymbol{u}_p \mid \boldsymbol{\beta})$ with respect to θ_p. The resulting marginal likelihood of $\boldsymbol{u}_p$ is

$$L_{\boldsymbol{u}_p}(\boldsymbol{\beta}) = \int_{-\infty}^{\infty} f(\boldsymbol{u}_p, \theta_p \mid \boldsymbol{\beta}) d\theta_p.$$

Often, the marginal likelihood is written in terms of the likelihood and population distribution, rather than the joint density. In this case, the marginal

likelihood of $\boldsymbol{u}_p$ is written

$$L_{\boldsymbol{u}_p}(\boldsymbol{\beta}) = \int_{-\infty}^{\infty} L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta}) f(\theta_p) d\theta_p.$$

The full likelihood is obtained by multiplying the likelihoods for the individual response patterns, resulting in

$$L_{\boldsymbol{u}}(\boldsymbol{\beta}) = \prod_{p=1}^{P} L_{\boldsymbol{u}_p}(\boldsymbol{\beta}) = \prod_{p=1}^{P} \int_{-\infty}^{\infty} L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta}) f(\theta_p) d\theta_p.$$

The marginal likelihood can be understood mathematically in the following way. Since

$$f(\boldsymbol{u}_p, \theta_p \mid \boldsymbol{\beta}) = L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta}) f(\theta_p)$$

the marginal likelihood of $\boldsymbol{u}_p$ can be expressed as

$$L_{\boldsymbol{u}_p}(\boldsymbol{\beta}) = \int_{-\infty}^{\infty} L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta}) f(\theta_p) d\theta_p = E_{\theta_p}[L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta})],$$

where $E_{\theta_p}[L_{\boldsymbol{u}_p}(\theta_p, \boldsymbol{\beta})]$ is the expected value of the joint likelihood over $f(\theta_p)$. This tells us that the contribution of person p's responses to the marginal likelihood of β is the *average* likelihood of his or her responses over the population. Thus, marginal maximum likelihood finds the value of $\boldsymbol{\beta}$ that maximizes the probability of the observed response patterns for test takers of average ability.

In R, marginal maximum likelihood estimation is provided by both the `mirt` (Chalmers, 2012) and `TAM` packages (Robitzsch et al., 2017). We demonstrate how to use these packages in Chapter 7 and Chapter 8, respectively. There is also an older package using marginal maximum likelihood estimation, `ltm` (Rizopoulos, 2006), but it is no longer maintained. Thus, we no longer recommend using it.

Just like conditional maximum likelihood, marginal maximum likelihood estimates the person parameters in a second step, often called *scoring*. The simplest way to do this would be to plug the item parameter estimates into the joint likelihood. Unfortunately, this tends to be inaccurate for small test sizes, so it is generally not recommended (Hoijtink & Boomsma, 1995). A popular method for dealing with this problem is to use Warm's (1989) weighted likelihood estimator. This estimator is constructed to have better performance for most test takers. Another alternative are Bayesian estimators, such as the maximum a posterior (MAP) and expected a posteriori estimators (EAP). All of these estimators are offered by both the `mirt` and `TAM` packages.

3.4 Bayesian Estimation

Bayesian estimation is an increasingly popular way of estimating the parameters of the Rasch model (Fox, 2010). Like joint maximum likelihood, Bayesian estimation simultaneously estimates both the person and item parameters. However, while joint maximum likelihood estimation finds the values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ by maximizing the joint likelihood, Bayesian estimation uses Bayes' rule to find the posterior density $f(\boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{U})$. For a primer on Bayesian inference, see the Appendix B.1.4.

For the Rasch model, Bayes' rule states that

$$f(\boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{U}) = \frac{\Pr(\boldsymbol{U} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) f(\boldsymbol{\theta}, \boldsymbol{\beta})}{\Pr(\boldsymbol{U})}. \tag{3.3}$$

The first term in the numerator, $\Pr(\boldsymbol{U} \mid \boldsymbol{\theta}, \boldsymbol{\beta})$ is the joint likelihood that we defined in Section 3.1. The second is the joint prior distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. The denominator is the average probability of the observed data over the joint prior distribution.

The use of prior distributions is a key distinction between the Bayesian approach and the "frequentist" approaches presented in the previous sections. Though theoretically we are free to choose the prior distribution to be whatever we want, it is typically chosen to reflect reasonable assumptions about the locations of the item parameters. For the Rasch model, we typically assume that the person parameters are independent draws from a normal distribution with a mean of zero and a variance of σ_θ^2. We choose a normal distribution for the reasons discussed in Section 3.3 for marginal maximum likelihood estimation. We set the mean to zero in order to fix the location of the latent scale. This is often abbreviated as $\theta_p \sim N(0, \sigma_\theta^2)$, using the distributional notation introduced in Section 2.3.2. In this expression, $N(0, \sigma_\theta^2)$ denotes a normal distribution whose mean is zero and whose variance is σ_θ^2.

Rather than assume a fixed value for σ_θ^2, we would like to infer it from the observed test data. To do this, we employ an inverse-χ^2 prior distribution for σ_θ^2. The inverse-χ^2 is the distribution of the random variable $1/Z$, when Z has a χ^2 distribution. The χ^2 distribution is a common distribution in statistics, because it is the sampling distribution of the test statistic for a number of common statistical tests. The inverse-χ^2 distribution is a common prior for variance parameters, because it only has non-zero density for positive values and has useful computational properties. The inverse-χ^2 distribution is shown for a number of different degrees of freedom ν_θ in Figure 3.1. The most common choice of ν_θ is 0.5, which is shown by the solid line. We will denote that σ_θ^2 has an inverse-χ^2 distribution with ν_θ degrees of freedom by writing $\sigma_\theta^2 \sim \text{Inv-}\chi^2(\nu_\theta)$.

The item parameters are also typically assumed to be independent draws from a normal distribution. The mean of this distribution is μ_β and its variance

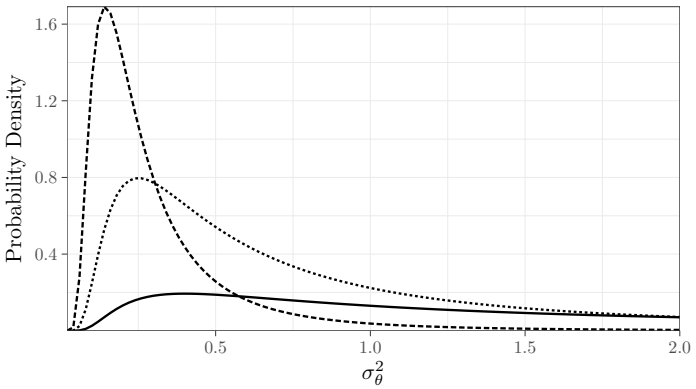

FIGURE 3.1: The inverse-
chi^2 distribution with degrees of freedom ν_0 equal to 0.5 (solid line), 2 (dotted line) and 3 (dashed line).

is σ^2_β. This is often abbreviated as $\beta_i \sim N(\mu_\beta, \sigma^2_\beta)$. We allow both the mean μ_β and the variance σ^2_β to be free parameters, and infer their values from the data. We use an improper uniform prior density for μ_β. The improper uniform density assigns equal density to every possible value of μ_β, i.e., the density of μ_β is 1 everywhere. It is called improper, because it is not a true probability distribution (its integral does not exist). We will write this as $f(\mu_\beta) \propto 1$. We again use an inverse-χ^2 distribution with ν_β degrees of freedom for σ^2_β, which we will write $\sigma^2_\beta \sim \text{Inv-}\chi^2(\nu_\beta)$.

Given these definitions, the joint prior distribution is

$$f(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2_\theta, \mu_\beta, \sigma^2_\beta) = \prod_{p=1}^{P} \phi(\theta_p \mid \mu_\beta, \sigma^2_\theta) \cdot \prod_{i=1}^{I} \phi(\beta_i \mid 0, \sigma^2_\beta) \cdot f_{1/\chi^2}(\sigma^2_\theta \mid \nu_\theta) \cdot f_{1/\chi^2}(\sigma^2_\beta \mid \nu_\beta).$$

where $\phi(\cdot \mid \mu, \sigma^2)$ is the probability density function of a normal distribution whose mean is μ and whose variance is σ^2 and $f_{1/\chi^2}(\cdot \mid \nu)$ denotes the probability density function of an inverse-χ^2 distribution with ν degrees of freedom. We did not include a prior term for μ_β, because its density is always 1. This is often written

$$\begin{aligned} \theta_p \mid \sigma^2_\theta &\sim N(0, \sigma^2_\theta) \\ \beta_i \mid \mu_\beta, \sigma^2_\beta &\sim N(\mu_\beta, \sigma^2_\beta) \\ f(\mu_\beta) &\propto 1 \\ \sigma^2_\theta &\sim \text{Inv-}\chi^2(\nu_\theta) \\ \sigma^2_\beta &\sim \text{Inv-}\chi^2(\nu_\beta). \end{aligned}$$

The joint posterior $f(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2_\theta, \mu_\beta, \sigma^2_\beta \mid \boldsymbol{u}, \nu_\theta, \nu_\beta)$ is defined by Equation (3.3). We can compute the numerator by substituting the joint likelihood for $\Pr(\boldsymbol{u} \mid \boldsymbol{\theta}, \boldsymbol{\beta})$ and $f(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2_\theta, \mu_\beta, \sigma^2_\beta \mid \nu_\theta, \nu_\beta)$ for the prior. The denominator, $\Pr(\boldsymbol{u} \mid \nu_\theta, \nu_\beta)$, cannot be computed analytically. To deal with this, we sample $f(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2_\theta, \mu_\beta, \sigma^2_\beta \mid \boldsymbol{u}, \nu_\theta, \nu_\beta)$ using Markov chain Monte Carlo (MCMC) methods. These methods provide a way to sample probability distributions that are only defined up to a proportionality constant. In Bayesian inference, that proportionality constant is the unknown denominator in Bayes' rule.

For the majority of applications, we do not need to deal with the details of implementing MCMC methods, as a number of software packages exist which automate this process. The best known of these packages are WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009), JAGS (Plummer, 2017) and Stan (Stan Development Team, 2016b). Each of these provides a way to specify the prior, likelihood and data, which defines the numerator in Bayes' rule. From there, each employs an MCMC algorithm to draw samples from the posterior distribution. In Chapter 9, we demonstrate how Stan can be used to sample the posterior of the Rasch model.

We can use samples from the posterior distribution to produce point and interval estimates of the parameters of interest. Bayesian analyses typically use the posterior mean as a point estimate of the unknown parameters. The posterior mean can be computed from posterior samples by computing the sample mean of each unknown parameter. Suppose that we have $T = 1000$ posterior samples and let $\theta_p^{(t)}$ and $\beta_i^{(t)}$ be the tth samples of θ_p and β_i. Then, the posterior mean estimates of θ_p and β_i are

$$\hat{\theta}_p = \frac{1}{T}\sum_{t=1}^{T} \theta_p^{(t)} \qquad \text{and} \qquad \hat{\beta}_i = \frac{1}{T}\sum_{t=1}^{T} \beta_i^{(t)}.$$

We can compute $(1 - \alpha) \times 100\%$ highest posterior density intervals by empirically computing the appropriate quantiles. In the case of a 95% highest posterior density interval, these are the 0.025 and 0.975 quantiles.

3.5 Item and Test Information

The Rasch ICC shown in Figure 3.2 illustrates that the slope of an ICC is steepest in the center, i.e. at the difficulty of the item. The larger the slope of the ICC, the larger the difference in probability for a given difference in ability. The larger differences in probability allow us to more accurately estimate the ability of a test taker when that test taker's ability is near the difficulty of the item than when it is not.

This also means that we will be more certain about person parameter

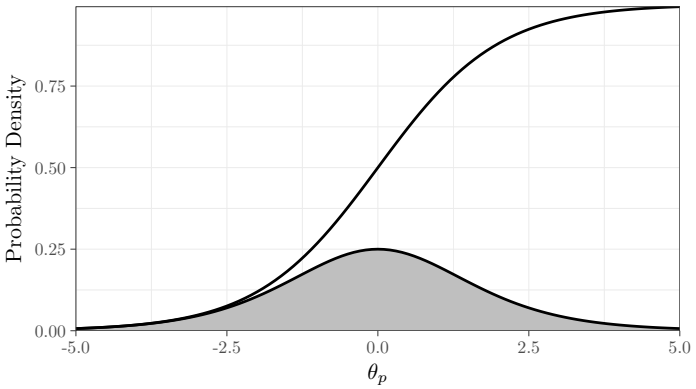

FIGURE 3.2: Item characteristic curve (top) and item information (bottom).

values near the difficulty of the item. Our certainty about the value of the person parameter is indicated by the width of its confidence interval, for the maximum likelihood estimators, and highest posterior density (HPD) interval, for the Bayesian estimator. The confidence interval for the estimated person parameter is constructed in a way that ensures it will cover the true value of the person parameter 95% of the time. The HPD interval is the shortest interval containing 95% of the posterior density. In order to contain the true value 100% of the time, the confidence interval or HPD would need to be infinitely wide. 95% represents a good comprise, in the same way that the 0.05 significance level represents a good compromise for tests.

The information an item provides about the latent trait is its discrimination, or slope, near the latent trait. Since its slope is equivalent to the derivative of a function, the information in the Rasch model is simply the derivative of the model equation,

$$\mathcal{I}(\theta_p) = \frac{\partial}{\partial \theta_p} \Pr(u_{pi} = 1 \mid \theta_p, \beta_i).$$

The shape of the item information curve is shown in Figure 3.2. The derivation of the logistic function resembles a normal distribution.[2] Its maximum value is at the item's difficulty, and it is highest in the vicinity of this point.

In any real application, we would use not only a single item but an entire test to estimate a person's ability. A test whose items are located near the ability of the test taker will provide a lot of information about the value of the person parameter. Items that are further away will provide less information.

[2]Remember that we have already discussed in Section 2.2.3 that the cumulative distribution function of the normal distribution has a very similar shape to the logistic ICC of the Rasch model. Accordingly, the derivative of the logistic ICC looks very similar to a normal distribution.

For the Rasch model the total information in a test is the sum of the individual item informations,

$$\mathcal{I}(\theta_p) = \sum_{i=1}^{I} \mathcal{I}(\theta_p),$$

which integrates all the information that each item provides about the ability of the person.

Most tests contain primarily items of moderate difficulty and only a few very easy and very difficult items. As a result, they will provide more information about test takers of average ability than test takers with very high or very low ability.

The variance of the estimate of θ_p is inversely related to the information that a test provides about a test taker. Intuitively, the more information we have, the more certain we are about the location of θ_p. This means that most tests will estimate the person parameter of the average test taker more precisely than the person parameters of an exceptional test taker.

Maximum likelihood estimators are asymptotically normally distributed (e.g. Casella & Berger, 2002). The corresponding confidence interval for the person parameter estimate is

$$\left[\hat{\theta}_p \pm z_{1-\alpha/2} \cdot \widehat{\mathcal{I}}(\theta_p)^{-1/2}\right].$$

Here $\widehat{\mathcal{I}}(\theta_p)$ is an estimate of the information that the test provides about θ_p. We can deduce from the expression for the confidence interval that $\widehat{\mathcal{I}}(\theta_p)^{-1/2}$ is the standard deviation of $\hat{\theta}_p$ and $z_{1-\alpha/2}$ determines the number of standard deviations needed to ensure the desired coverage. For example, when we want a 95% confidence interval, we set $\alpha = 0.05$ resulting in $1-\alpha/2 = 0.975$. Then, $z_{1-\alpha/2} = z_{0.975}$, the 97.5% quantile of the normal distribution, which is 1.96. This determines the width of the confidence interval.

The Bayesian posterior distribution is also influenced by the information in a test. In this case, the posterior distribution of the average test taker will have a smaller variance than the posterior distribution of an exceptional test taker. This leads to shorter HPD intervals for typical test takers by comparison to extreme test takers. For very large samples, the HPD interval for the person parameter will be similar to the confidence interval for the person parameter. However, for most tests, the two intervals will differ, as a result of the prior distribution.

As a final note, the fact that different tests provide different information about a test taker does not contradict specific objectivity. Specific objectivity tells us how the probability of a correct response relates to test taker ability and item difficulty. By constrast, information tells us how much a set of test responses tell us about the location of the ability and difficulty parameters. For a single set of response probabilities, we can sample tests offering differing amounts of information. Thus, a test whose true ability and difficulty parameters satisfy specific objectivity can result in response patterns that provide

more and less information about the locations of those ability and difficulty parameters.

3.6 Sample Size Requirements

The estimation of item parameters based on an observed sample of responses is often termed the calibration of the items. In general, a larger calibration sample allows a more accurate estimation of the item parameters, although other factors affect the accuracy of the estimation as well. For instance, the difficulty of an item can in general be estimated more accurately if the item is neither too easy nor too difficult for the person sample that worked on it.

Several publications have addressed the question which sample size is typically required for working with the Rasch model. For instance, De Ayala (2009) gives the rough calibration guideline that a calibration sample should contain at least several hundred respondents, and mentions an earlier paper from B. D. Wright (1977) that stated that a calibration sample of 500 would be more than adequate.

We agree with De Ayala (2009) that such guidelines should not be interpreted as hard-and-fast rules, but that an adequate sample size depends on the conditions and goals of the analysis. A more elaborate method for determining the necessary sample size is power analysis. Here, the goal of the analysis and the desired risks of false-positive and false-negative results or the desired accuracy need to be formalized before the analysis. The necessary sample size is then determined based on these considerations. Publications that address power analysis specifically for the Rasch model are Draxler (2010) and Draxler and Alexandrowicz (2015). Users of R can also carry out simulation studies such as those reported in the literature themselves to replicate and extend the results reported there (see Mair, 2018, Section 4.5, for exemplary R code for the 2PL model using the `mirt` package).

3.7 Exercises

1. Figure 3.3 (left) shows the likelihood for a range of values for the person parameter θ_p when the item parameters are known. Determine the maximum likelihood estimate of θ_p from the graph.
2. (a) Compare joint maximum likelihood to conditional maximum likelihood.
   (b) Compare conditional maximum likelihood to marginal maximum likelihood.

3. Determine which curve in Figure 3.3 (right) is the prior and which is the posterior.
4. Suppose you wanted to use the posterior median to estimate $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, rather than the posterior mean. How could you do this using MCMC samples?
5. In this advanced exercise—that requires a little more enthusiasm for mathematical brainteasers as well as a little more time to solve—we derive an additional estimation method that allows to calculate the item parameters of the Rasch model by hand (or by calculator).

   – Consider two items 1 and 2 under the Rasch model, with item difficulty parameter β_1 and β_2. Consider a person with ability parameter θ_1. What is the probability that this person solves item 1, but not item 2? What is in turn the probability that this person solves item 2, but not item 1? What is the ratio of these two probabilities?

   – We now define two new terms $\epsilon_1 = \exp(-\beta_1)$ and $\epsilon_2 = \exp(-\beta_2)$. How can the ratio calculated in the previous step be expressed by ϵ_1 and ϵ_2? Does this ratio depend on the ability parameter?

   – To estimate the ϵ parameters (and thus the item difficulty parameters β), we now agree that $\prod_j \epsilon_j = 1$. Show that this is equivalent to $\sum_j \beta_j = 0$.

   – Suppose we have a sufficiently large sample of test takers working on items 1 and 2, and some other items. Let n_{12} denote the number of respondents that solve item 1, but not item 2. Analogously, let n_{21} denote the number of respondents that solve item 2, but not item 1. How can we estimate $\frac{\epsilon_1}{\epsilon_2}$ based on n_{12} and n_{21}?

   – Given the relationship found in the last step and considering $\prod_j \epsilon_j = 1$, how can we estimate ϵ_i for an arbitrary test item?

   In the literature, this method is called the explicit procedure for item parameter estimation (G. H. Fischer & Scheiblechner, 1970).

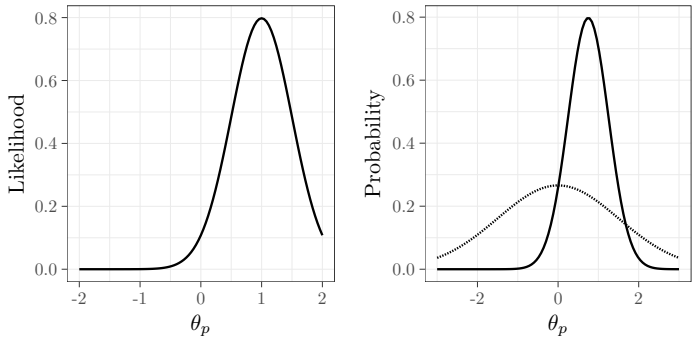

FIGURE 3.3: Figures for the end of chapter exercises. Likelihood, prior and posterior distributions for θ_p for known item parameters.