In or out; a journey through Rasch fit statistics.

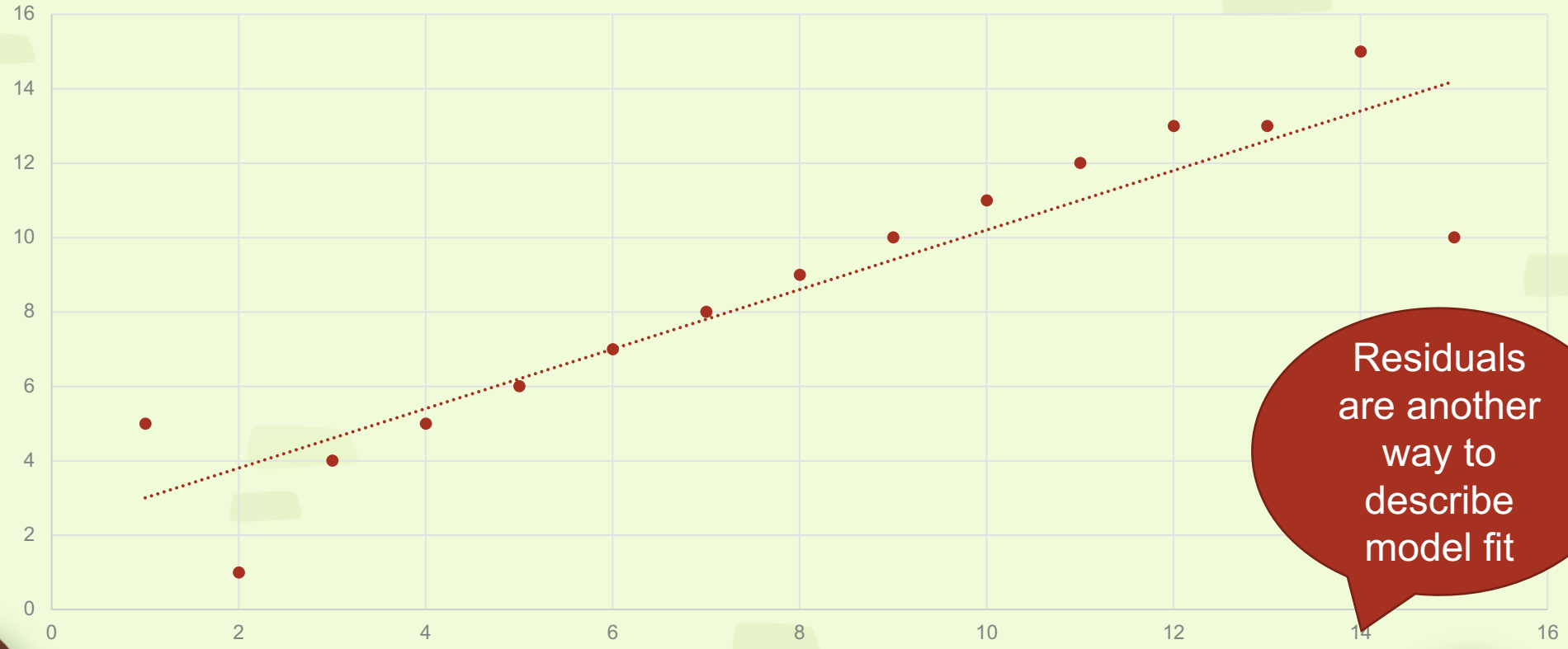
You will be able to…

- Say **what** Infit mnsq, Infit zstd, Out mnsq, Out zstd are.
- Understand **how** we get them.
- Understand what we **use** them for.

Let’s look at the Winsteps output table

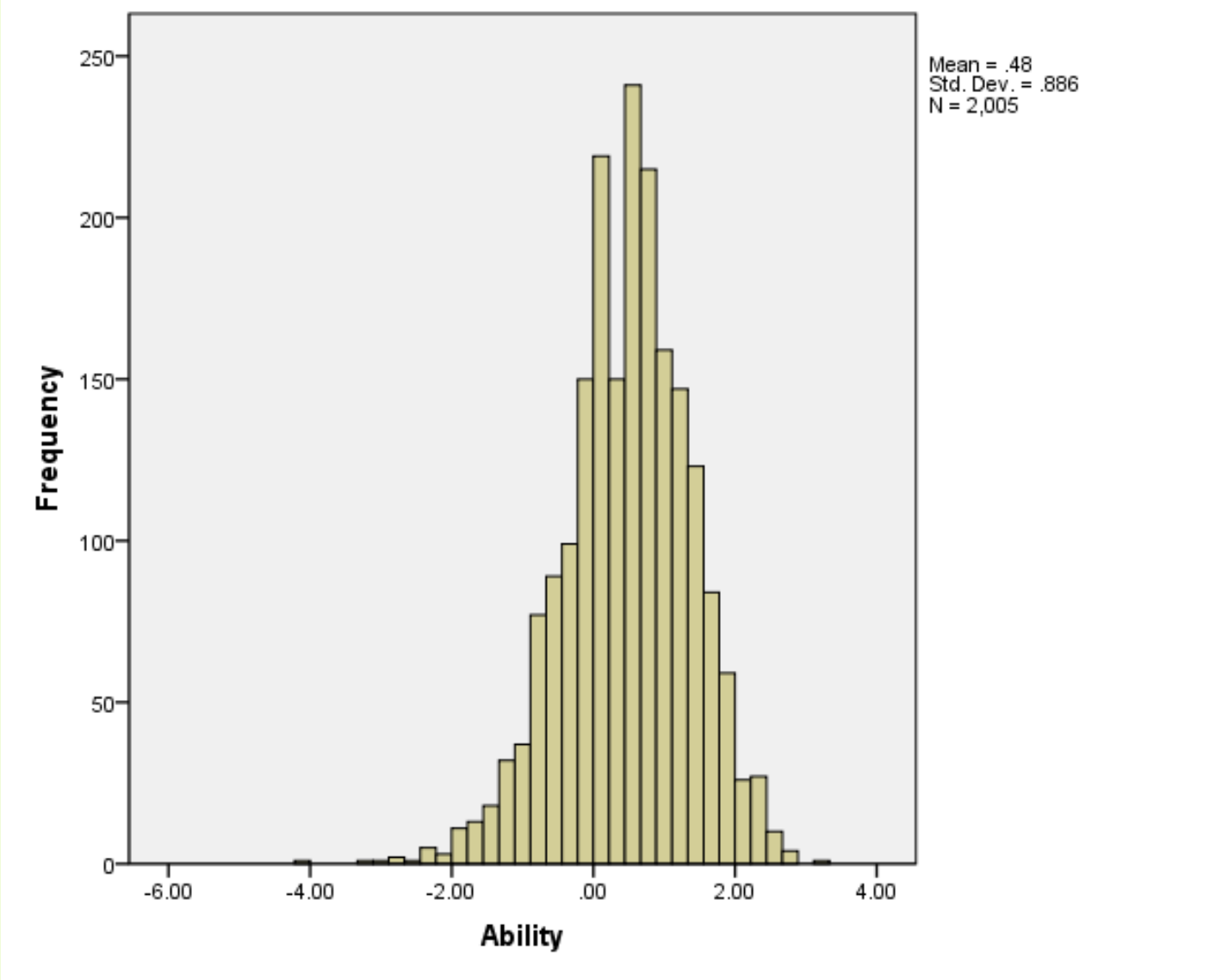
- Winsteps run

Regression example



Residuals are another way to describe model fit

Ability distribution



Infit MNSQ

- Why mean square? – In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors or deviations (https://en.wikipedia.org/wiki/Mean_squared_error)
- What do we mean by this?

Infit MNSQ = sum (residual^2) / sum (modeled variance)

- Winsteps fit statistics table
- Imagine an item with categories j=0 to m. According to the Rasch model, every category has a probability of being observed, Pj.
- Then the expected value of the observation is E = sum (j * Pj)
- The model variance (sum of squares) of the probable observations around the expectation is V = sum (Pj * (j - E) **2). This is also the statistical information in the observation. For dichotomies, these simplify to E = P1 and V = P1 * P0 = P1*(1-P1). The infit mean-square is the accumulation of squared residuals divided by their expectation.


This is just like…?
Chi square statistic!

Chi Square Statistic

	More skilled (theta > 0.5)	Average (theta between -0.5 & +0.5)	Less skilled (theta < -0.5)	Row total
Answer correctly (1)	10 (9)	5 (9)	15 (12)	30
Answer incorrectly (0)	5 (6)	10 (6)	5 (8)	20
Column total	15	15	20	Grand total: 50

$$(O\text{-}E)^2 / E$$

Outfit MNSQ

This is similar to INFIT MNSQ

… but accounting for less observations towards the end of the distribution.

OUTFIT MNSQ

- Infit mean-square = sum (residual*2) / sum (modeled variance)
- Outfit mean-square = sum (residual*2 / model variance) / (count of observations)
- Winsteps fit statistics table
- Infit was an innovation of Ben Wright's (G. Rasch, 1980, Afterword). Ben noticed that the standard statistical fit statistic (that we now call Outfit) was highly influenced by a few outliers (very unexpected observations). Ben need a fit statistic that was more sensitive to the overall pattern of responses, so he devised Infit. Infit weights the observations by their statistical information (model variance) which is higher in the center of the test and lower at the extremes. The effect is to make Infit less influenced by outliers, and more sensitive to patterns of inlying observations.

Infit MNSQ ZSTD & Outfit MNSQ ZSTD

- Infit MNSQ conforms to a Chi Square distribution
- Distribution changes with df
- Wilson and Hilferty 1931

$$t_i = (v_i^{1/3} - 1)(3/q_i) + q_i/3$$

- Where vi is the infit or outfit mnsq statistic
- qi = sum of model kurtosis minus model variance (over all the people who took the item) divided by the sum of the model variance squared (over all the people)

How do we use these statistics

- The goal is to detect misfit just as we detect outliers in regression.
- Another way of conceptualizing “infit mean square” is to view it as the ratio between observed and predicted variance. For example, when infit mean square is 1, the observed variance is exactly the same as the predicted variance. When it is 1.3, it means that the item has 30% more unexpected variance than the model predicted (Lai et al., 2003).
- In different parts of the online manual, Winsteps recommends that any mean squared above 1.0 (Winsteps & Rasch measurement Software, 2010a) or 1.5 (Winsteps & Rasch measurement Software, 2010b) is considered too big and "noise" is noticable, Lai et al. (2003) suggests using 1.3 as the demarcation point, but many other psychometricians do not recommend setting a fixed cut-off (Wang, & Chen, 2005); instead, a good practice is to check all mean squared visually.

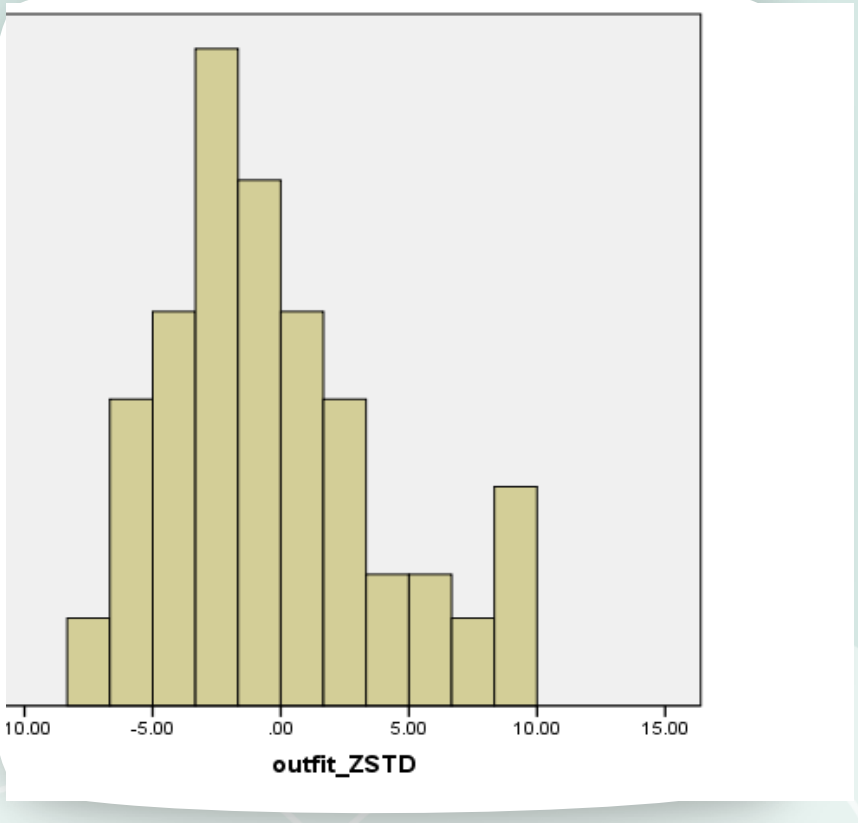
Winsteps fit statistics table – plot item fit stats

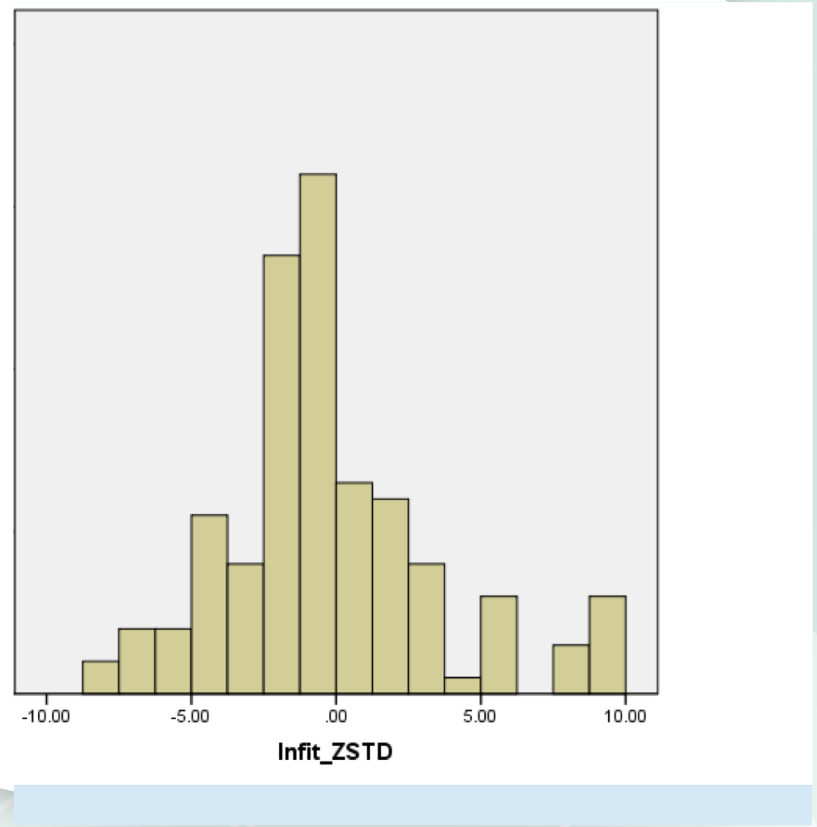
Contradictory messages

- Seen the case where the infit says there is a good fit but outfit is saying the fit is poor.
- Winsteps fit statistics table – ZTSD vs. MNSQ gives contradictory message

Reconciling these contradictory messages relies on our diagnosis strategy

- Regression standardised residuals – useful for indicating model fit. If the residuals form a normal distribution with the mean as zero, with approximately the same number of residuals above and below zero, we can tell that there is no systematic discrepancy. However if the distribution of residuals is skewed, it is likely that there is a **systematic bias,** and the regression model is invalid. While item parameter estimation, like regression, will not yield an exact match between the model and the data, the distribution of standardized residuals informs us about the goodness or badness of the **model** fit.
- A plot of ZSTD is useful for informing model fit


10.00
-5.00
.00
5.00
10.00
15.00
outfit_ZSTD


-10.00
-5.00
.00
5.00
10.00
Infit_ZSTD

Strategy for diagnosis

- Use ZSTD for model fit use Infit MNSQ for item fit.
- Start with person fit and ZSTD. Depending on purpose remove suspicious examinees.
- Plot ZSTD (histogram) and MNSQ (line chart) to identify observations that do not accord with the whole.
- Do item diagnosis after person diagnosis. Look at Infit MNSQ and Outfit MNSQ. Contradictory messages encountered key questions to ask:1) to what degree they differ from one another, and 2) does this difference lead to contradictory **conclusions** regarding the fitness of certain items.
- Parallel coordinate plot of sorting in excel. If there is a discrepancy, determining whether or not to trust the infit or outfit will depend on what your goal is. If the target audience of the test is examinees with average skill-level, an infit model index may be more informative.

Outcome

- When misfits are found, one should check the key, the distracters, and the question content first. Farish (1984) found that if misfits are mechanically deleted just based on chi-square values or standardized residuals, this improves the fit of the test as a whole, but worsens the fit of the remaining items.

What is your primary concern? Statistical fit or productive measurement?

- Statistical fit is like "beauty". Productive measurement is like "utility". Statistical fit is dominated by sample size. It is like looking at the data through a microscope. The more powerful the microscope (= the bigger the sample), the more flaws we can see in each item. For the purposes of beauty, we may well scrutinize our possessions with a microscope. Is that a flaw in the diamond? Is that a crack in the crystal? There is no upper limit to the magnification we might use, so there is no limit to the strictness of the statistical criteria we might employ. The nearer to 1.0 for mean-squares, the more "beautiful" the data. In practical situations, we don't look at our possessions through a microscope. For the purposes of utility, we are only concerned about cracks, chips and flaws that will impact the usefulness of the items, and these must be reasonably obvious. In terms of mean-squares, the range 0.5 to 1.5 supports productive measurement. However, life requires compromise between beauty and utility. We want our cups and saucers to be functional, but also to look reasonably nice. So a reasonable compromise for high-stakes data is mean-squares in the range 0.8 to 1.2.