

Statistical Analysis of IOI Scoring

Gordon V. Cormack

24 January 2006

University of
Waterloo



What are IOI Tests *Meant* to Measure?

Achievement?

Skill?

Knowledge?

Strategy?

Mind Reading?

Luck?

Testing Theory

Precision

How much of the measurement is chance?

Reliability

Same test, same subjects, same result?

Accuracy

How close is measurement to the *true* value?

Validity (Internal Validity)

Does the quantity we measure reflect what we want to know?

Generalizability (External Validity; Transferability)

Does the test work for other subjects?

What *Do* IOI Tests Measure?

IOI 2006

Sum of scores

128 cases over 6 problems

64 “50% rule” tests

64 “non-50% rule” tests

More Abstract

Percentile Rank (0% - 100%)

Gold: 91.66% -

Silver: 75% - 91.66%

Bronze: 50% - 75%

Empty Handed: 0% - 50%

Does the ranking measure chance?

Method:

Repeat (materially) same test several times

Measure variance, standard deviation

Predict 95% Confidence Intervals

If the test were repeated 100 times result would be expected to fall in the interval 95 times

Difficulty:

Can't repeat the test

Logistics & resources

Participants learn (poor reliability)

Statistical Model

Infinite Hypothetical Population of tests & results

Our Model

Imaginary infinite pool of test cases having identical distribution of scores when applied to the contestants' submissions.

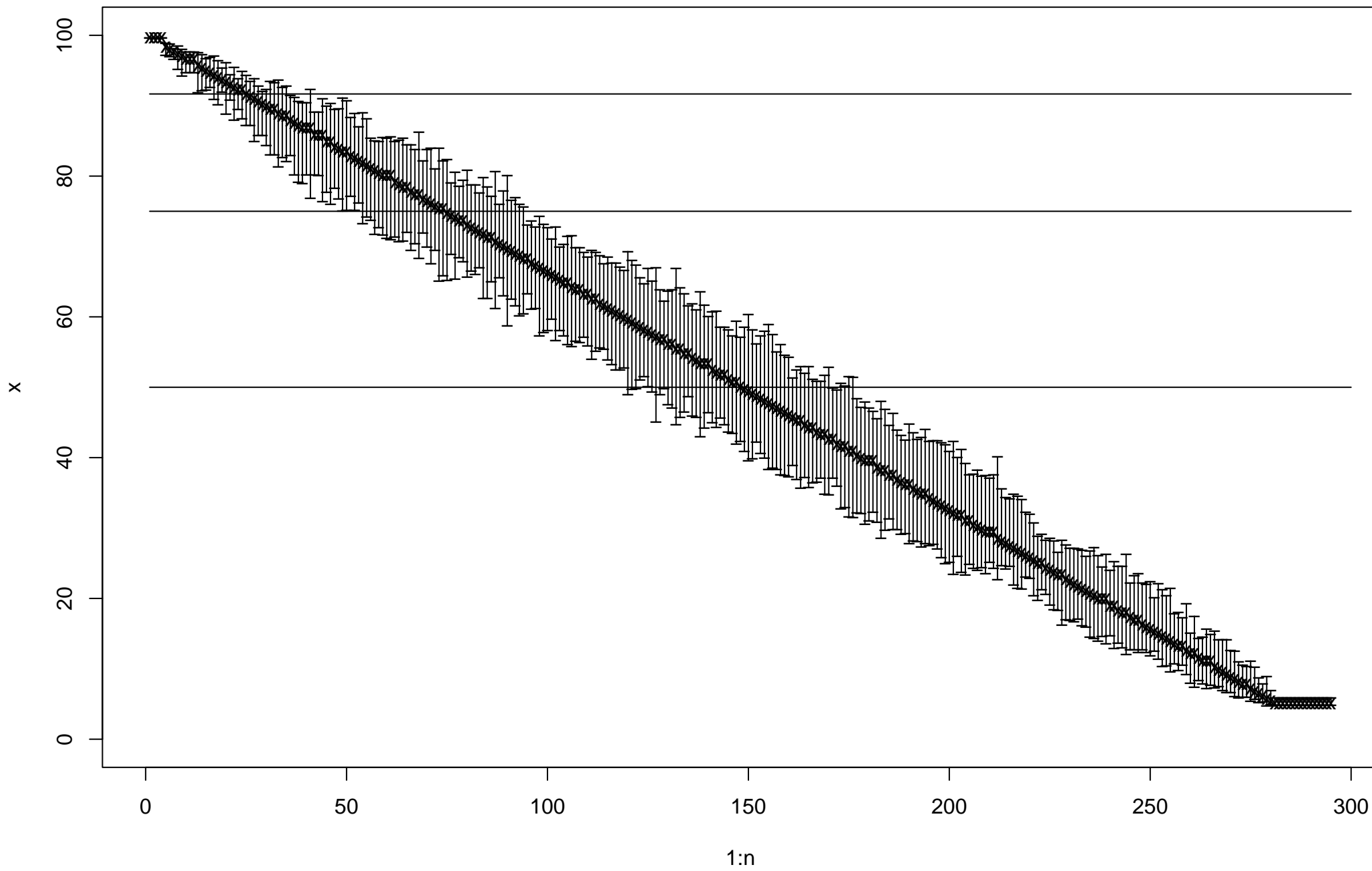
The Bootstrap

n imaginary 128-test-case IOIs

Test cases (and results) drawn *with replacement* from the actual cases.

Raw Results

Contestant	Percentile	C. I.
Can01	69.26	62.51 – 76.56
Can02	77.36	68.29 – 86.19
Can03	93.92	90.11 – 96.41
Can04	61.15	53.87 – 68.49



By Eye?

Average Difference in Scores (Ranks)?

Average Absolute Difference in Scores (Ranks)?

Inversions?

A ahead of B vs. B ahead of A?

Kendall's Tau Correlation

Insensitive to *Significance* of inversion

RMS (Root Mean Square) Difference in Ranks?

RMS Difference in Logit Ranks?

$\log (x / (100\% - x)$

models *probability* well in a linear space

converts to *odds* (easily multiplied)

then takes *logarithm* (easily added)

common in epidemiology

basis of *logistic regression*

equivalent differences:

1% - 2%

10% - 20%

98% - 99%

Difference Between Bootstraps & *Real* Results

RMS Difference of Logit ranks

0.189 (0.133 - .244)

Converted to *odds ratio*

1.207 (1.142 – 1.277)

Interpretation

A random repetition of the IOI may be expected to change ranks by about 20%, 95% of the time

Effect of 50% Rule (FPR)

Eliminate *half* (i.e. Reduce FPR cases to 33.3%)

Mean OR: 1.02

Double (i.e. Increase FPR cases to 66.7%)

Mean OR: 1.02

Eliminate *all* (i.e. No FPR cases)

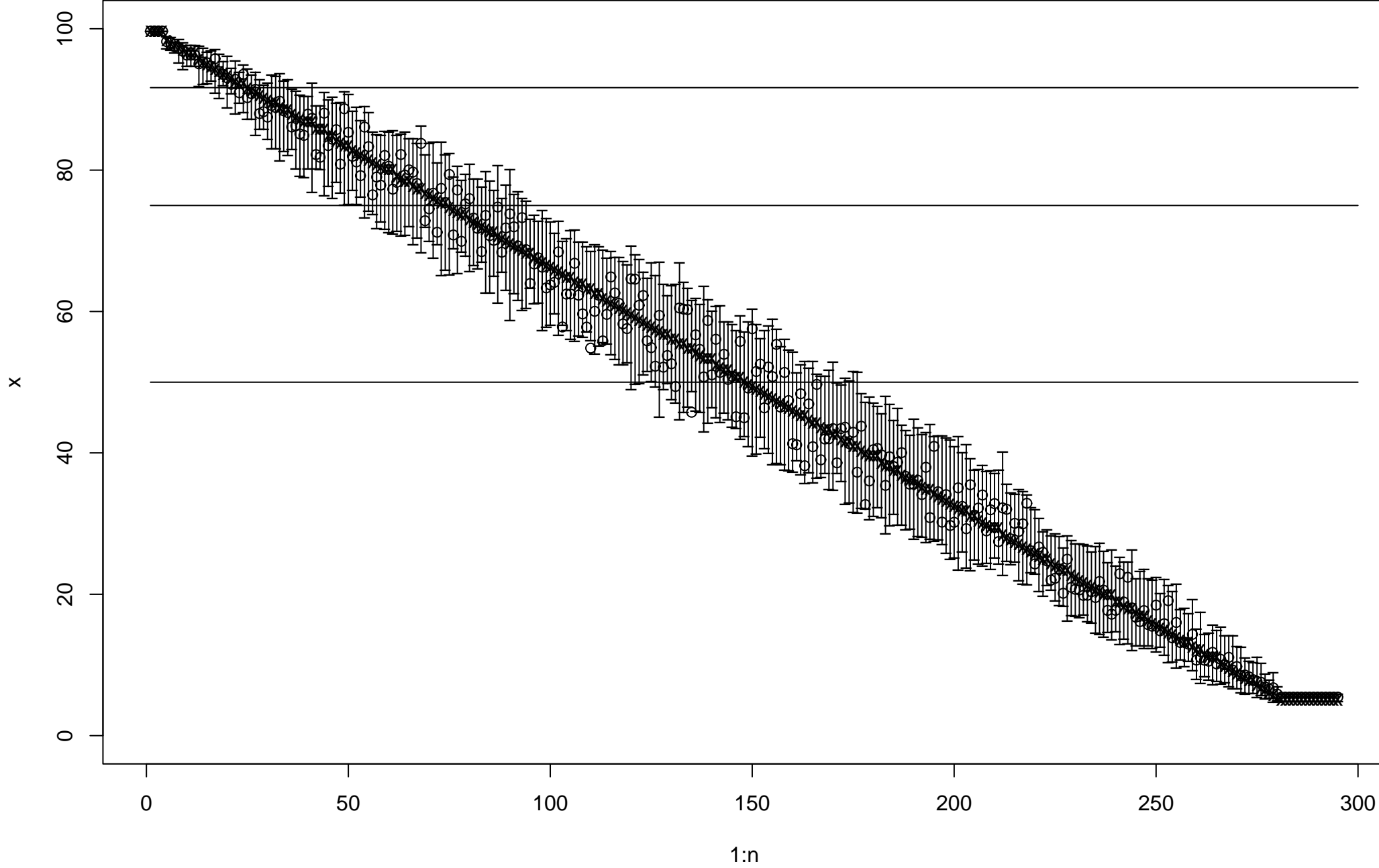
Mean OR: 1.31

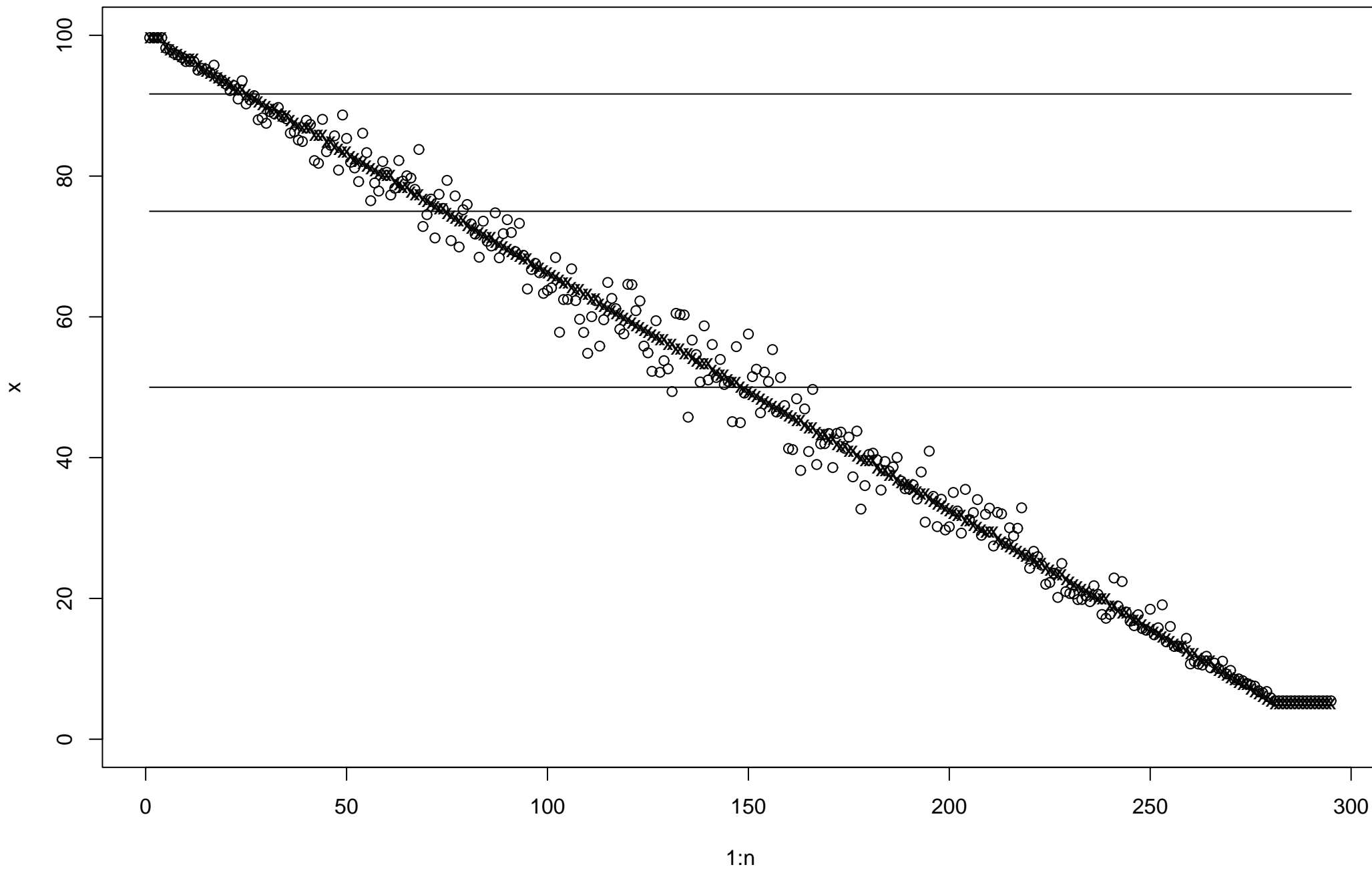
Eliminate *all but* (i.e. Only FPR cases)

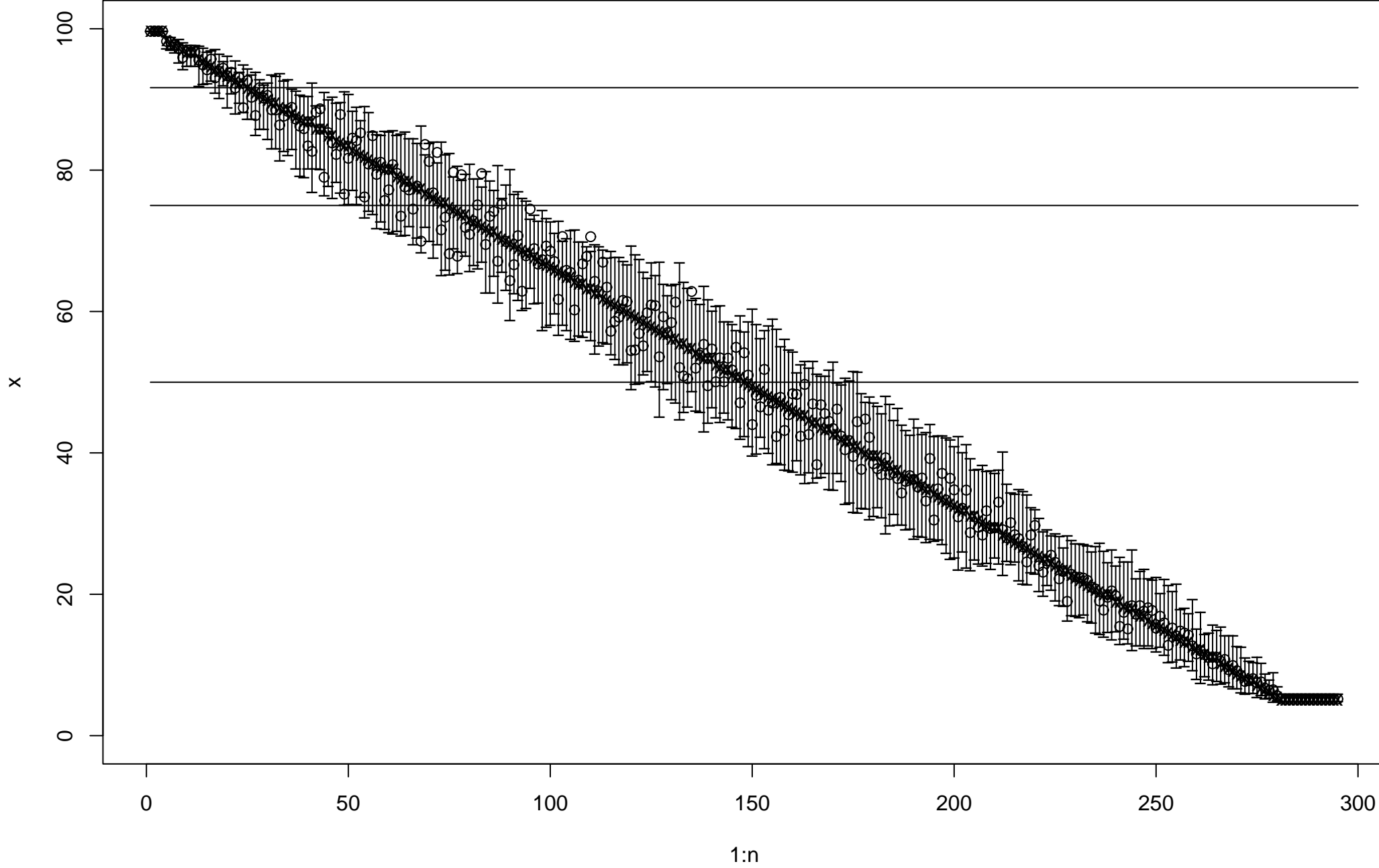
Mean OR: 2.11

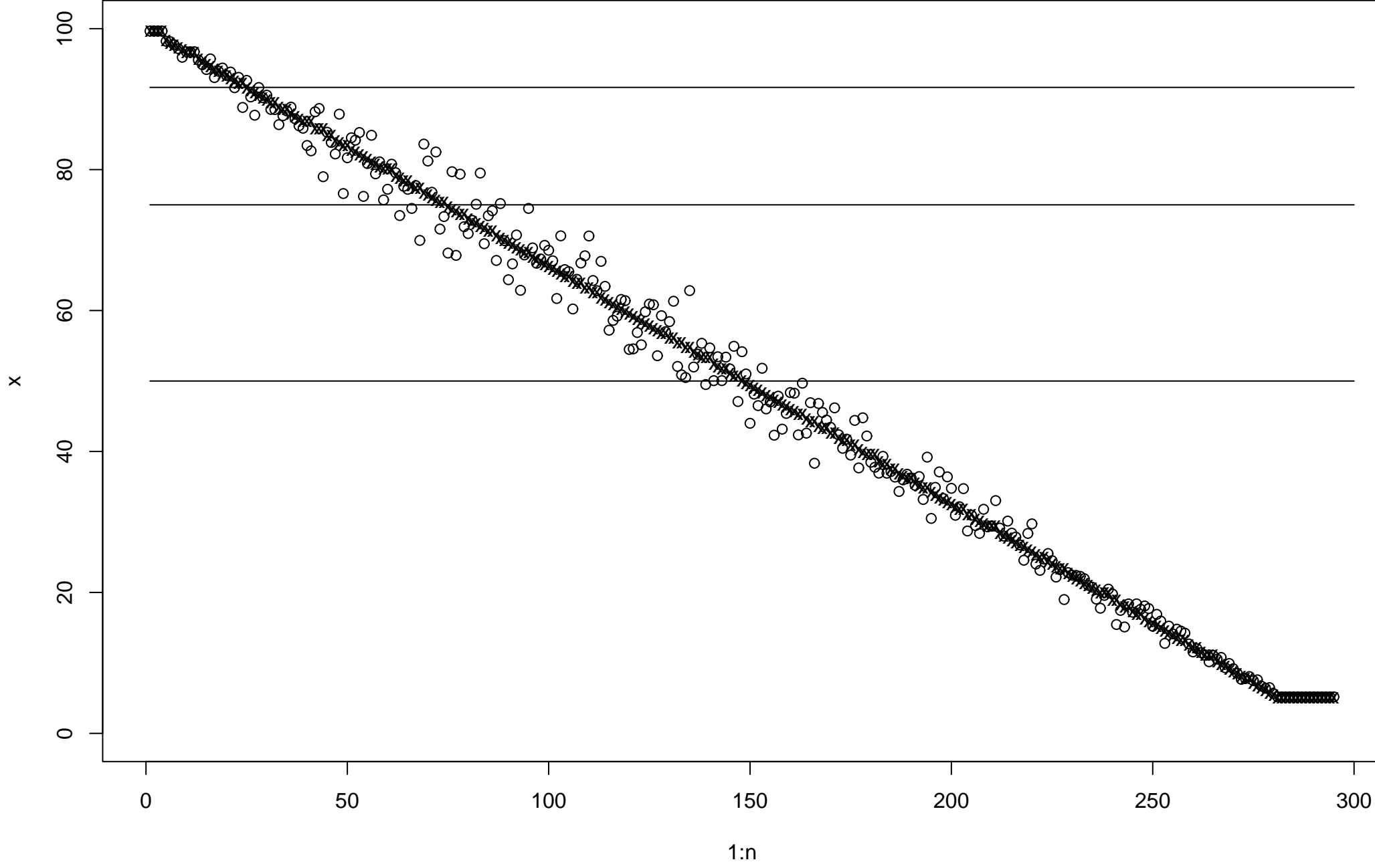
Eliminate 75% of FPR or 75% of non-FPR

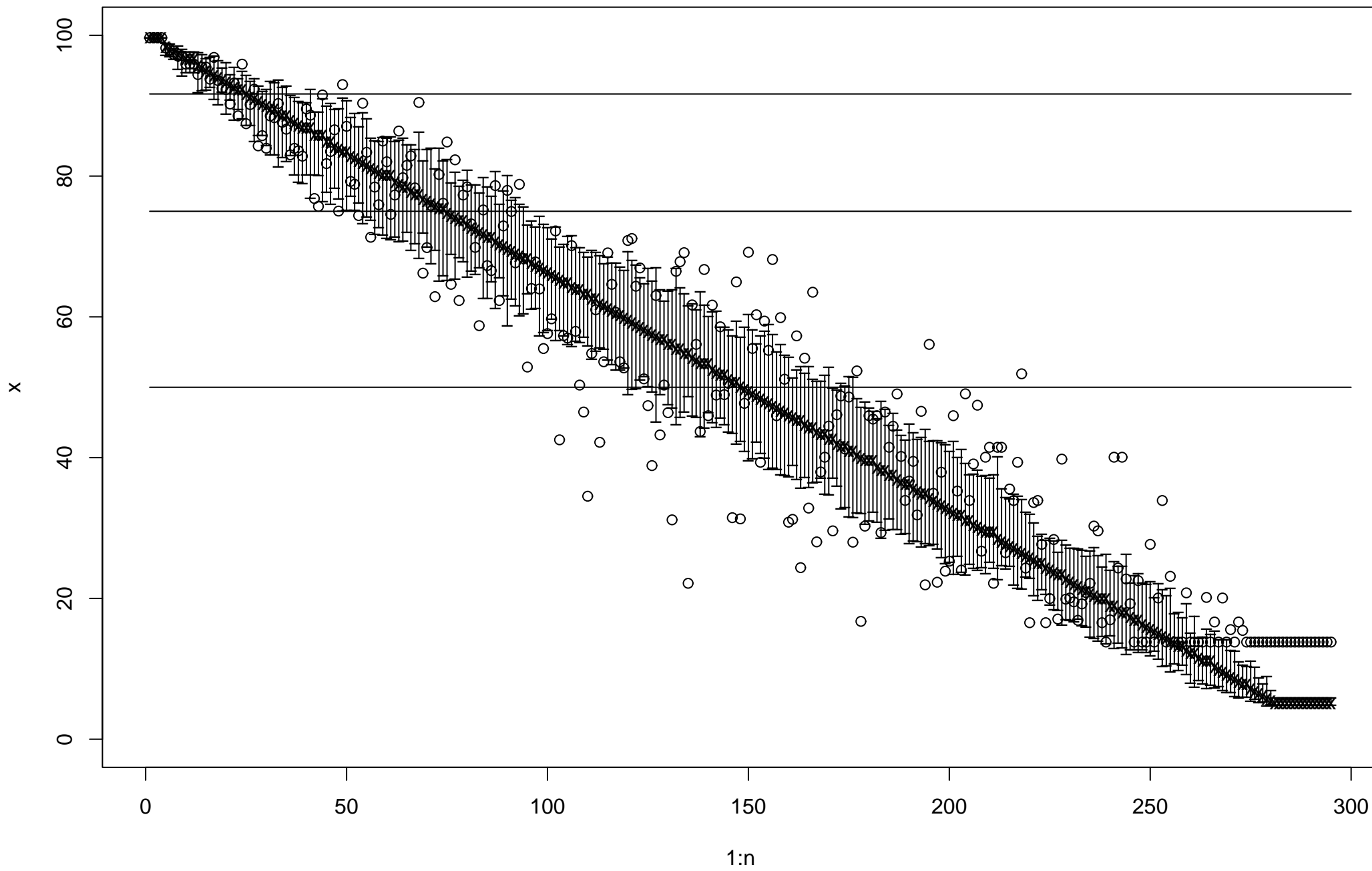
Mean OR: 1.06 in either case

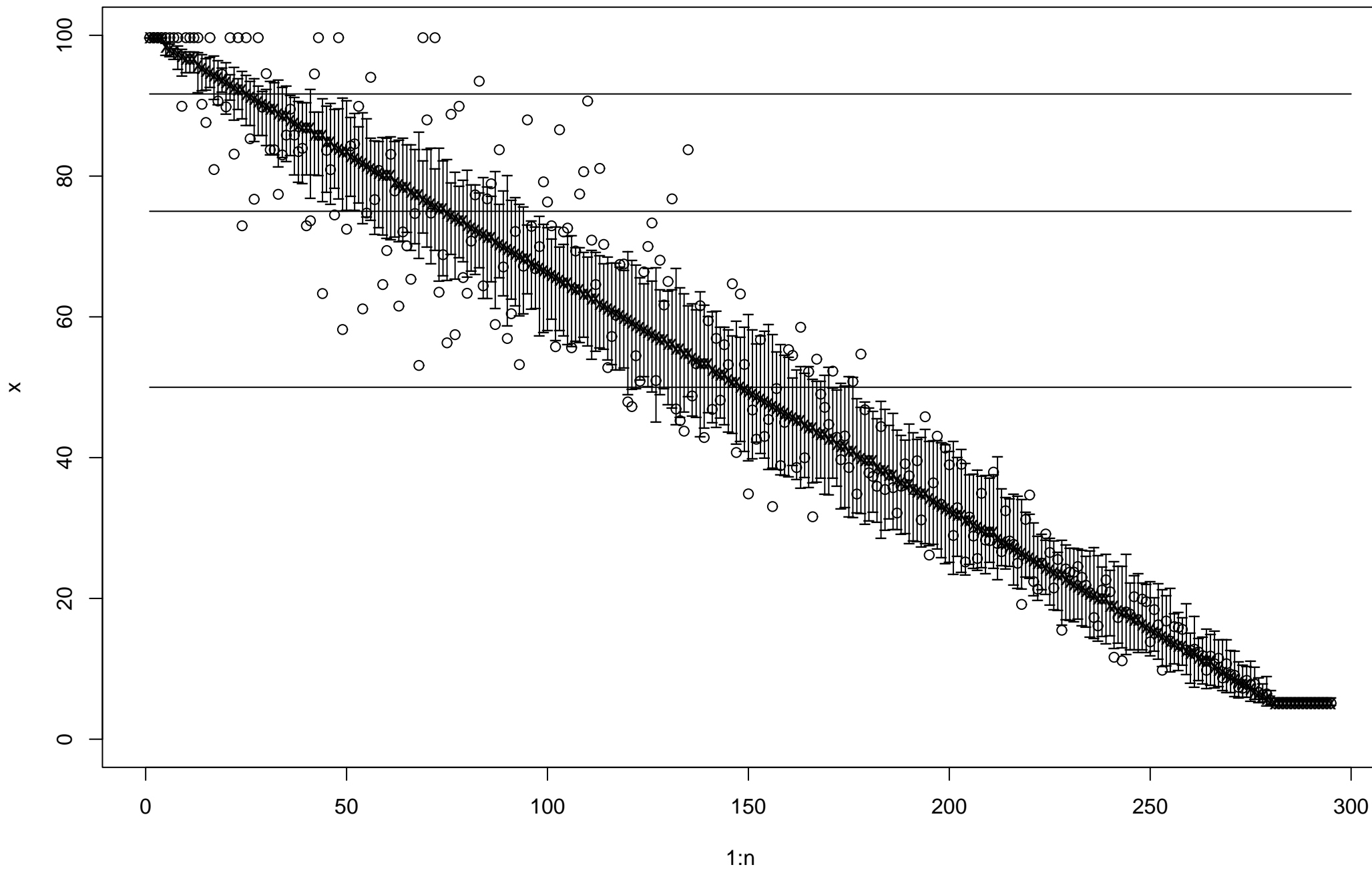












Report Confidence Intervals

Gee I was lucky!

On another day I might've won a medal!

Minimize Chance

But explicitly acknowledge as inevitable

Honourable Mention

Huge uncertainty right at medal cutoff

Right tail is thinner than median!

No real difference between fringes of adjacent categories