Cambridge Assessment
Network

Critical values for Rasch item fit statistics

(joint work with Mike Horton and Guido Makransky)

Karl Bang Christensen | Univ. of Copenhagen | 21 March 2019

Contents

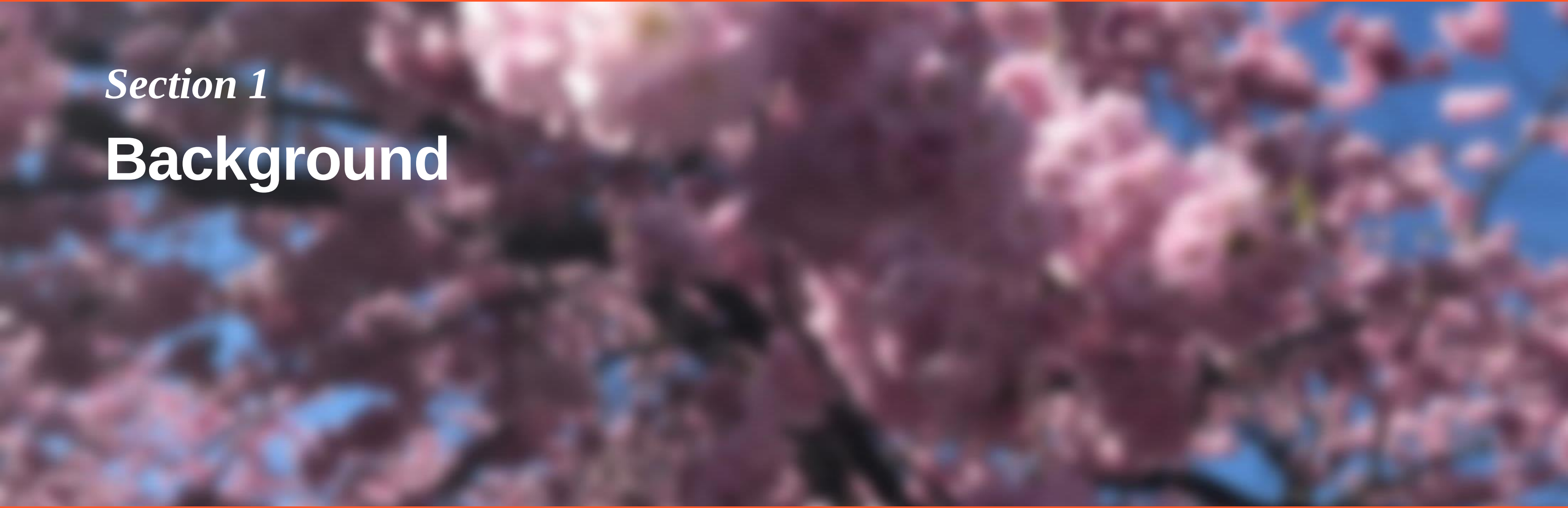
Background

Simulation study

Motivating example: AMTS

Parametric bootstrap

Recommendations



Section 1

Background

Background

Item fit statistics are arguably the most relevant and certainly the most commonly reported feature in Rasch validation studies.

Many item fit statistics exist, most widely used are the mean square item fit statistics in Winsteps and the fit residual, Chi-square and ANOVA item fit statistics in RUMM2030.

Many papers about performance of Winsteps fit statistics, few discussing RUMM2030.

Winsteps

- The null distribution for item mean squares depend on a number of factors and thus it is not appropriate to set a single set of critical values [Wu, Adams, 2013]
- Rules of thumb like $1 + 2/\sqrt{N}$ and $1 + 6/\sqrt{N}$ have been proposed, but further research is needed to establish the exact type I error rates for these approximate critical values [Smith, Schumacker, Bush, 1998]
- The fact that these rule-of-thumb values are not universally appropriate led Wolfe (2013) to identify that:

> ...users are faced with a quandary. How does one interpret a fit statistic if the distribution of the values of that statistic, and hence the range of reasonable values, is not known?

RUMM2030

Typical application

- FitResidual for each item should be in the interval *[-2.5, 2.5]*
- item chi-square evaluated using P-value after Bonferroni adjustment
- item ANOVA fit statistics using P-value after Bonferroni adjustment

No adjustment formulas for critical values have been published.

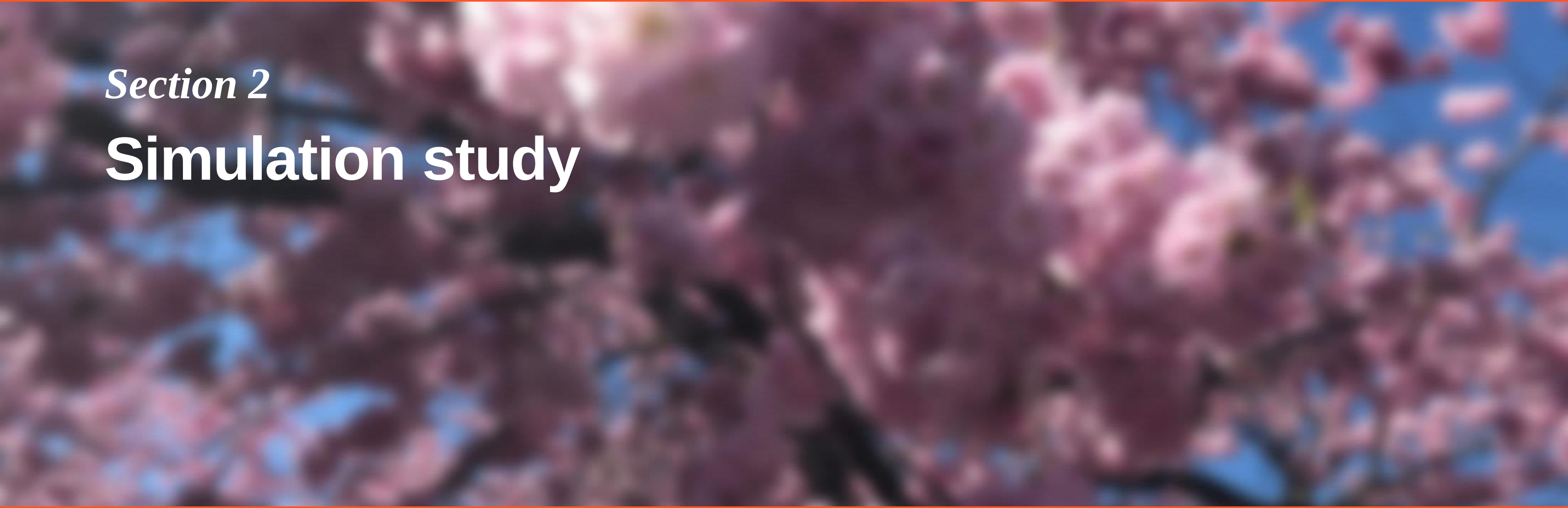
  - Evaluation of performance in samples drawn with replacement from real data [Smith er al, 2008]. This tells us nothing about type I error rates.
  - Very small simulation study [Hagell & Westergren, 2016] (N=25).

Summary

We do a Rasch analysis to identify anomalies. If we have, say, 10 items and no pre-specified hypothesis about a misfitting item there is an inherent risk of type I errors

- Do we have justification for saying that INFIT>1.3, FitResid>2.5, P<0.05 constitues an anomaly?
- Bonferroni adjustment is only correct if the P-values are correct. We have no way of adjusting for multiple testing

Will look only at fit statistics from RUM2030. All results also apply to infit, outfit and their t-transformed values



Section 2

Simulation study

Simulation study

10 dichotomous items, 300 respondents

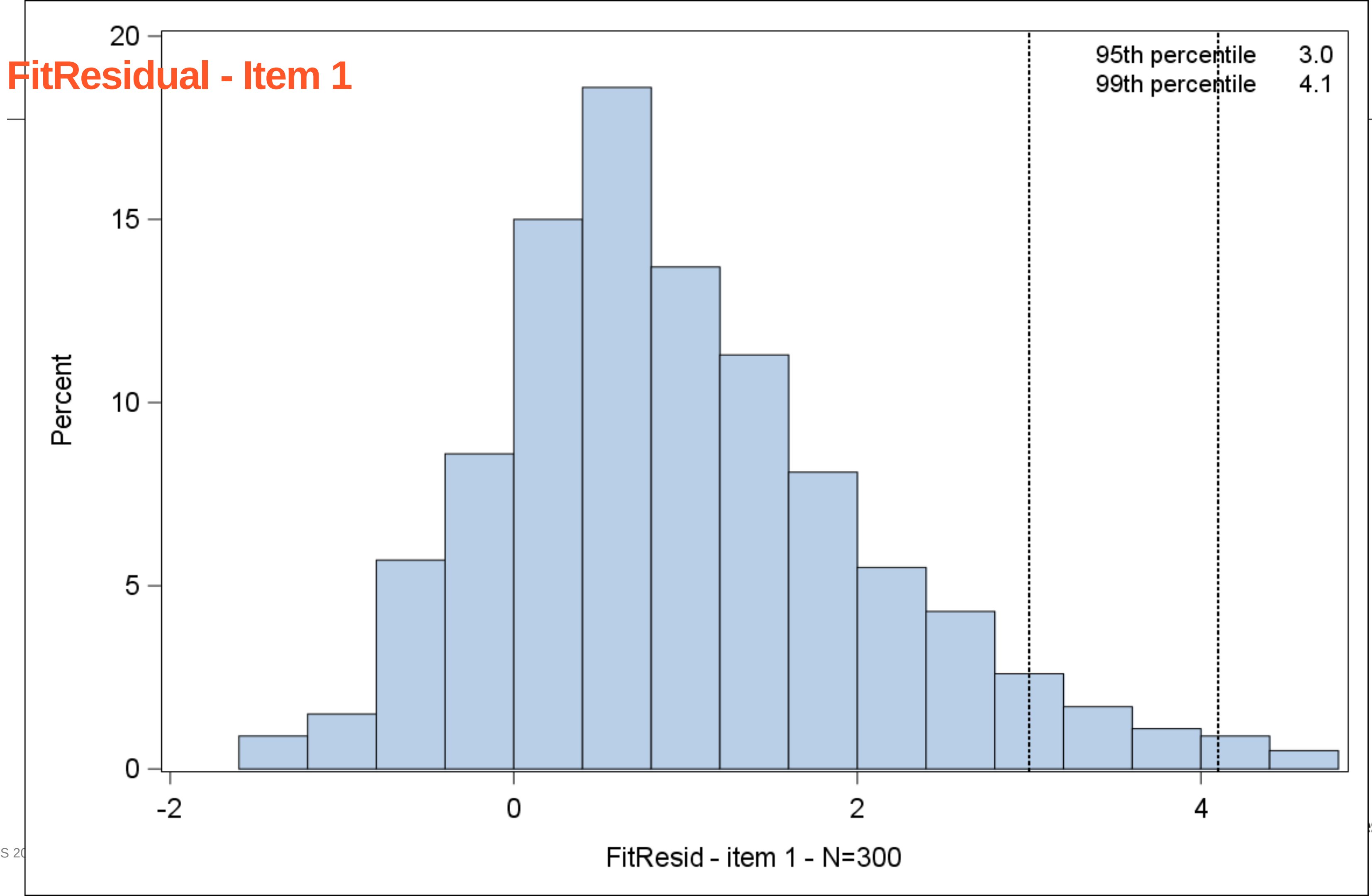
Simulate a data set where the Rasch model fits

Compute RUMM item fit statistics

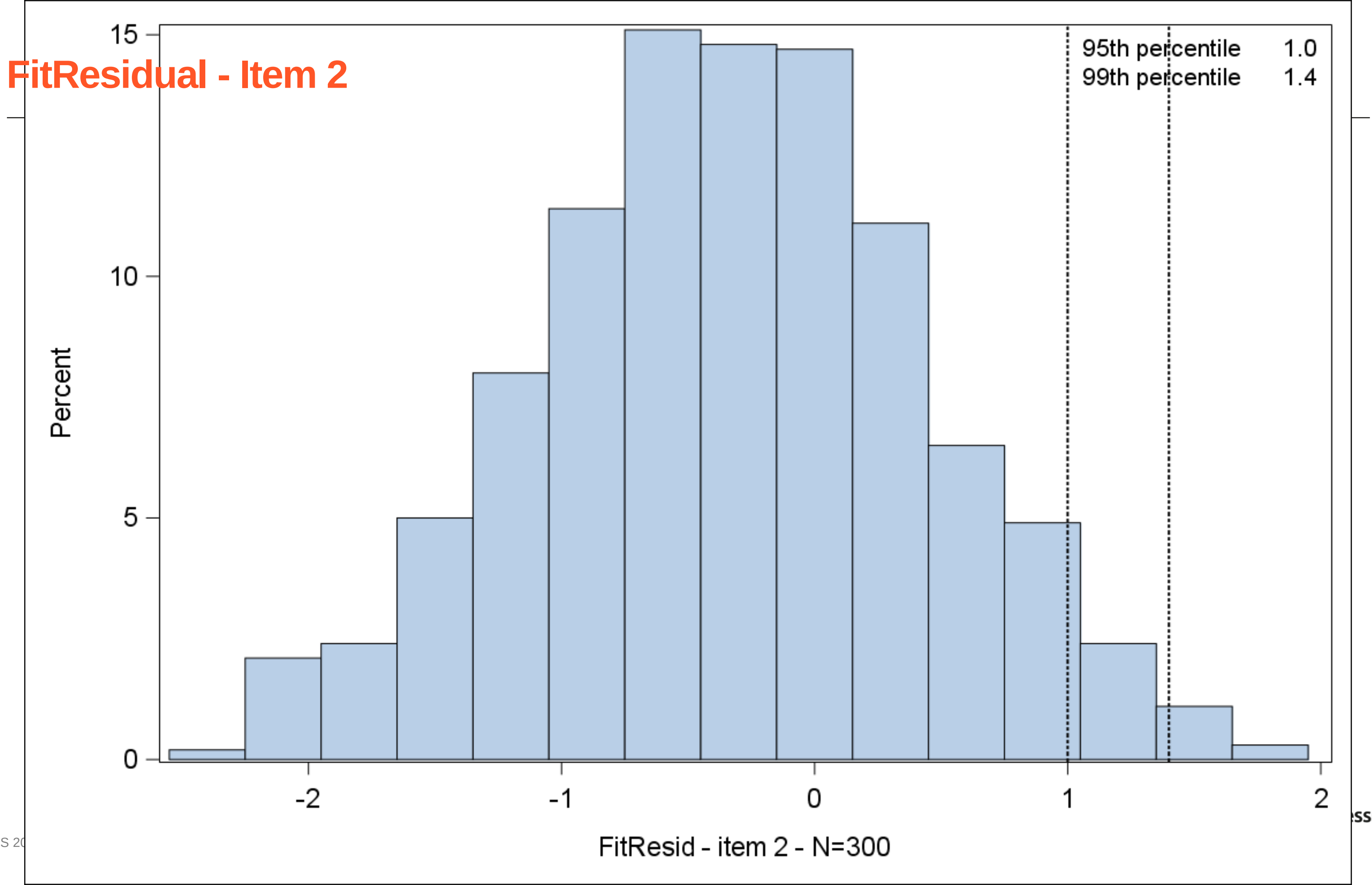
- FitResidual – look at interval *[-2.5, 2.5]*
- Fvalue (using three class intervals) – F-test statistic with (2, N-3) degrees of freedom [CV: ~3.0]
- Chisq (using three class intervals) – Chi-square distribution with 2 degrees of freedom [CV: 5.99]

Repeat 1000 times

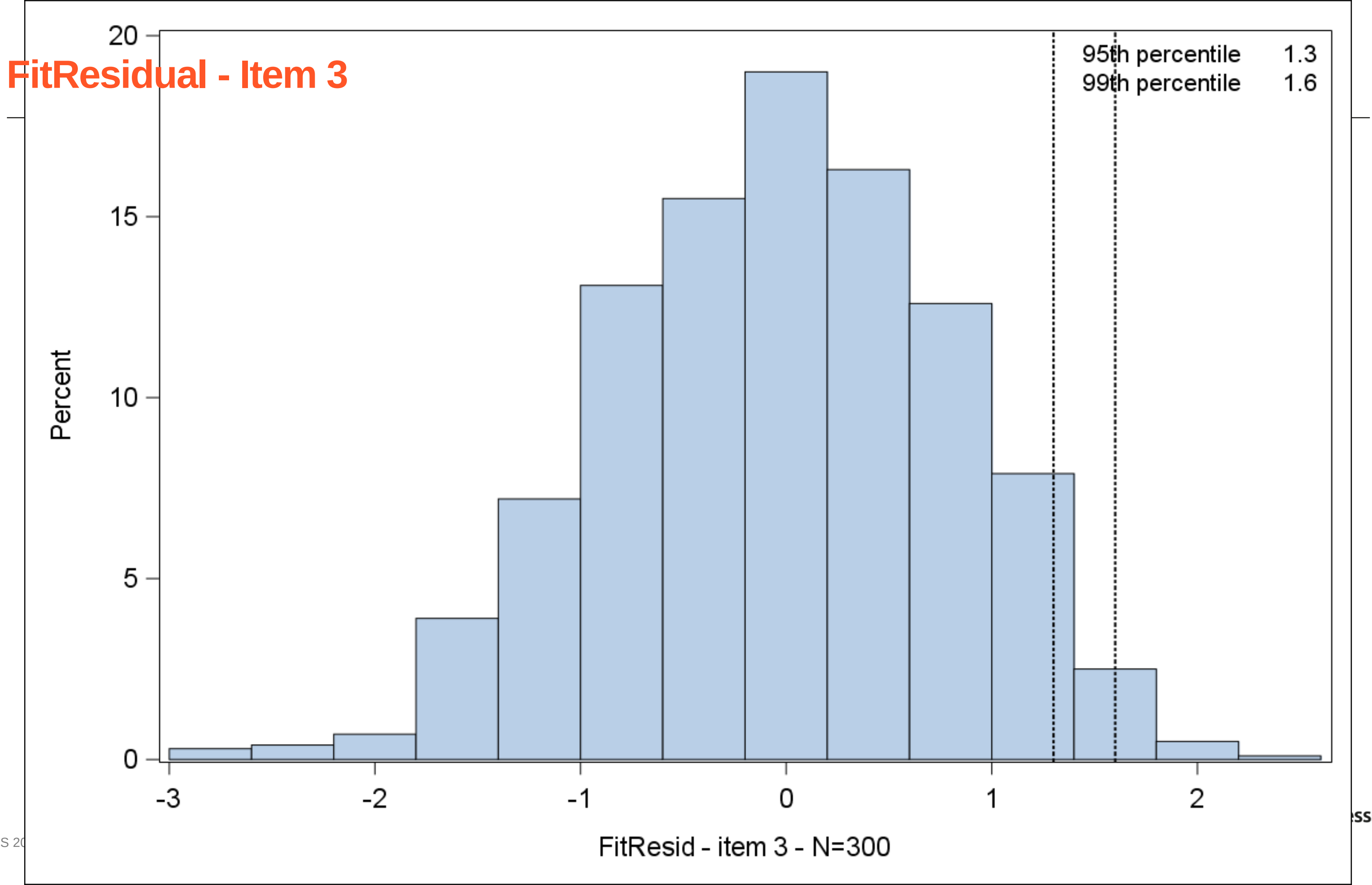
FitResidual - Item 1



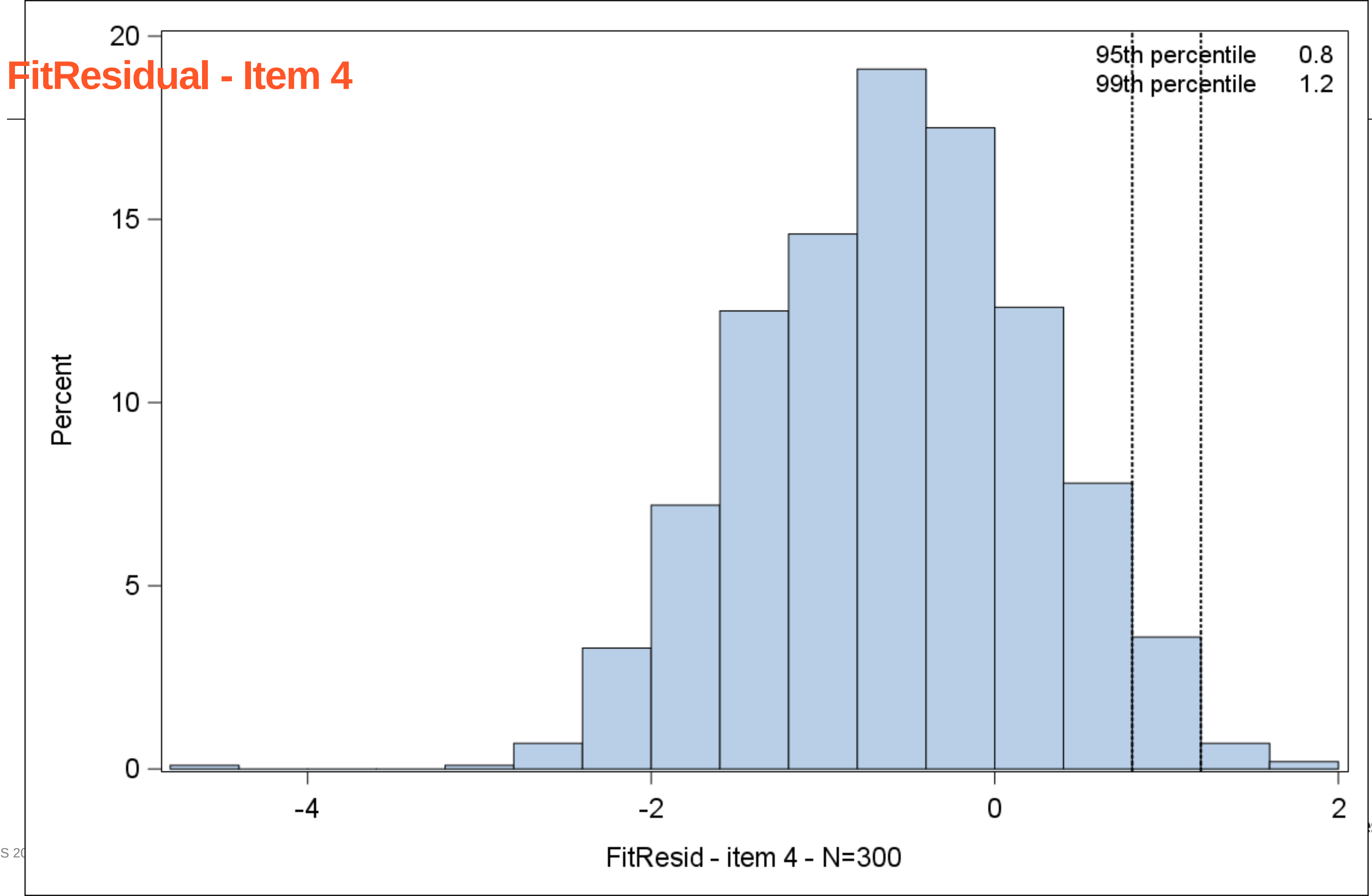
FitResidual - Item 2



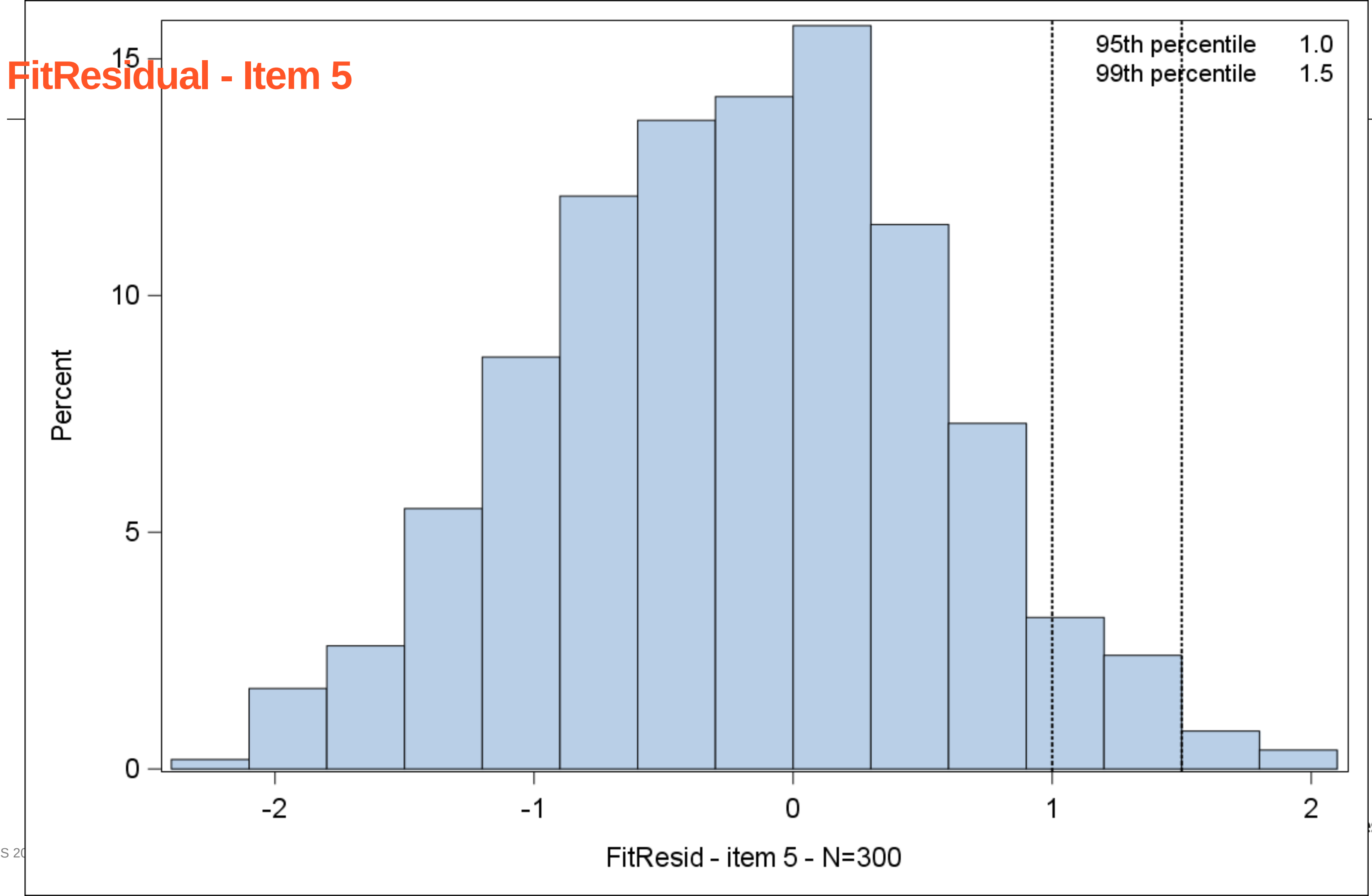
FitResidual - Item 3



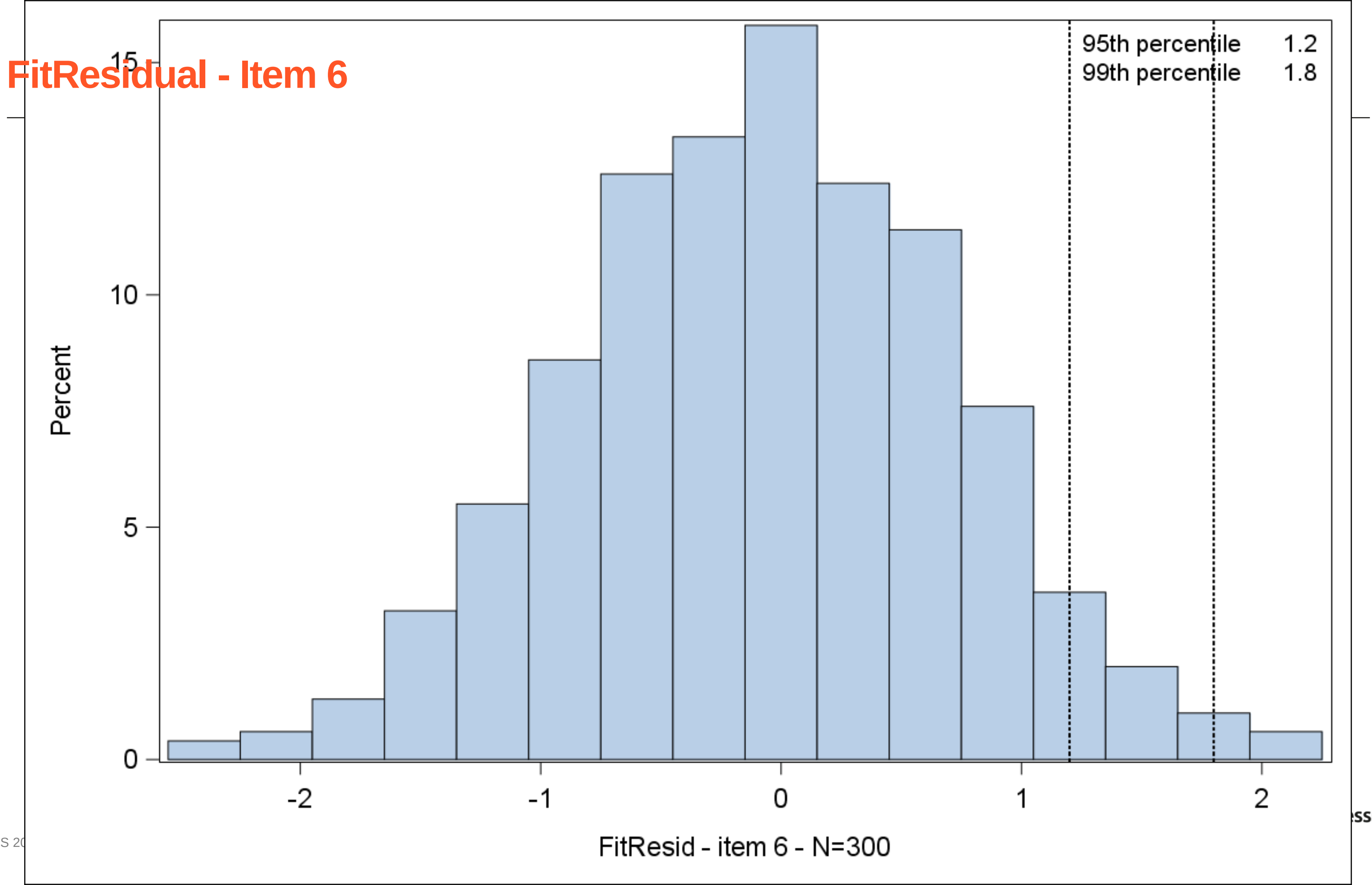
FitResidual - Item 4



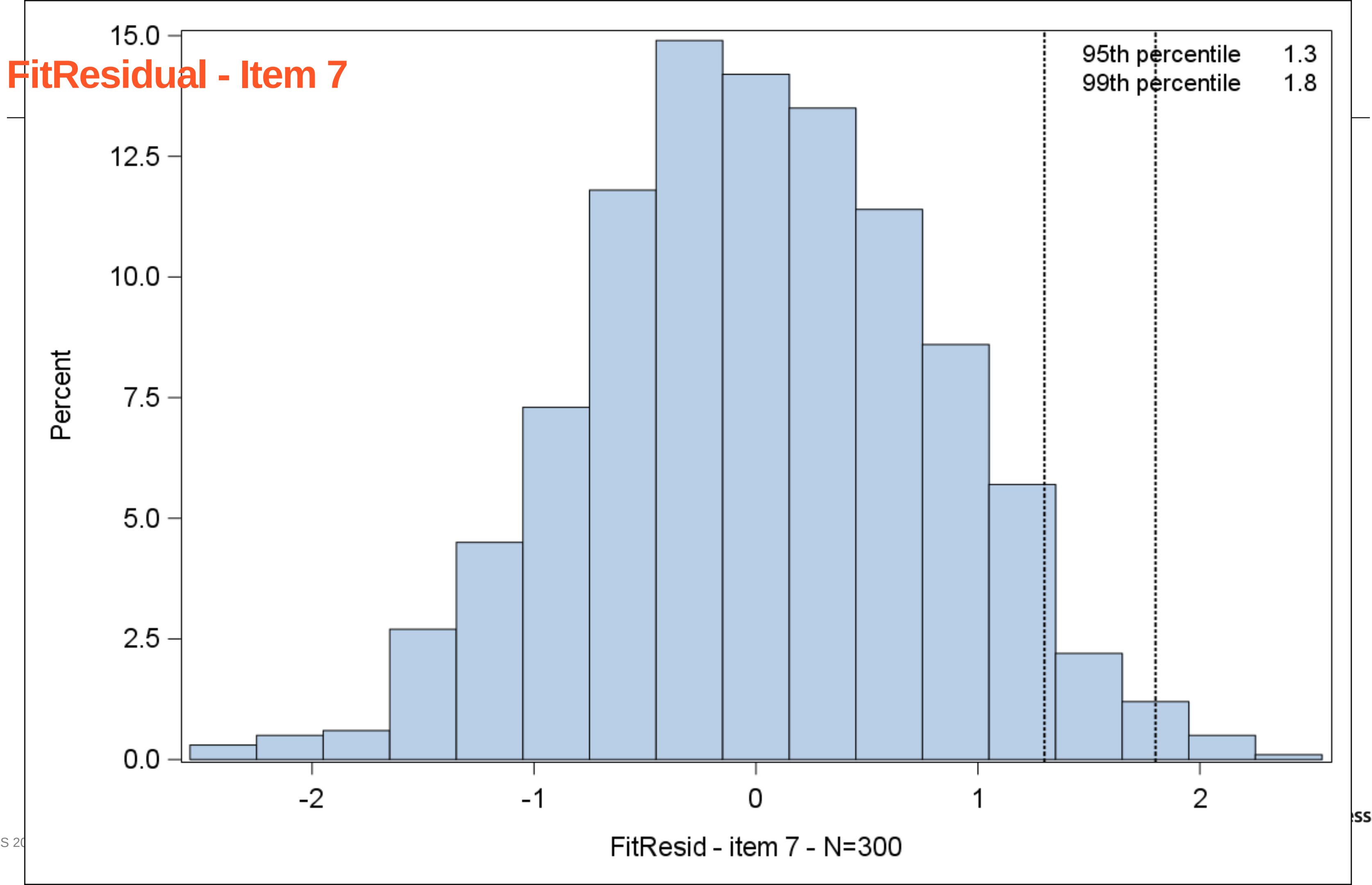
FitResidual - Item 5



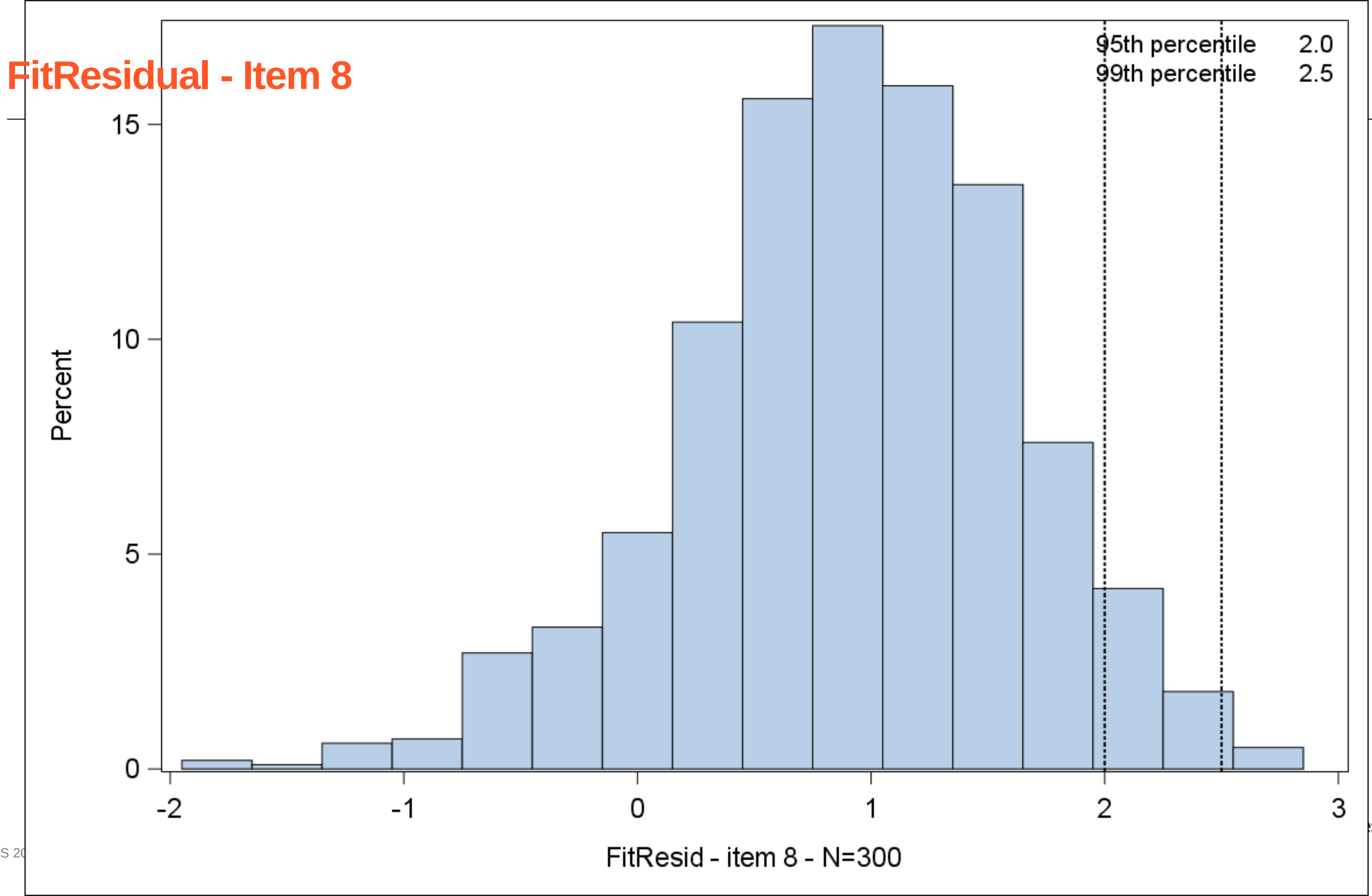
FitResidual - Item 6



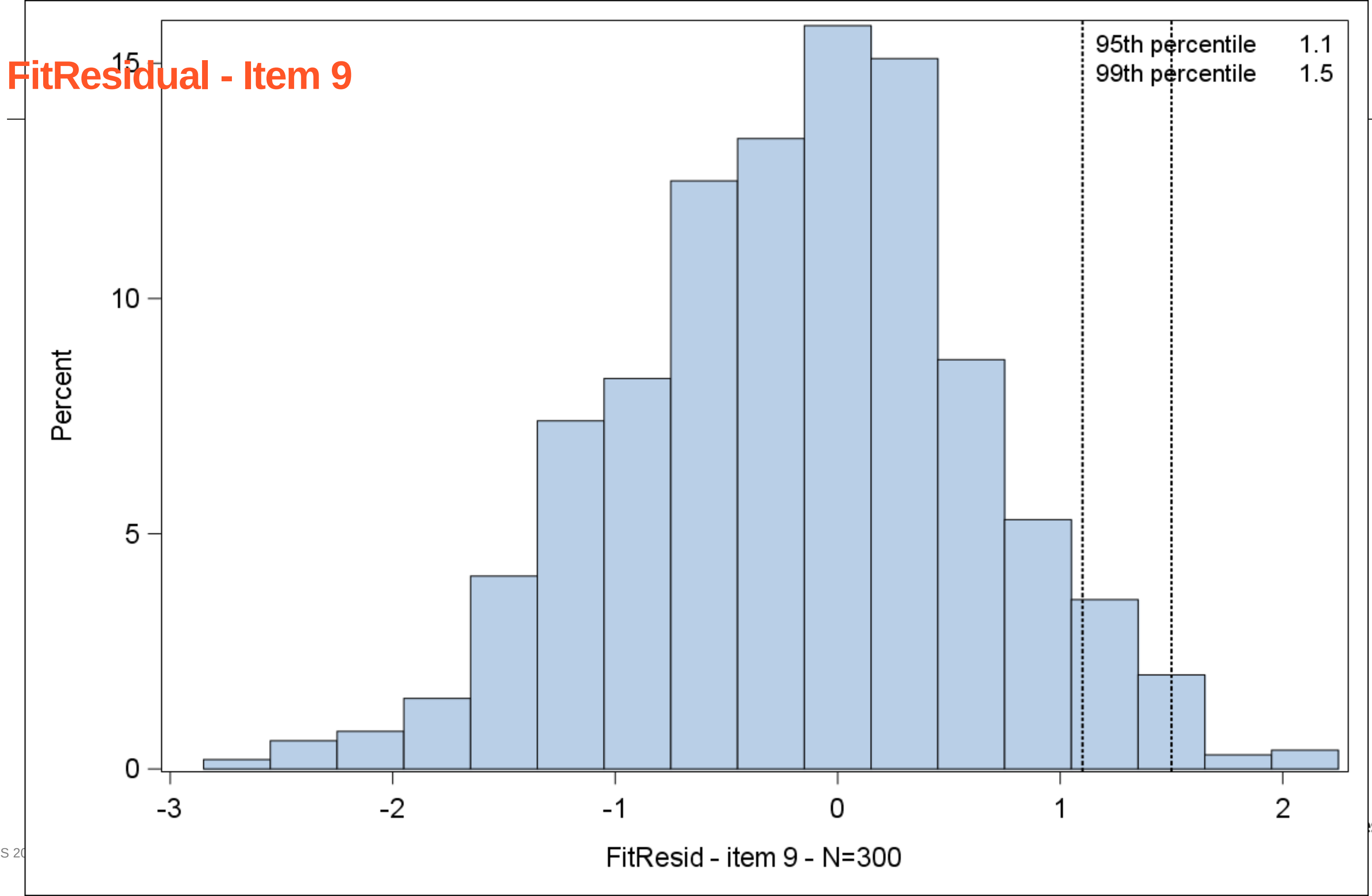
FitResidual - Item 7



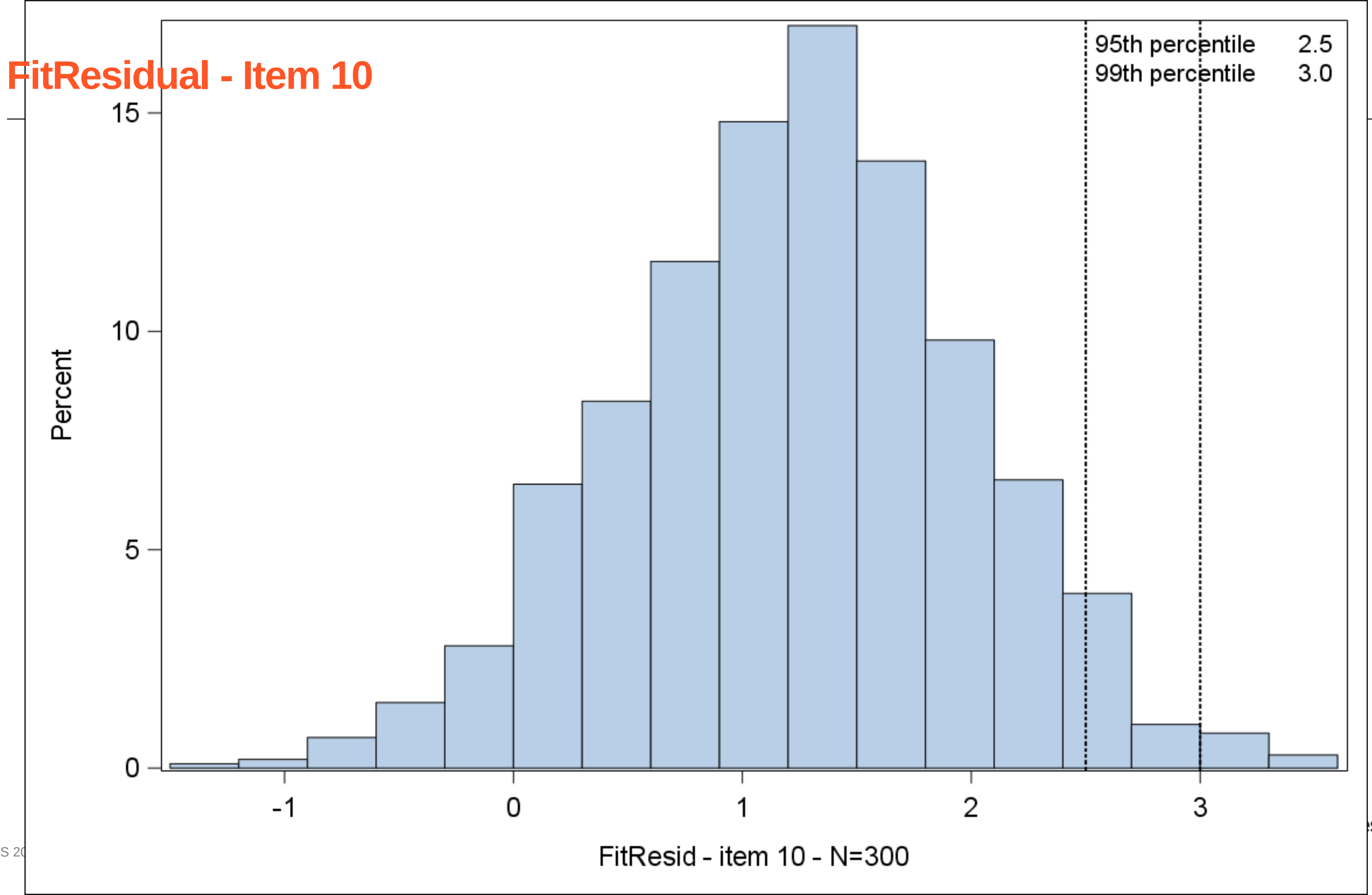
FitResidual - Item 8



FitResidual - Item 9



FitResidual - Item 10



FitResidual - Summary

item	Mean	(SD)	1st	99th
1	0.8	(1.0)	-1.2	3.7
2	-0.3	(0.8)	-2.1	1.4
3	-0.0	(0.8)	-2.0	1.7
4	-0.7	(0.9)	-2.9	1.1
5	-0.1	(0.8)	-2.0	1.6
6	-0.1	(0.8)	-1.9	1.6
7	-0.0	(0.8)	-1.8	1.7
8	0.7	(0.7)	-1.1	2.2
9	-0.1	(0.8)	-2.0	1.7
10	1.0	(0.8)	-0.8	2.7

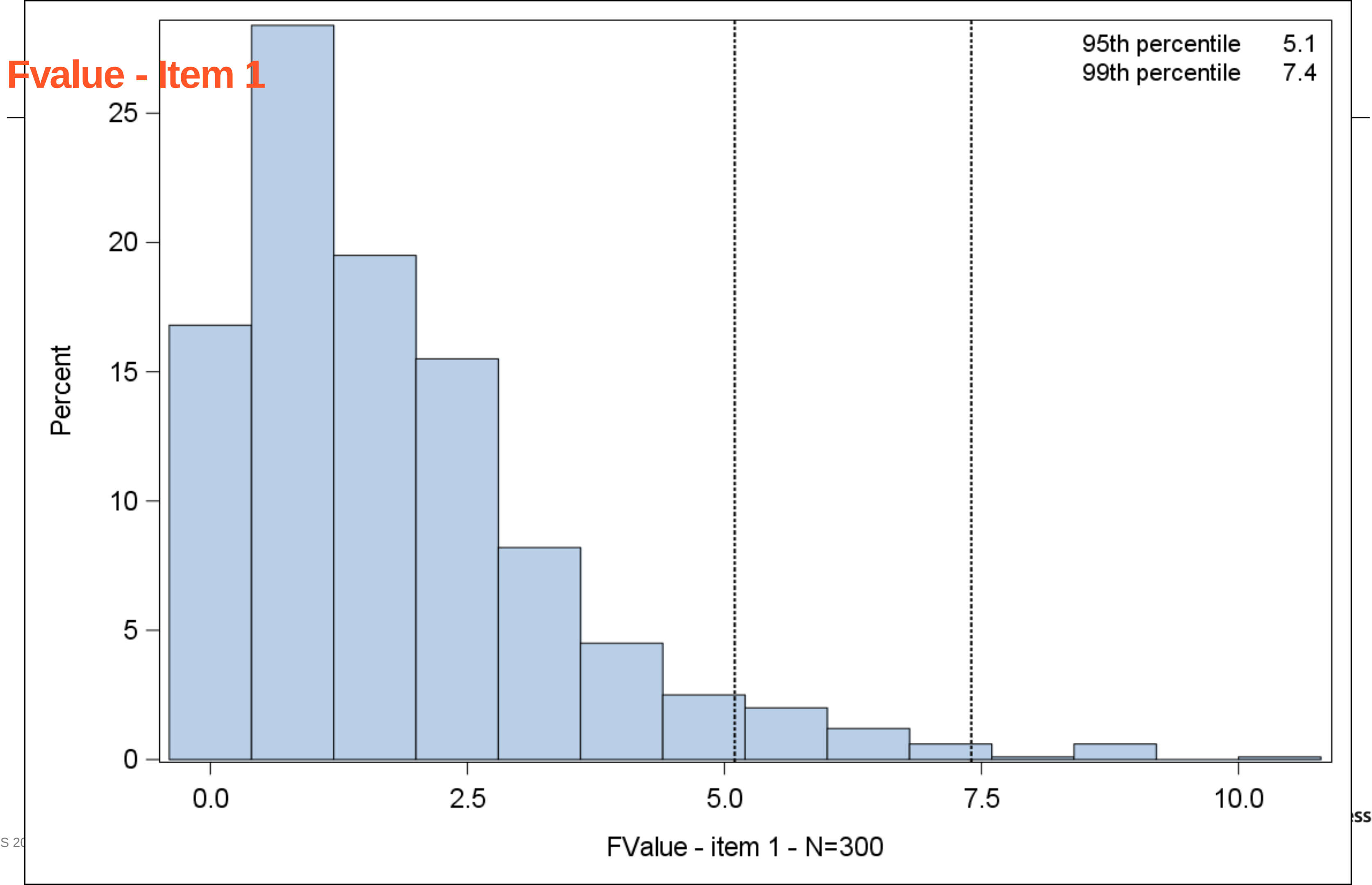
No single critical value is appropriate

Using the interval

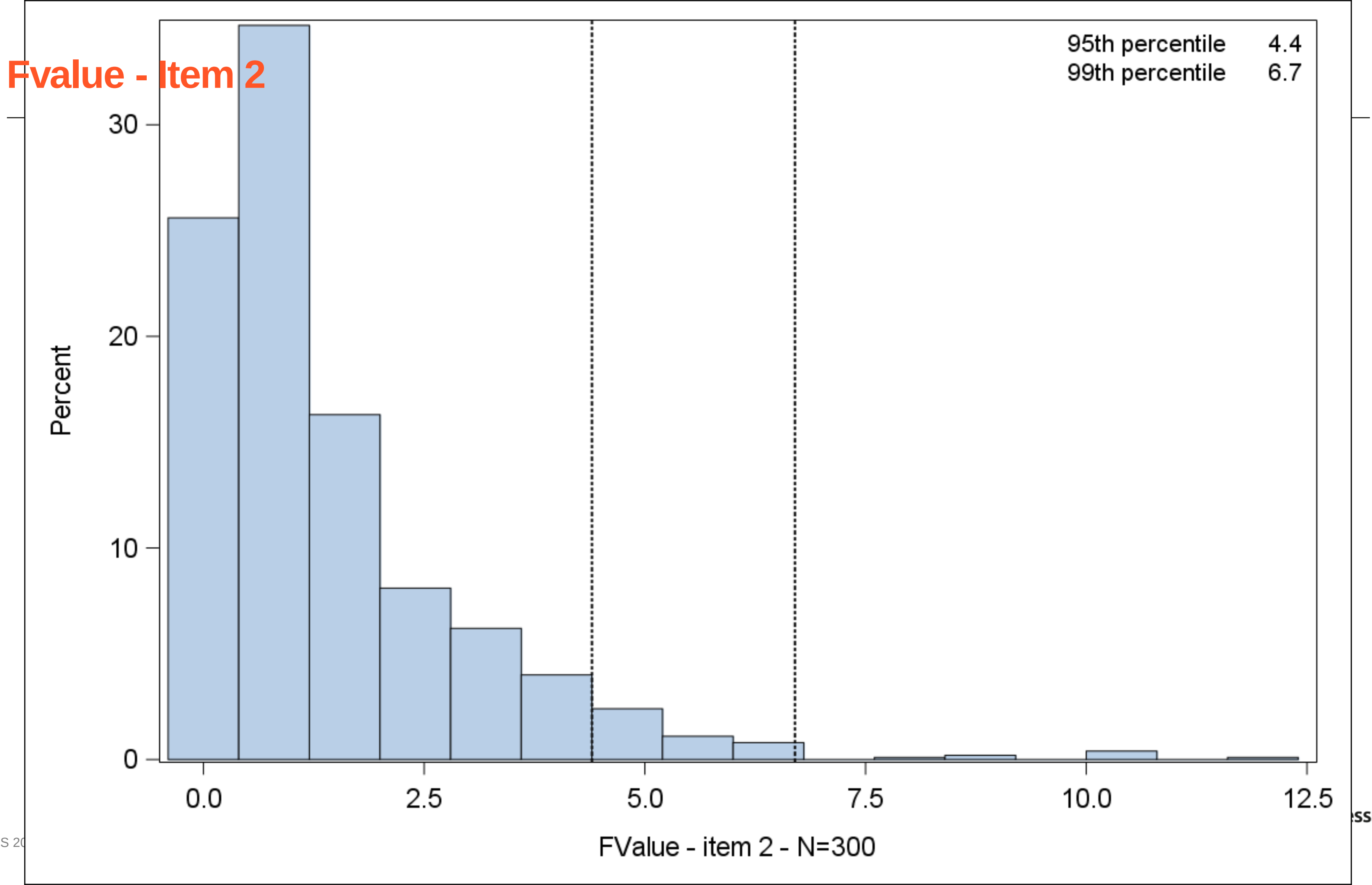
[-2.5, 2.5]

as the acceptable range means that some items will very often be rejected, while others will very rarely be rejected.

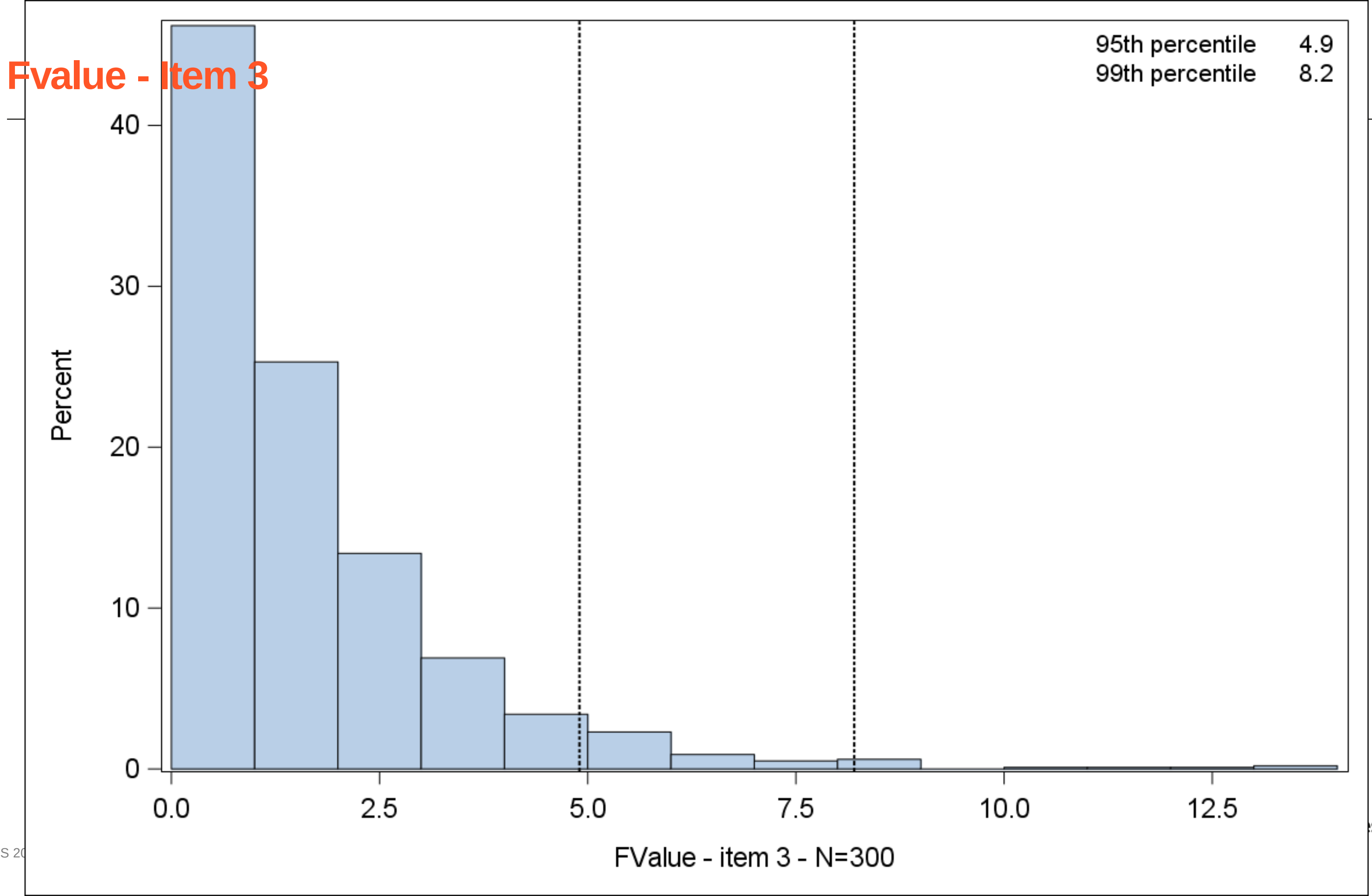
Fvalue - Item 1



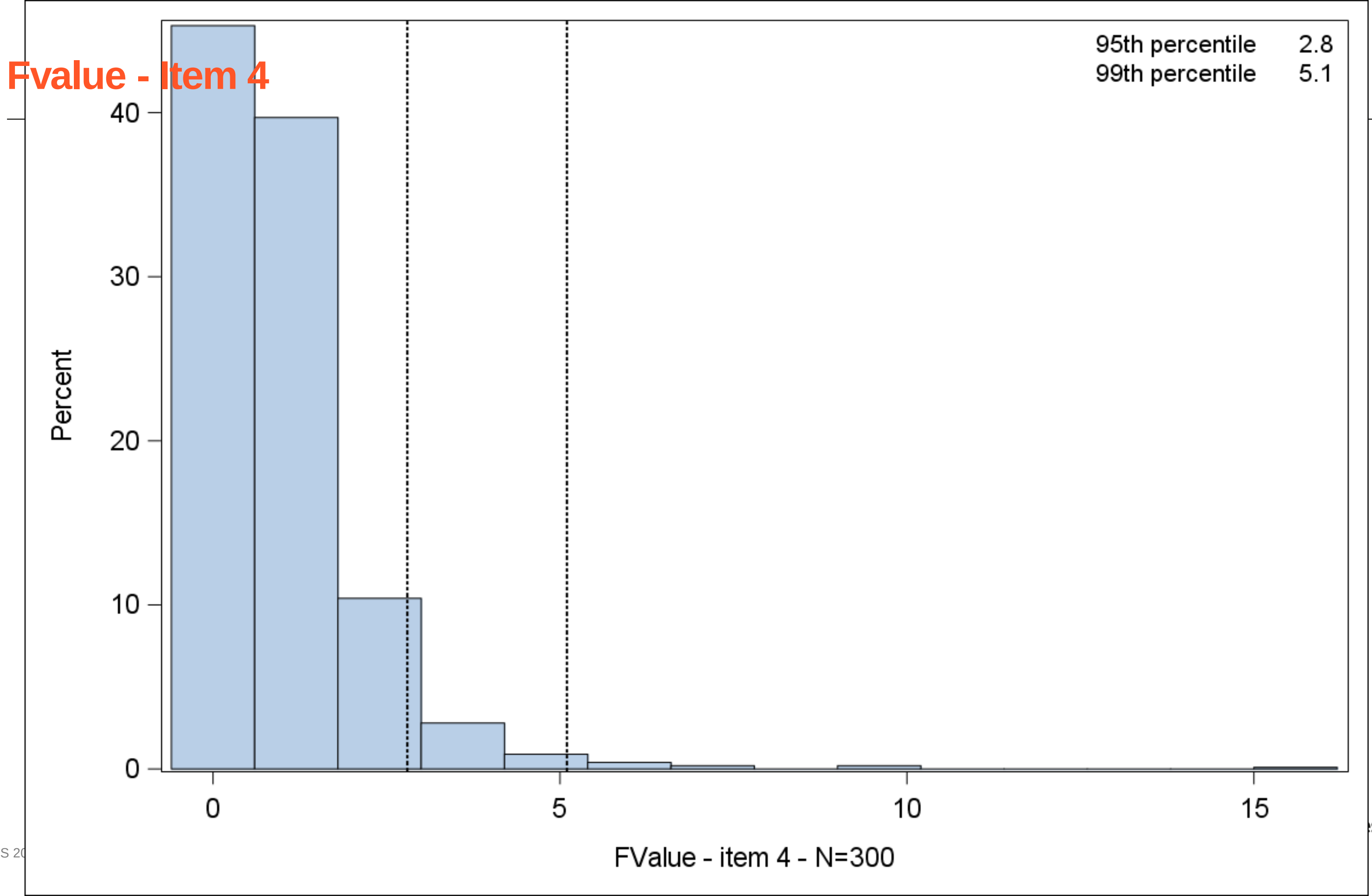
Fvalue - Item 2



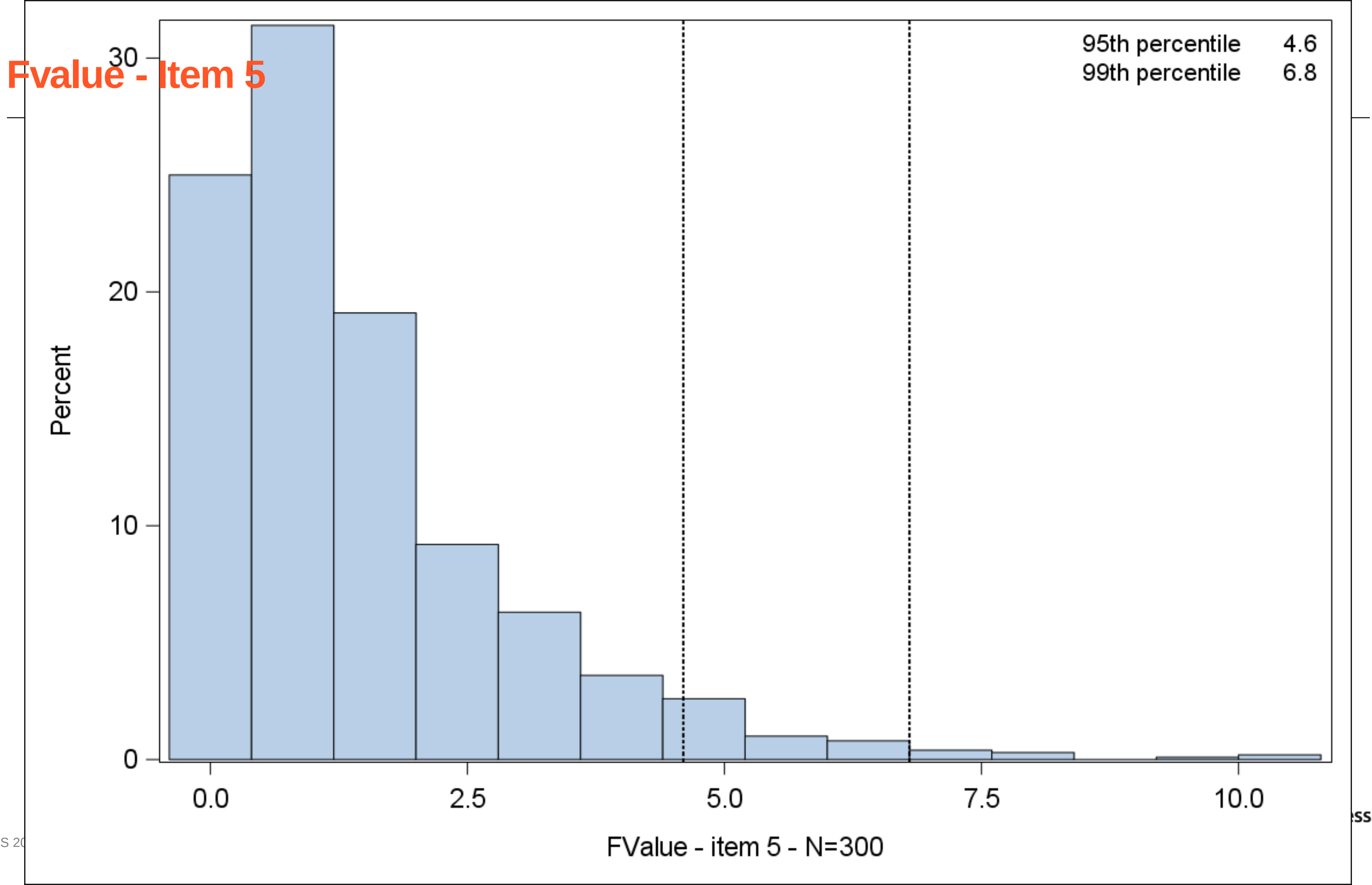
Fvalue - Item 3



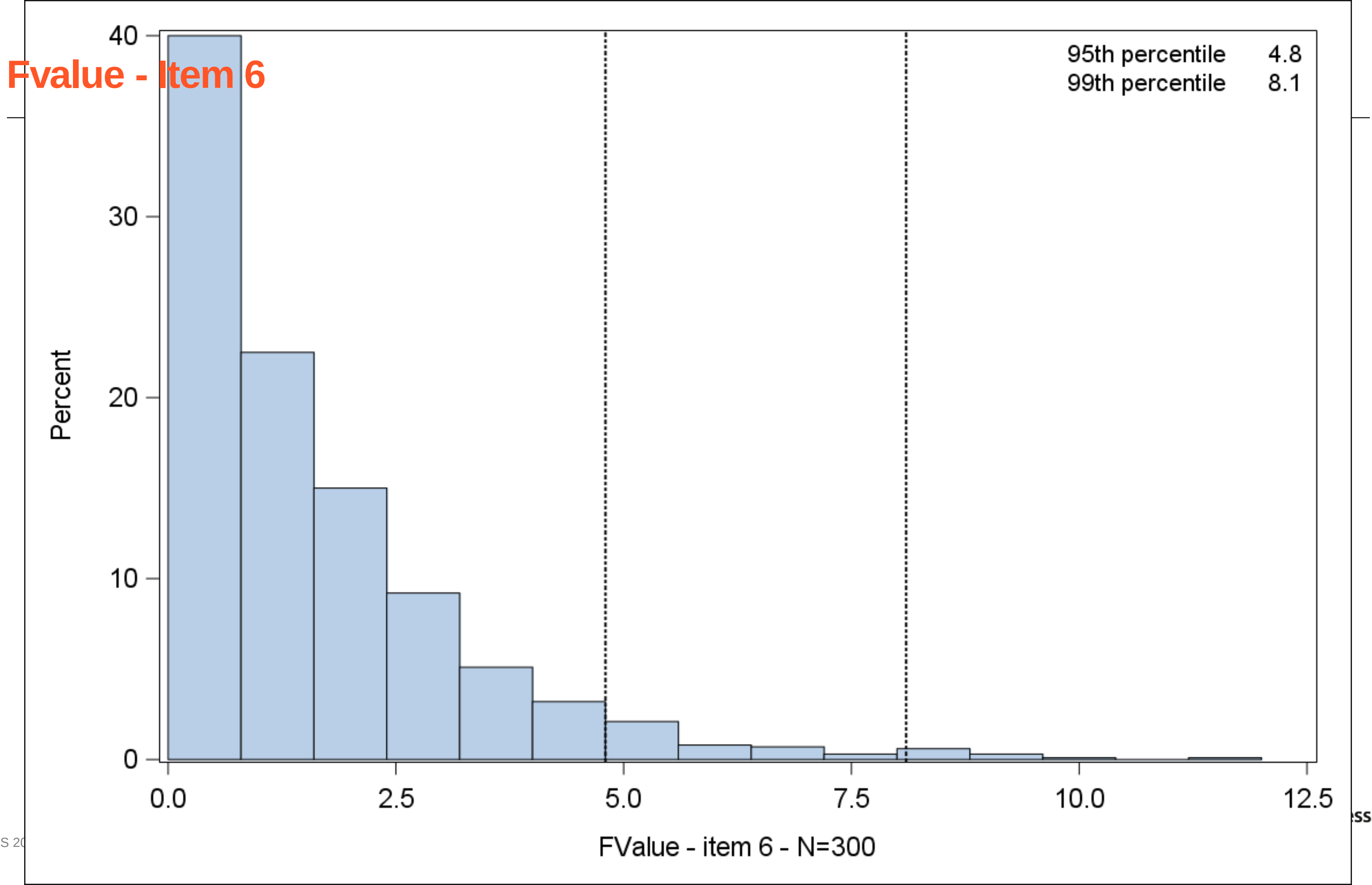
Fvalue - Item 4



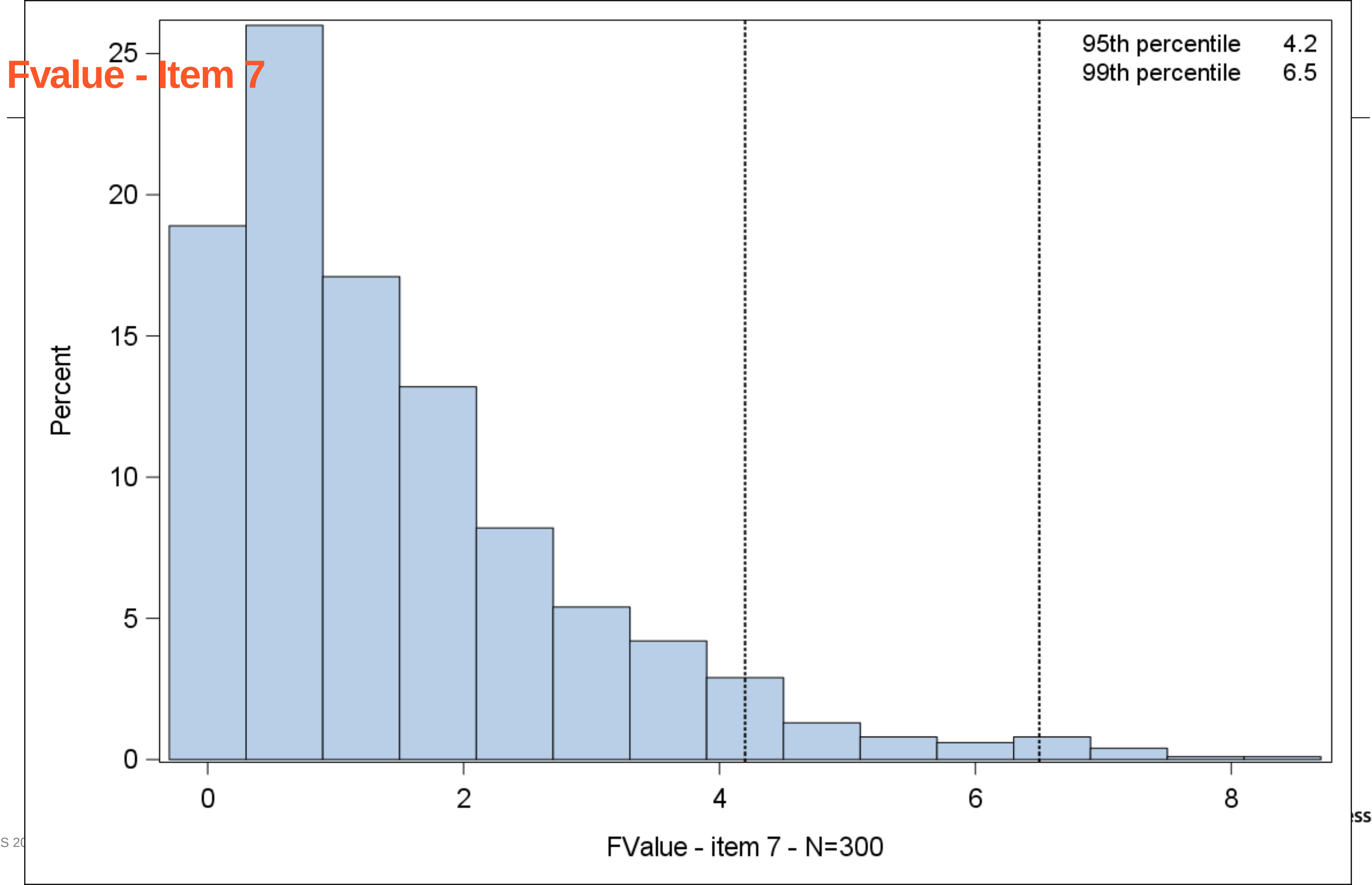
Fvalue - Item 5



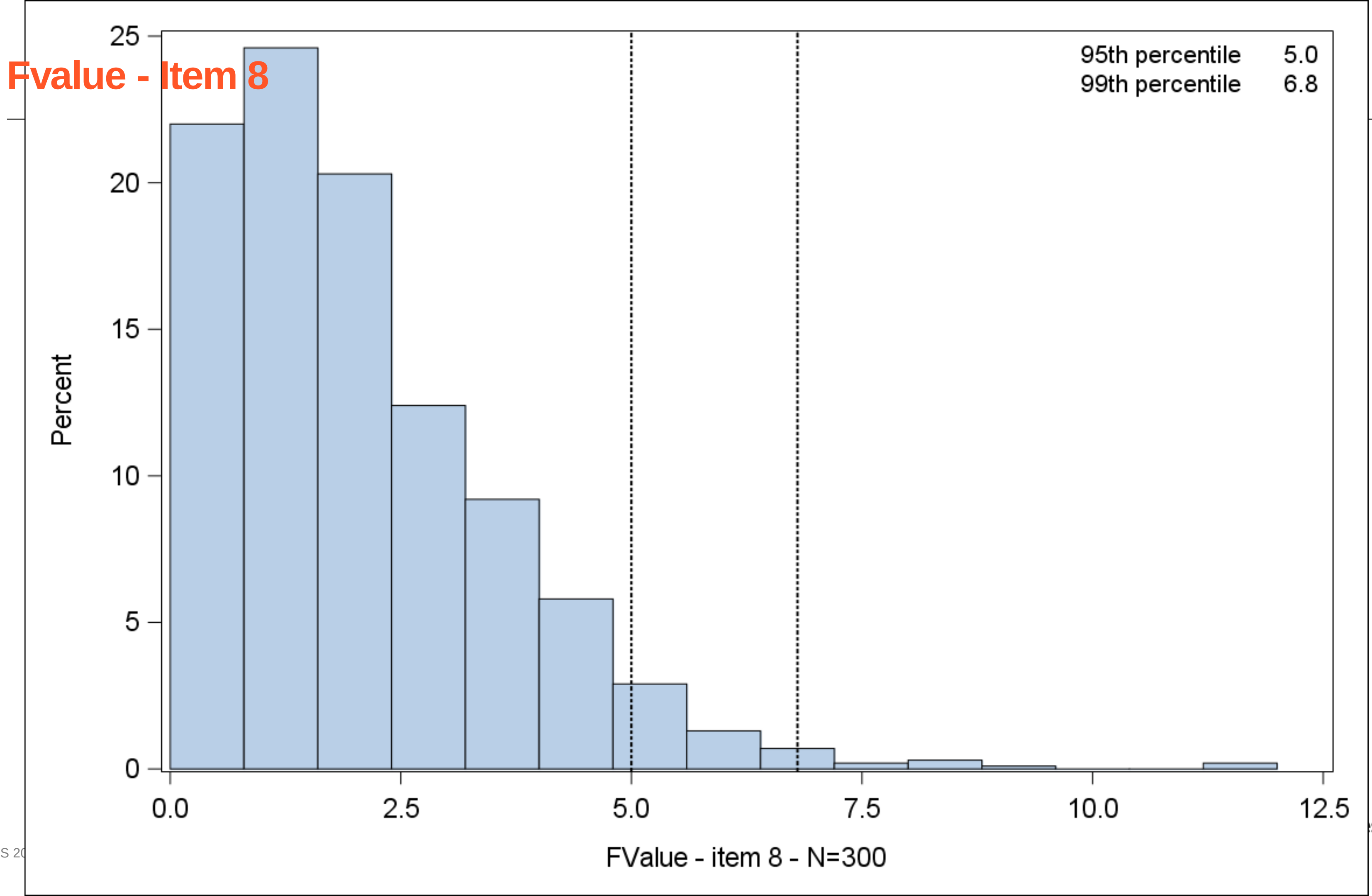
Fvalue - Item 6



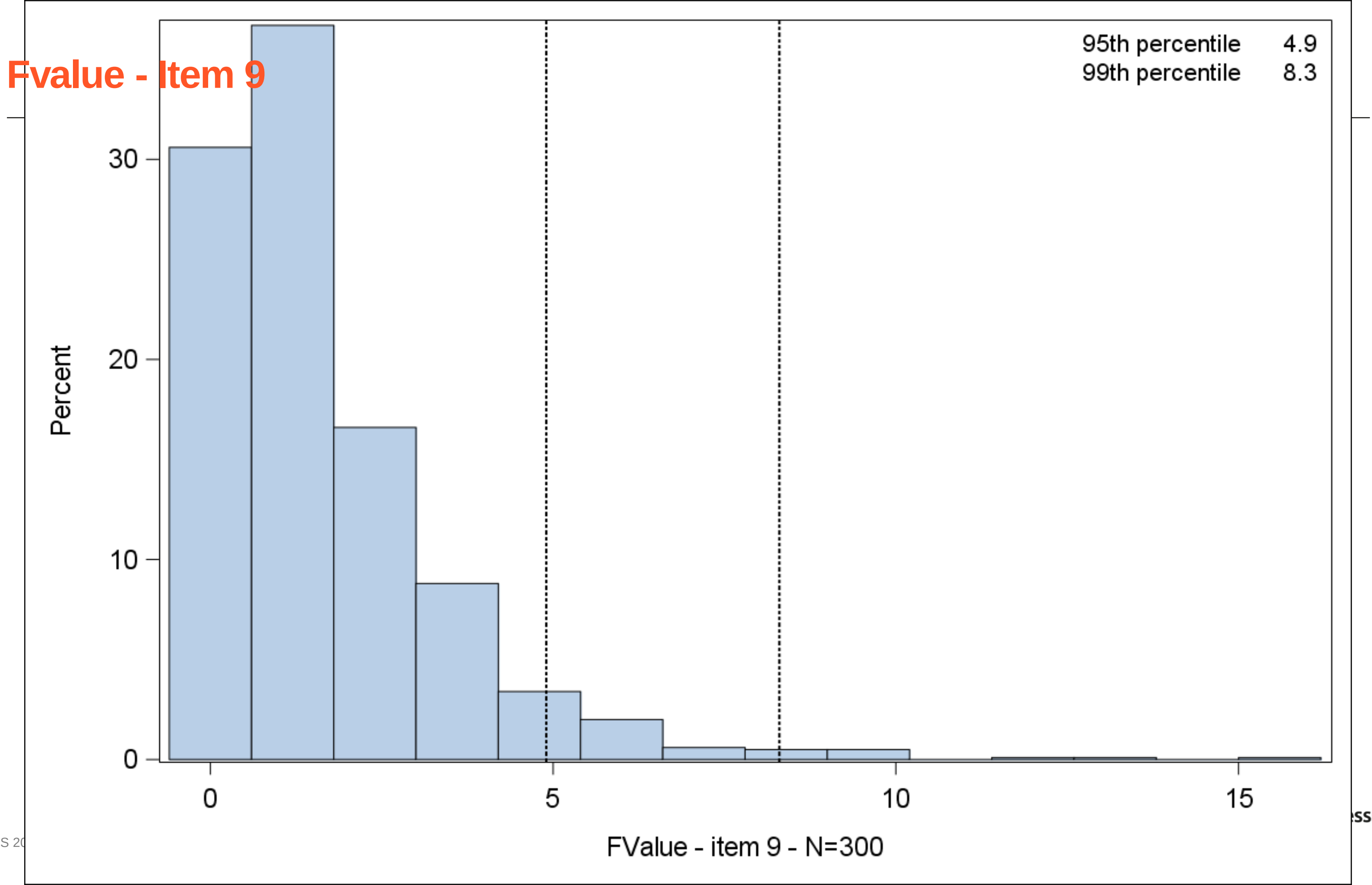
Fvalue - Item 7



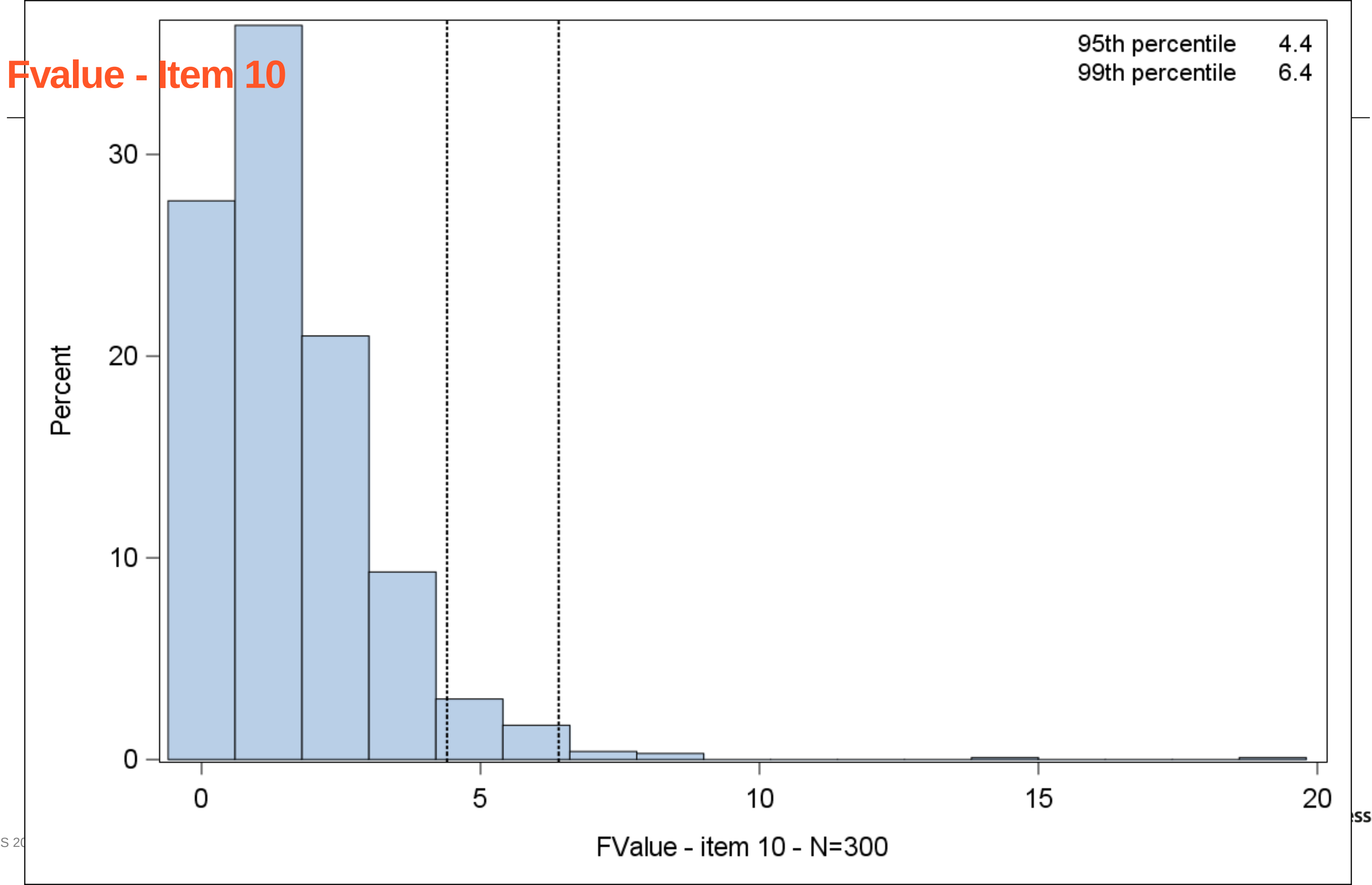
Fvalue - Item 8



Fvalue - Item 9



Fvalue - Item 10



ssment

Fvalue - Summary

item	Mean	(SD)	95th	99th
1	1.8	(1.6)	5.1	7.4
2	1.4	(1.5)	4.4	6.7
3	1.6	(1.7)	4.9	8.2
4	1.0	(1.1)	2.8	5.1
5	1.5	(1.5)	4.6	6.8
6	1.6	(1.6)	4.8	8.1
7	1.5	(1.4)	4.2	6.5
8	2.1	(1.6)	5.0	6.8
9	1.6	(1.7)	4.9	8.3
10	1.6	(1.6)	4.4	6.4

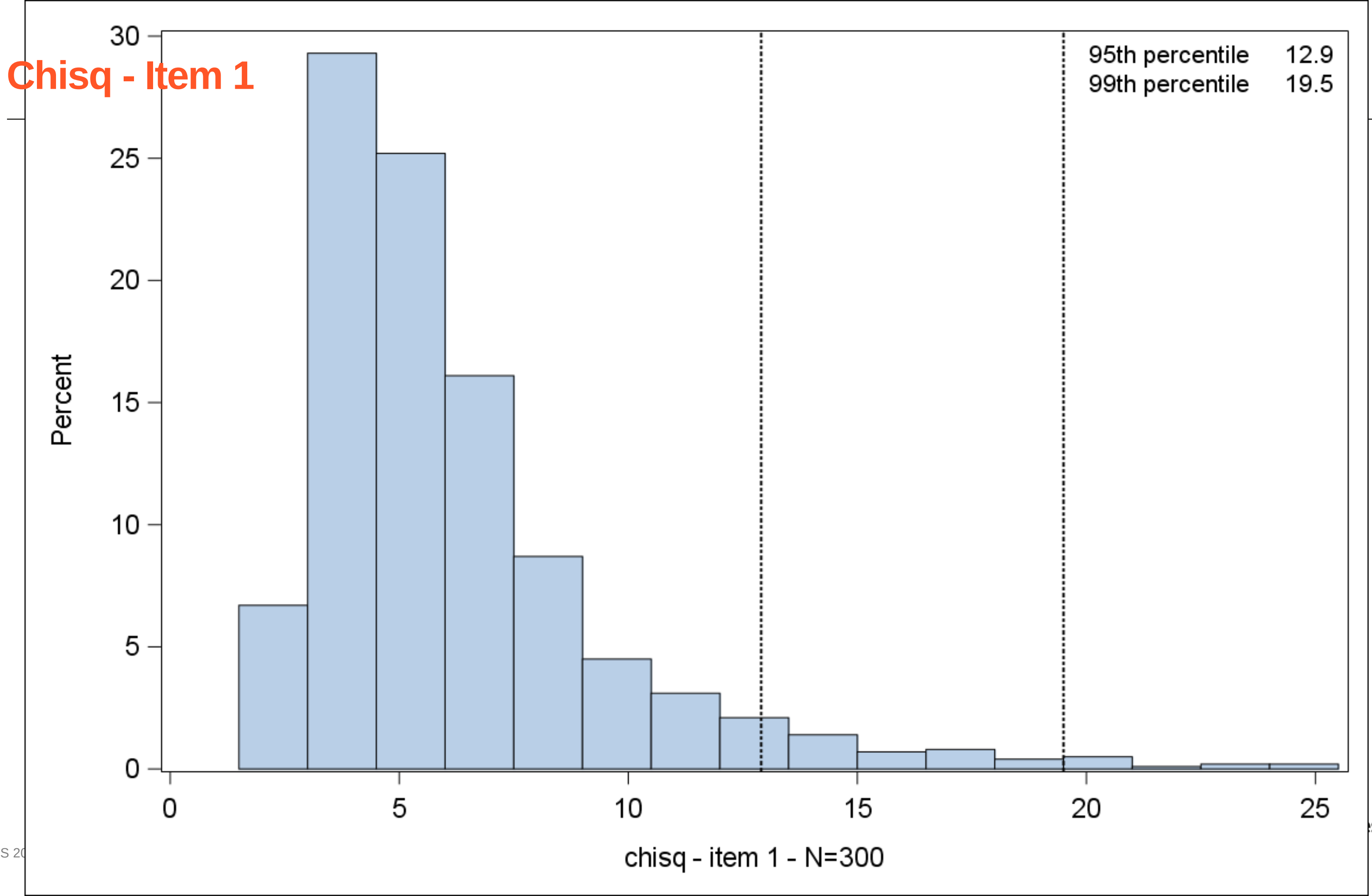
No single critical value is appropriate

The assumption that the item fit statistic has an F-distribution with (2, N-3) degrees of freedom is incorrect.

[Critical value for a test at the 5% level is somewhere around 3.0]

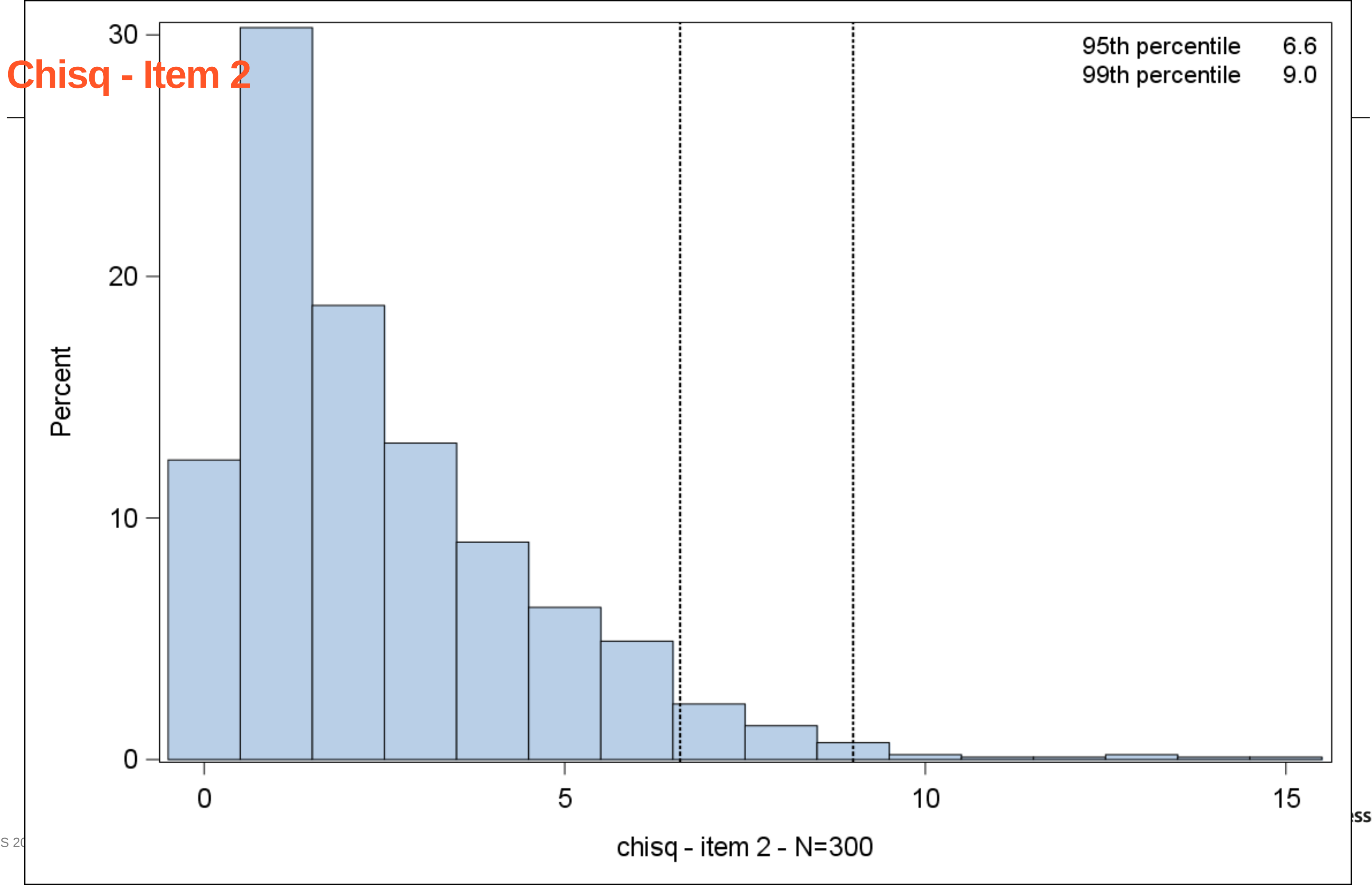
Looks like the F-test will very often reject item fit.

Chisq - Item 1

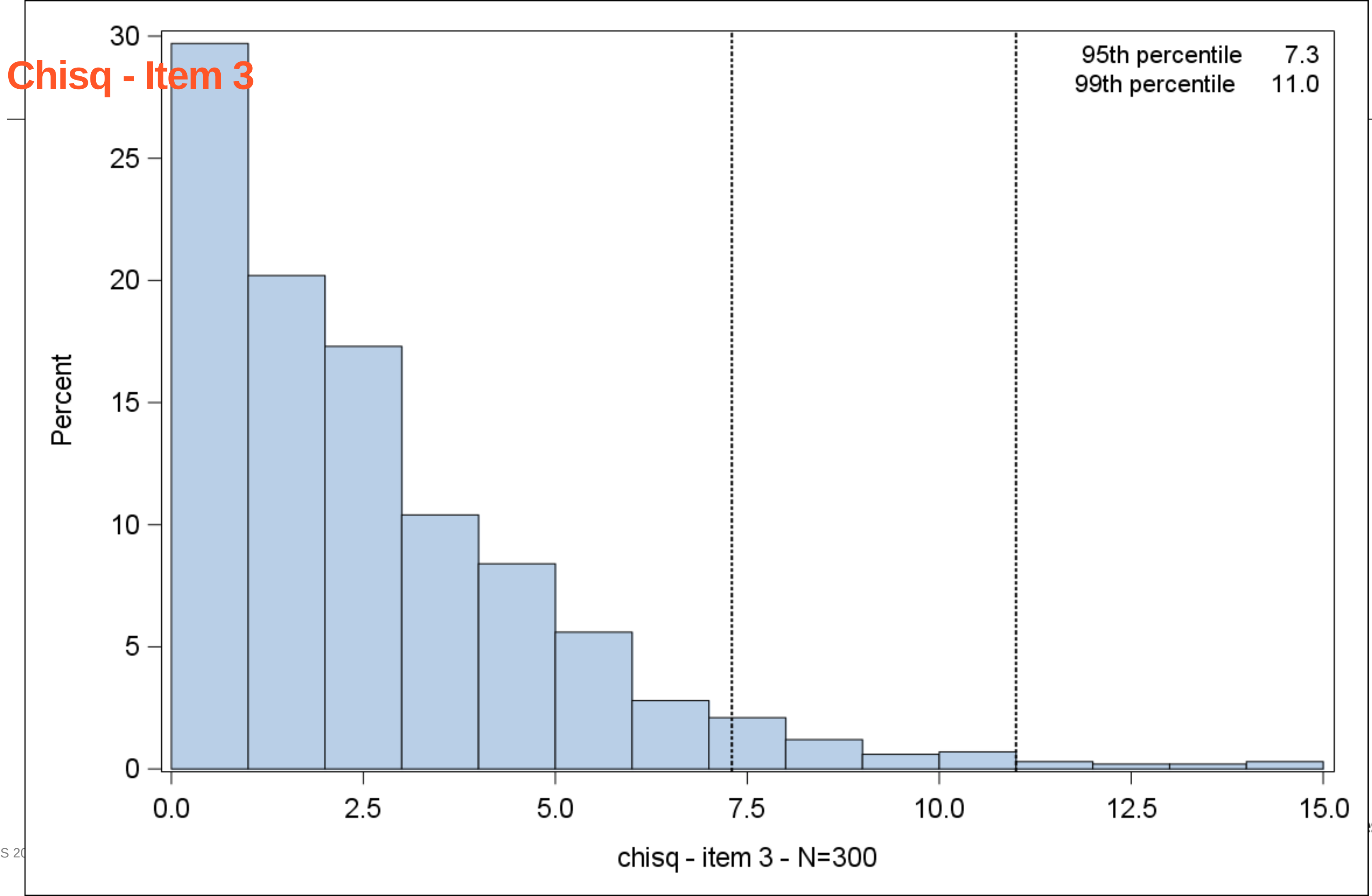


30
25
20
15
10
5
0
Percent
0
5
10
15
20
25
chisq - item 1 - N=300
95th percentile 12.9
99th percentile 19.5

Chisq - Item 2

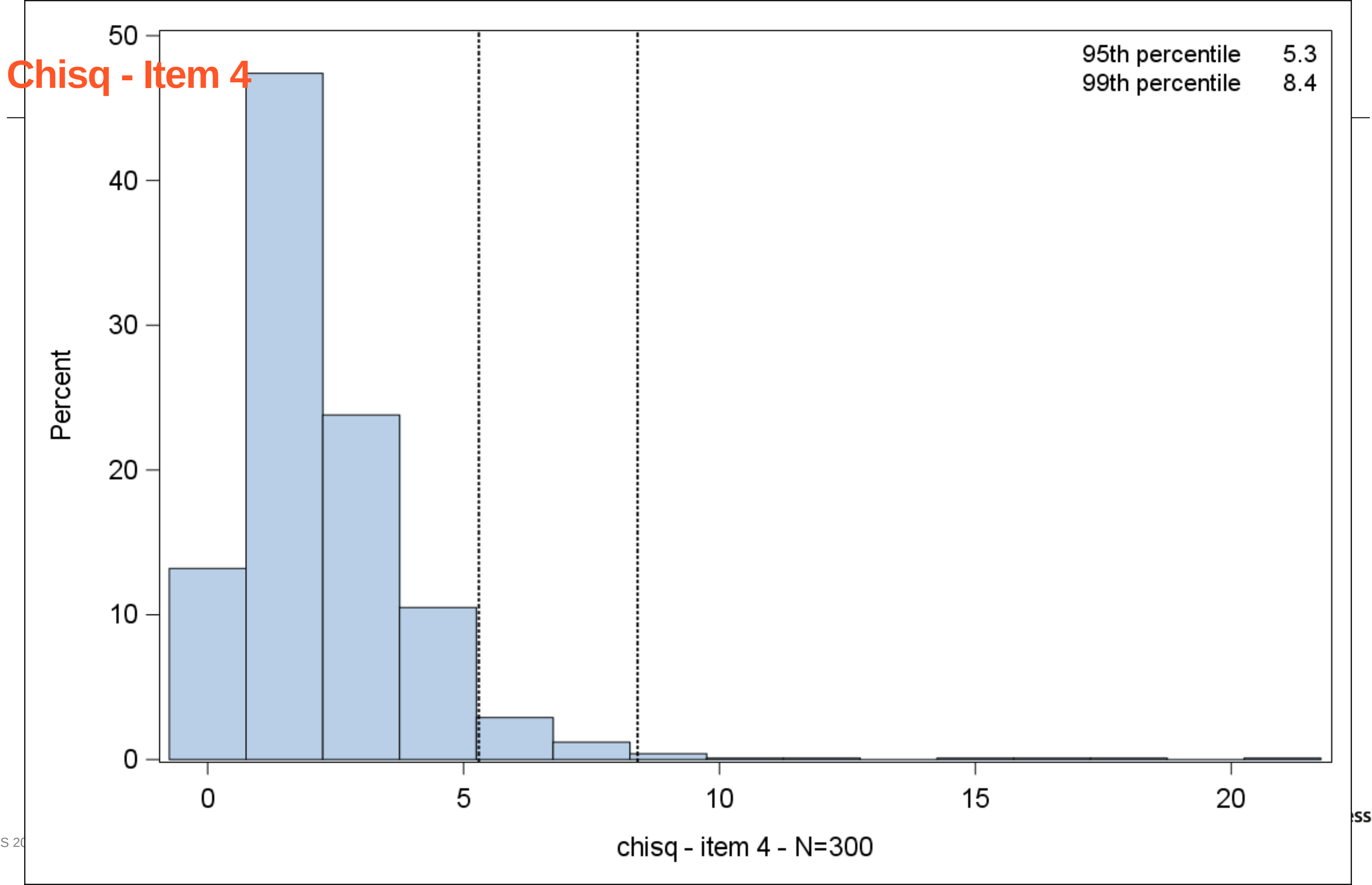

95th percentile 6.6
99th percentile 9.0
30
20
10
0
Percent
0
5
10
15
chisq - item 2 - N=300

Chisq - Item 3

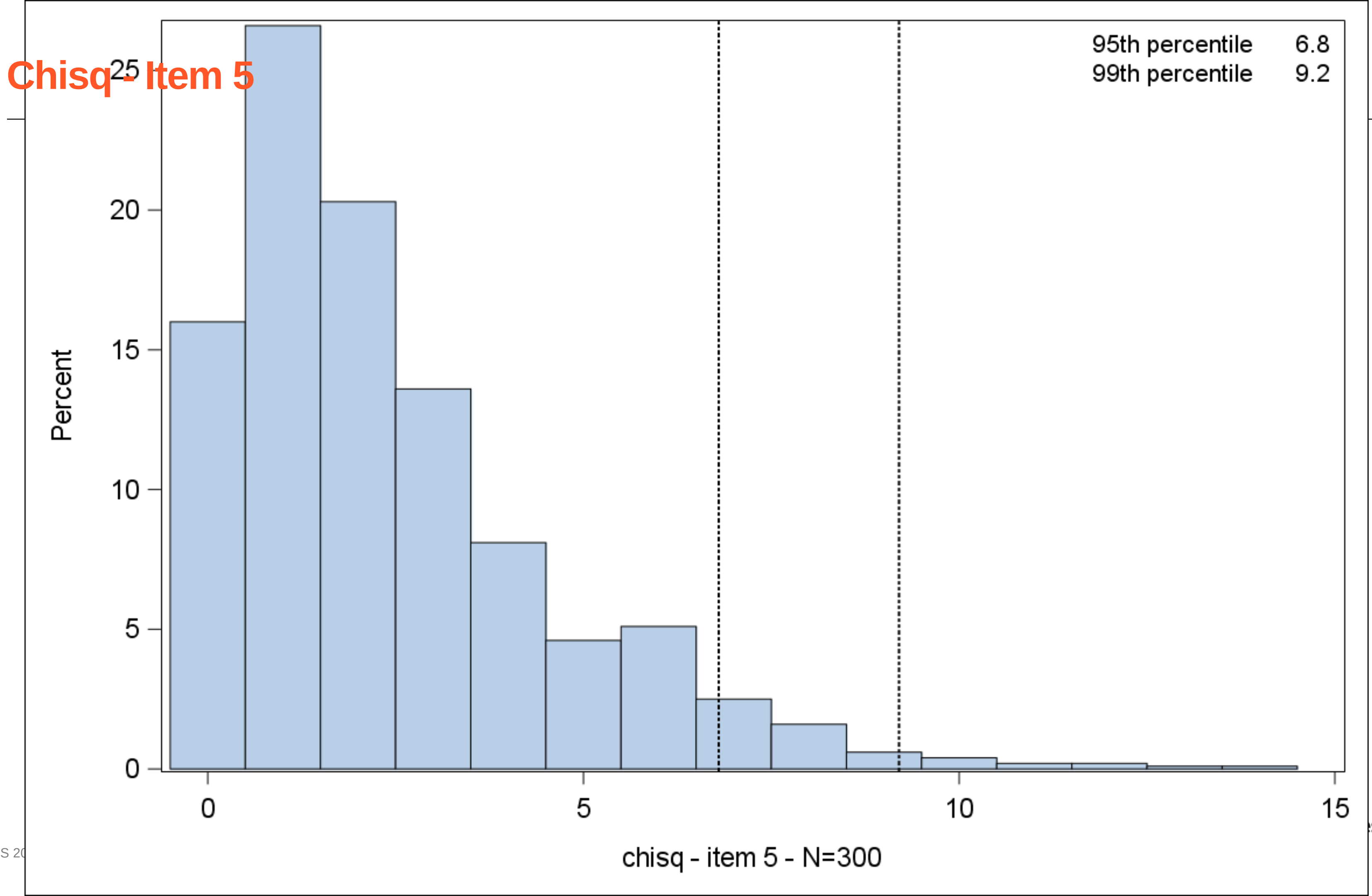


95th percentile 7.3
99th percentile 11.0
Percent
30
25
20
15
10
5
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
chisq - item 3 - N=300

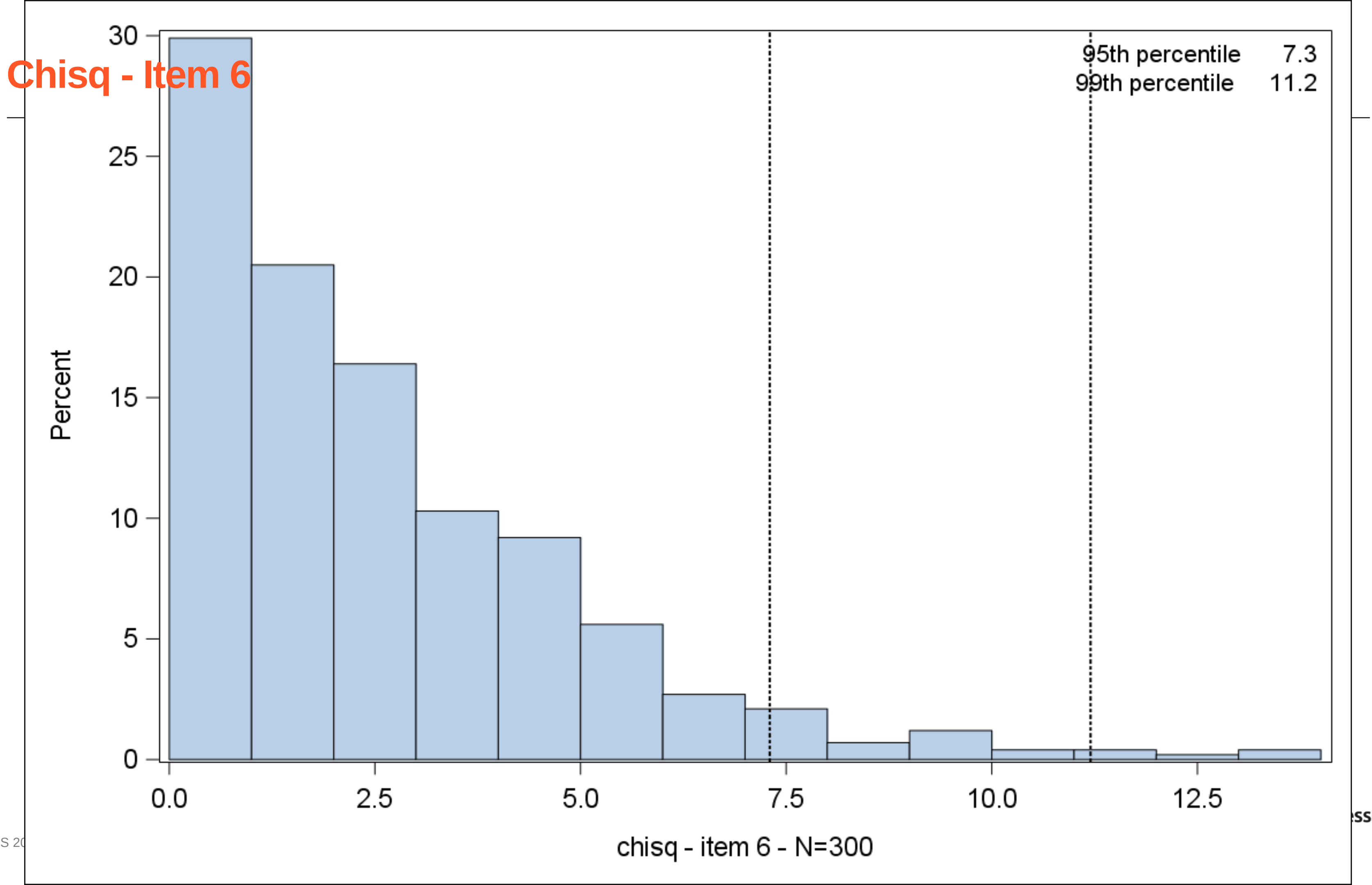
Chisq - Item 4


95th percentile 5.3
99th percentile 8.4
Percent
50
40
30
20
10
0
0
5
10
15
20
chisq - item 4 - N=300

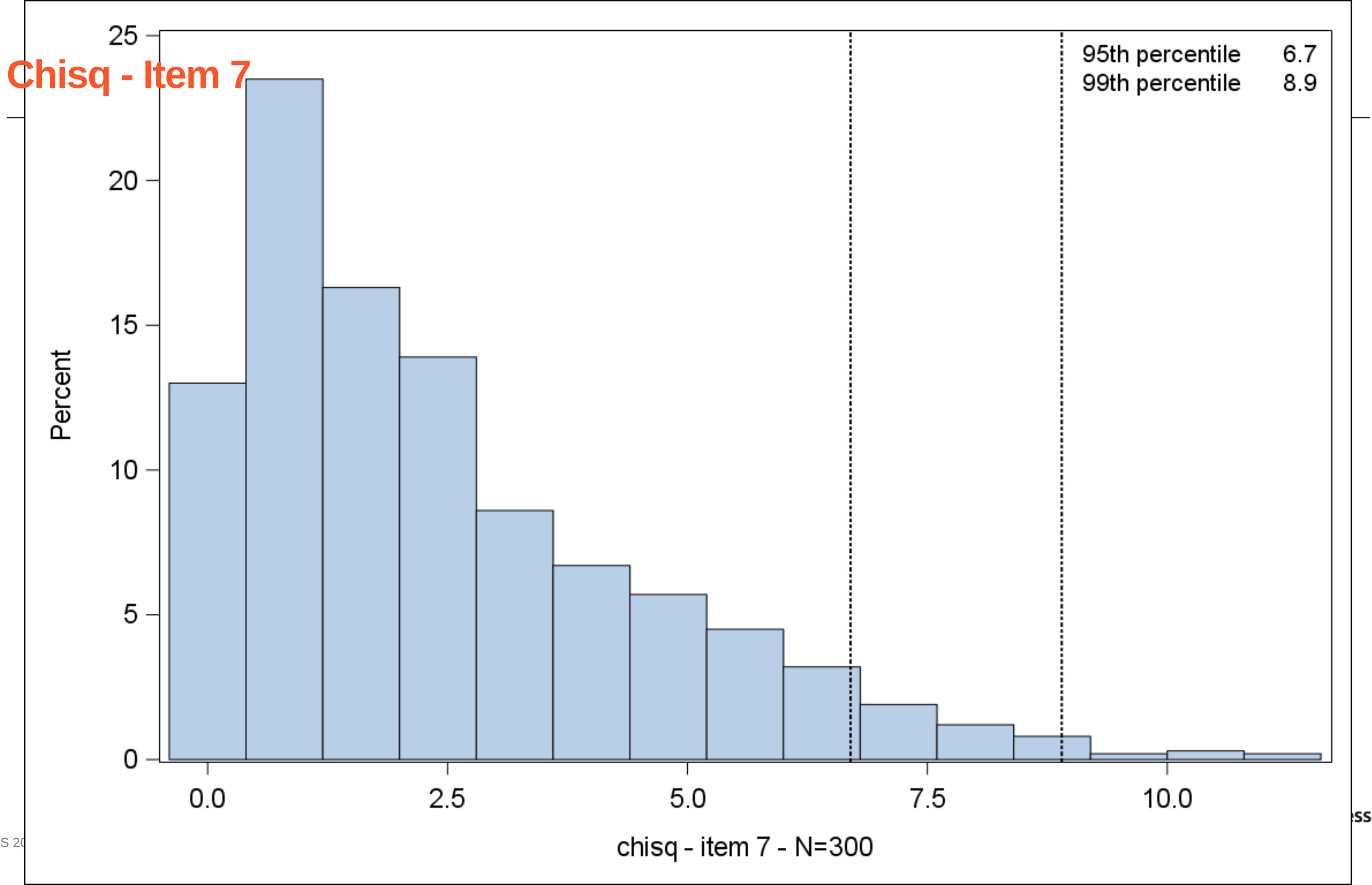
Chisq - Item 5


95th percentile 6.8
99th percentile 9.2
Percent
0
5
10
15
20
25
0
5
10
15
chisq - item 5 - N=300

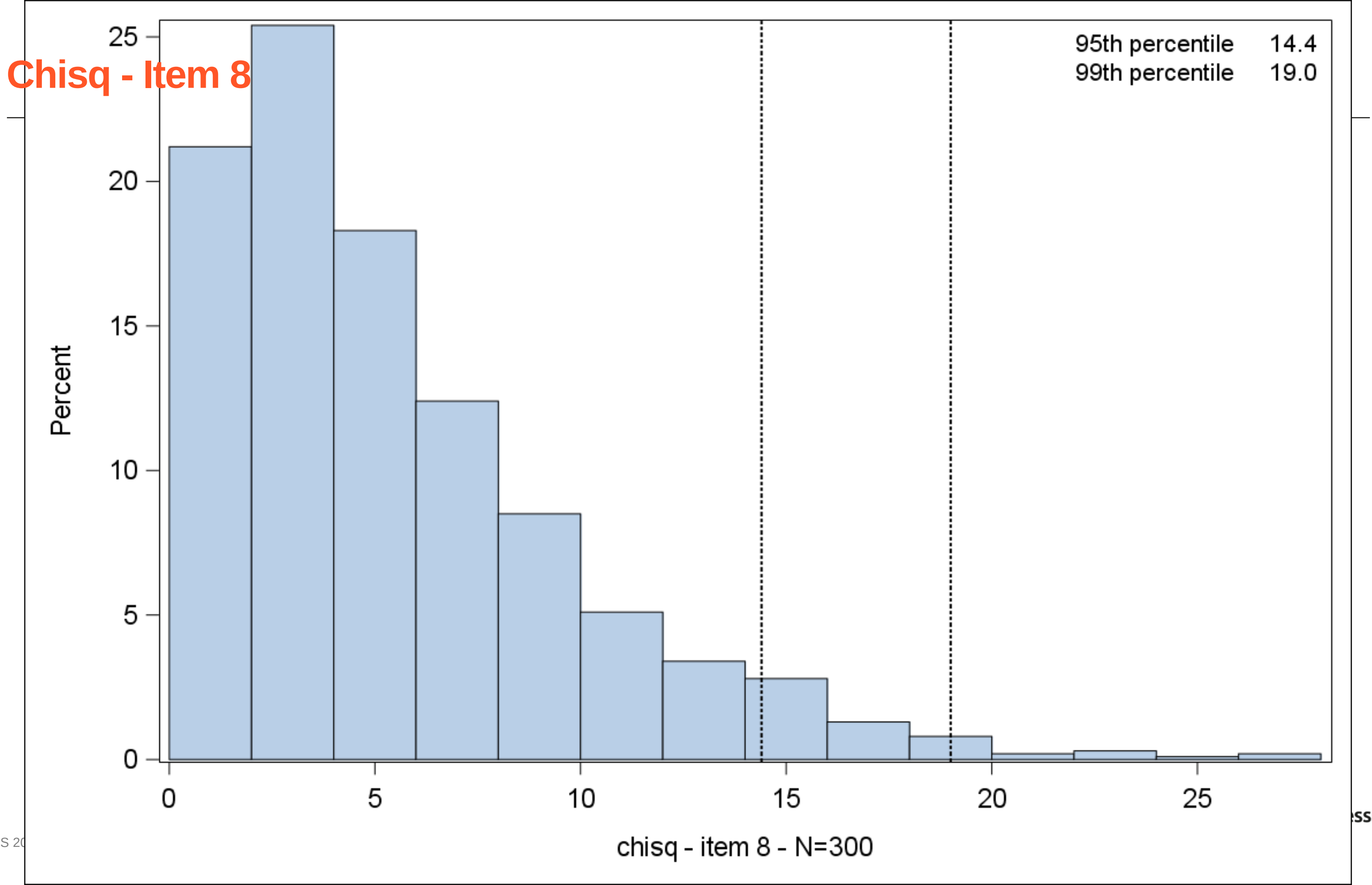
Chisq - Item 6


95th percentile 7.3
99th percentile 11.2
Percent
30
25
20
15
10
5
0
0.0
2.5
5.0
7.5
10.0
12.5
chisq - item 6 - N=300

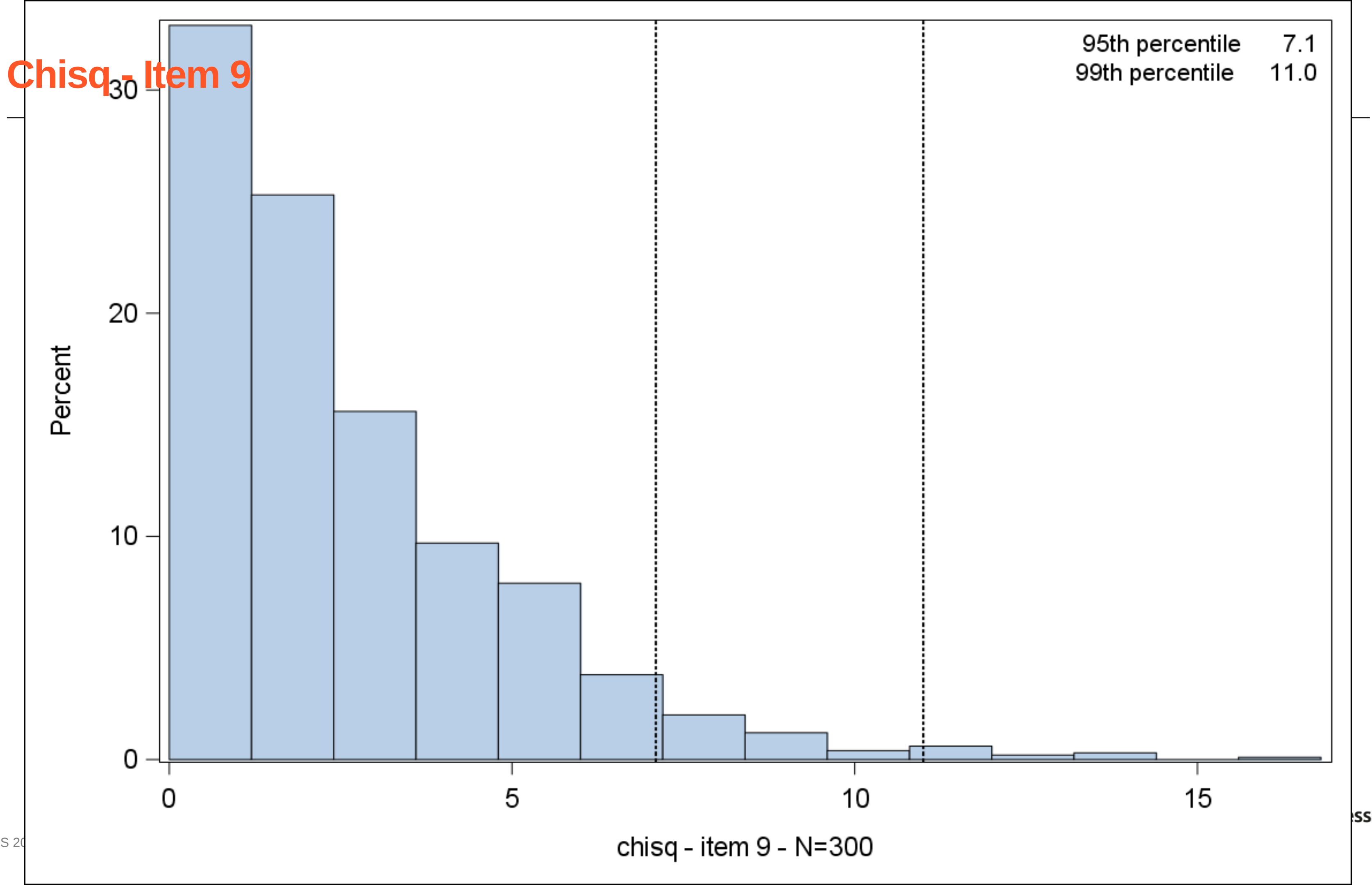
Chisq - Item 7


95th percentile 6.7
99th percentile 8.9
Percent
0
5
10
15
20
25
0.0
2.5
5.0
7.5
10.0
chisq - item 7 - N=300

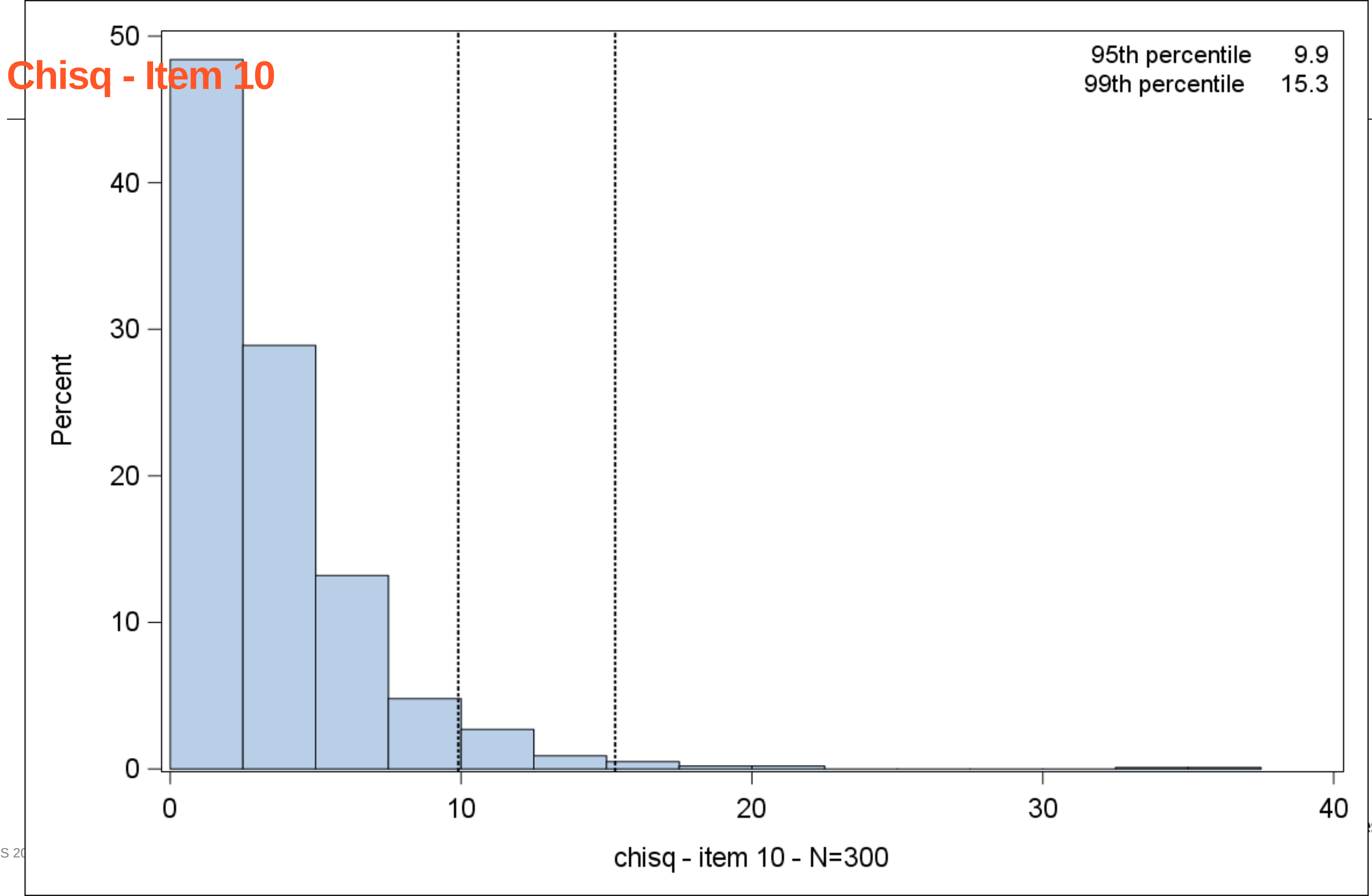
Chisq - Item 8


95th percentile 14.4
99th percentile 19.0
Percent
0
5
10
15
20
25
0
5
10
15
20
25
chisq - item 8 - N=300

Chisq - Item 9


95th percentile 7.1
99th percentile 11.0
Percent
0
10
20
30
0
5
10
15
chisq - item 9 - N=300

Chisq - Item 10



Chisq - Summary

item	Mean	(SD)	95th	99th
1	6.2	(3.3)	12.9	19.5
2	2.5	(2.1)	6.6	9.0
3	2.6	(2.4)	7.3	11.0
4	2.3	(1.9)	5.3	8.4
5	2.4	(2.2)	6.8	9.2
6	2.6	(2.3)	7.3	11.2
7	2.5	(2.1)	6.7	8.9
8	5.5	(4.3)	14.4	19.0
9	2.6	(2.4)	7.1	11.0
10	3.5	(3.4)	9.9	15.3

No single critical value is appropriate

The assumption that the item fit statistic has a chi-squared distribution with 2 degrees of freedom is incorrect.

Critical value for a test at the 5% level is 5.99 – this will lead to inflated type I error rates.

Summary

We do a Rasch analysis to identify anomalies. If we have, say, 10 items and no pre-specified hypothesis about a misfitting item there is an inherent risk of type I errors

No justification for saying that FitResid>2.5 or P<0.05 constitues an anomaly

Bonferroni adjustment is only correct if the P-values are correct. We have no way of adjusting for multiple testing.

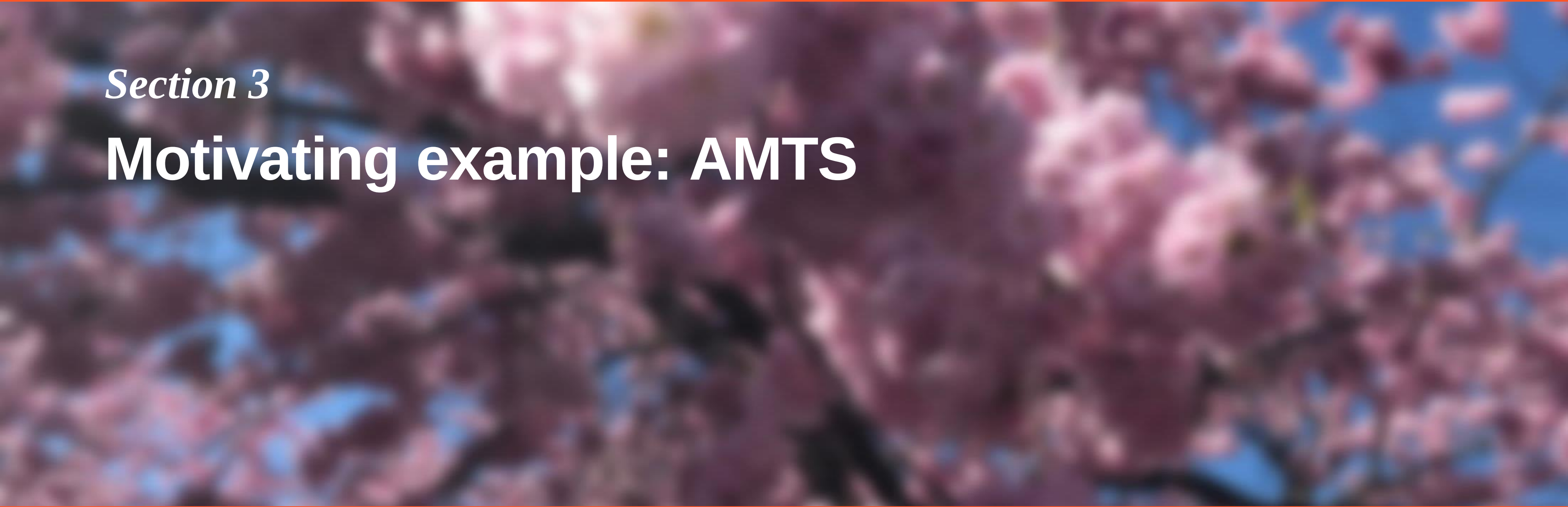
How did we end up in this mess ?

None of the item fit statistics distinguish between the true person location value and the estimated person location value.

No theoretical justification for the claims about the asymptotic distribution.

The item fit statistics are used only because they are implemented in WINSTEPS and RUMM2030.

Item fit statistics likely to have correct asymptotic distributions have been proposed [Wright, Panchapakesan, 1969; …; Christensen, Kreiner, 2013]



Section 3

Motivating example: AMTS

Real data example: Abbr. Mental Test Score (AMTS)

Item	Loc	FitResid		ChiSq	DF	Prob	F-stat	DF-1	DF-2	Prob
1	-0,7	-1,65		4,824	2	0,08964	3,765	2	143	0,025504
2	0,1	-0,09		0,477	2	0,78784	0,082	2	142	0,921527
3	1,9	1,02		4,083	2	0,12984	1,582	2	143	0,209222
4	-0,6	-1,15		1,608	2	0,44754	0,955	2	143	0,387131
5	0,2	-1,78		6,717	2	0,03479	4,500	2	143	0,012721
6	-1,7	-0,62		1,124	2	0,57014	0,238	2	143	0,788361
7	0,5	-3,34		16,213	2	0,00030	17,188	2	143	0,000002
8	-0,1	0,81		1,521	2	0,46746	0,486	2	143	0,615916
9	0,2	-0,85		4,334	2	0,11450	2,425	2	143	0,092114
10	0,4	1,80		1,466	2	0,48057	0,754	2	143	0,472505

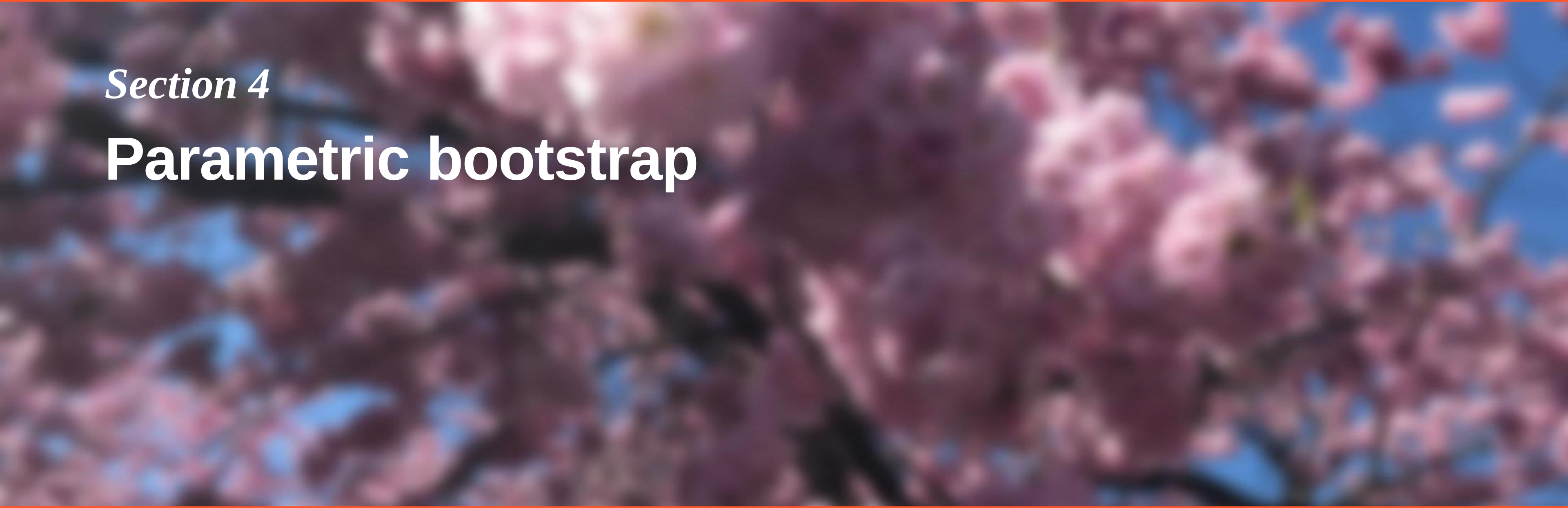
How can we know what constitues an anomaly

Have no prespecified hypothesis pointing out a single item

Use parametric bootstrap [Su, Sheu, Wang, 2007; Wolfe, 2008; 2013; Seol, 2016]:

1. Simulate a data set where the Rasch model fits use sample size and item locations from AMTS example.
2. Calculate the 10 values of Chisq (one for each item)
3. Store the minimum and maximum values

Repeat 1000 times



Section 4

Parametric bootstrap

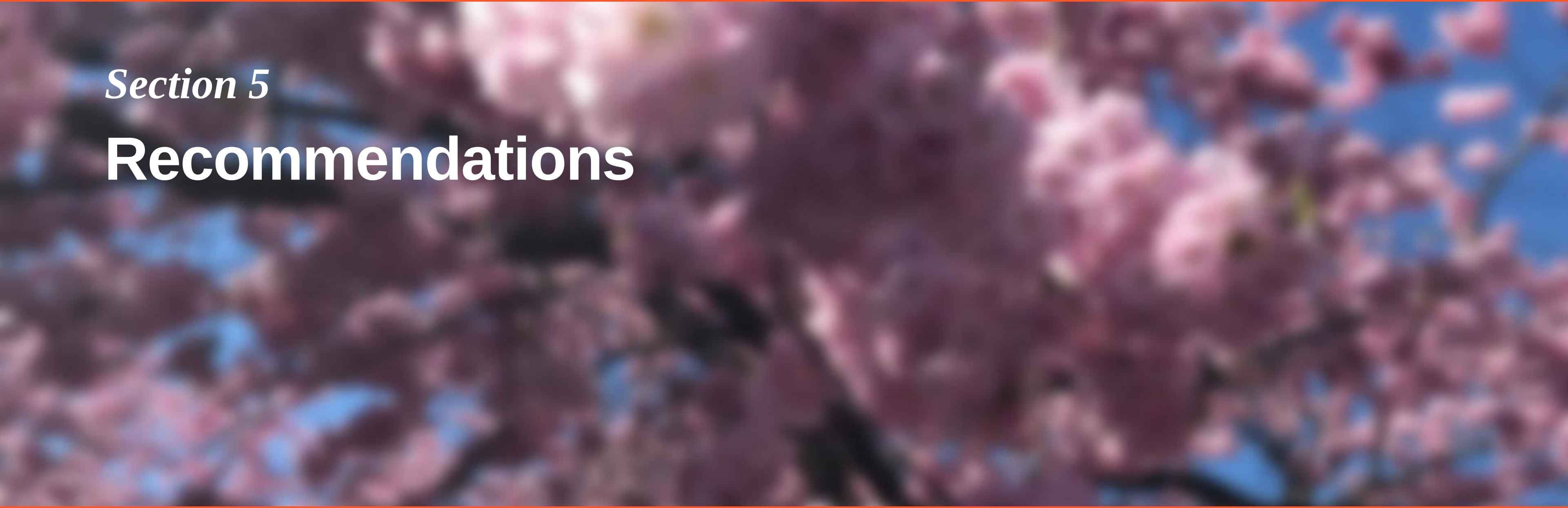
Minimum and maximum value of FitResid

	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Maximum
FitResid	1.7	0.6	0.7	1.0	2.9	3.7	4.7
FValue	4.5	2.0	1.7	2.1	8.4	11.8	19.3
Chisq	8.2	3.8	3.1	3.9	15.1	23.3	37.8

Observed minimum value min(FitResid) = -3,34 is clearly an anomaly

Observed max(Fvalue) = 17,188 is clearly an anomaly

Observed max(Chisq) = 16,213 is also an anomaly



Section 5

Recommendations

Recommendations

Interpret fit statistics with caution.

Look out for our simulation studies aimed at providing guidelines (work in progress)

References

- Christensen, Kreiner (2013). Item Fit Statistics. https://doi.org/10.1002/9781118574454.ch5
- Hagell, Westergren (2016). *Journal of Applied Measurement*, 17, 416-431. http://www.ncbi.nlm.nih.gov/pubmed/28009589
- Seol (2016). *Psychological Reports*, 118, 937-956. https://doi.org/10.1177/0033294116649434
- Smith et al (2008). *BMC Medical Research Methodology,* 8:33 https://doi.org/10.1186/1471-2288-8-33
- Smith, Schumacker, Bush (1998). *Journal of Outcome Measurement*, 2 , 66-78. https://www.ncbi.nlm.nih.gov/pubmed/9661732
- Su, Sheu, Wang, (2007). *Journal of Applied Measurement*, 8, 190-203. https://www.ncbi.nlm.nih.gov/pubmed/17440261
- Wolfe (2008). Applied Psychological Measurement, 3, 585-586. https://doi.org/10.1177/0146621607312308
- Wolfe (2013). *Journal of applied measurement*, 14, 1-9. http://www.ncbi.nlm.nih.gov/pubmed/23442324
- Wright, Panchapakesan (1969). *Educational and Psychological Measurement*, 29, 23-48. https://doi.org/10.1177/001316446902900102

Thank you

You can see these slides on my homepage

http://biostat.ku.dk/~kach/