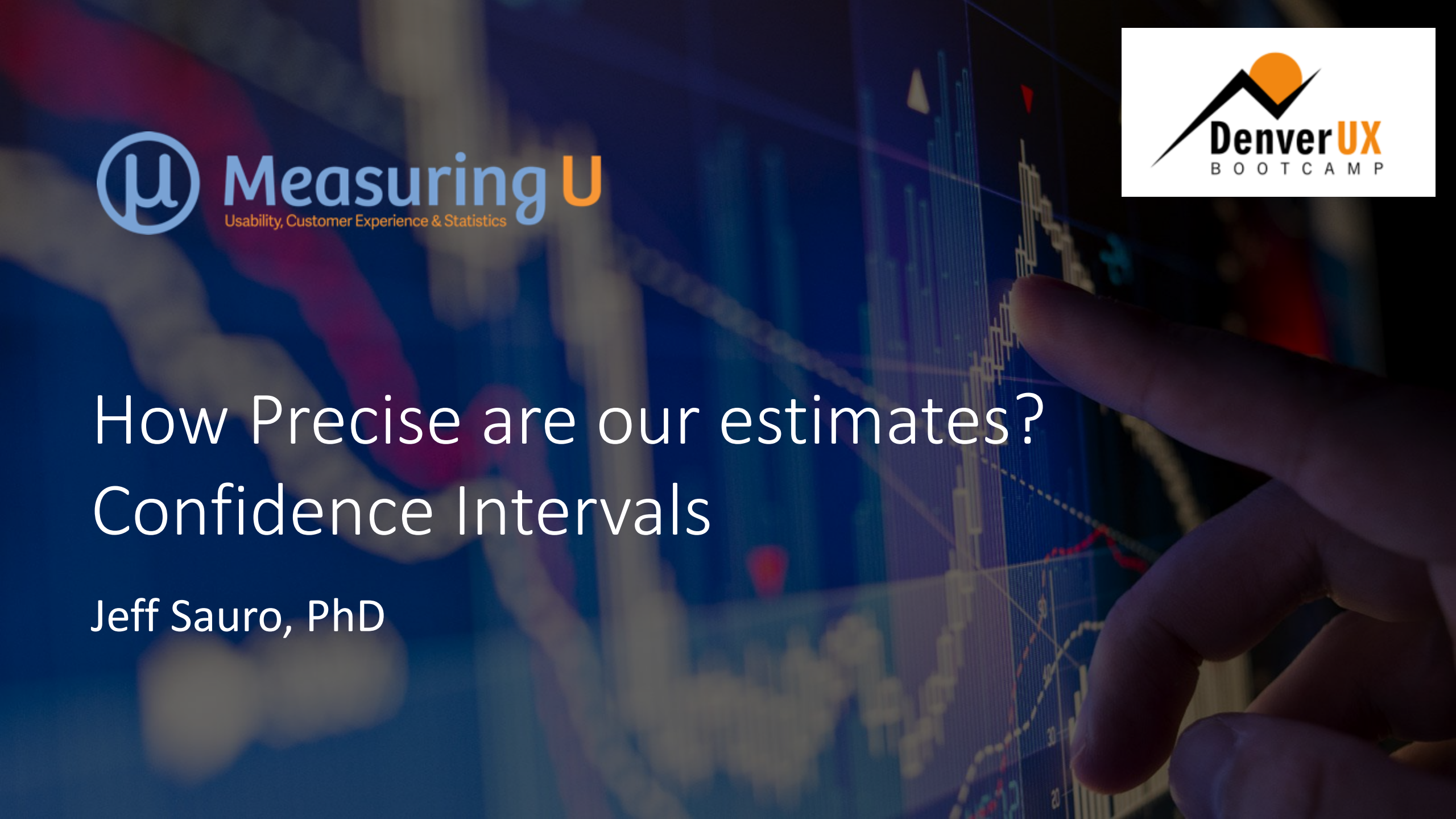
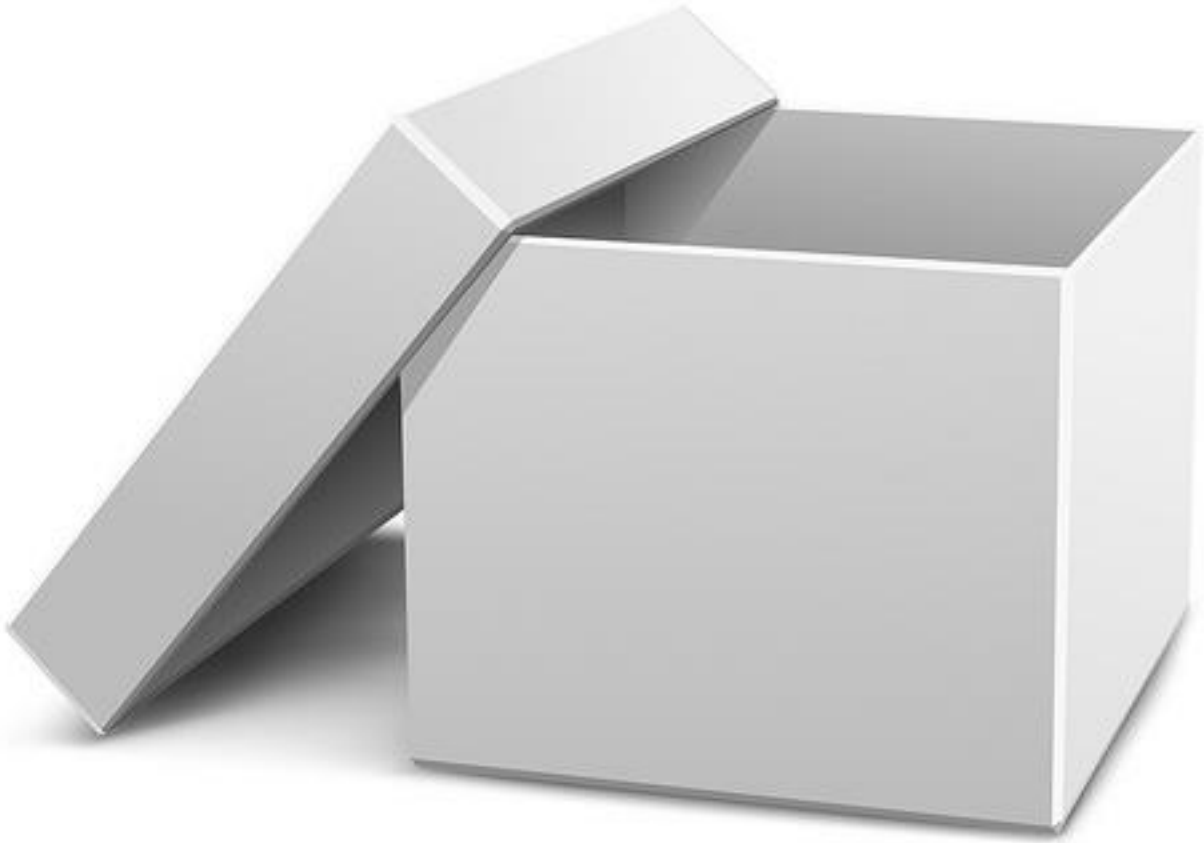
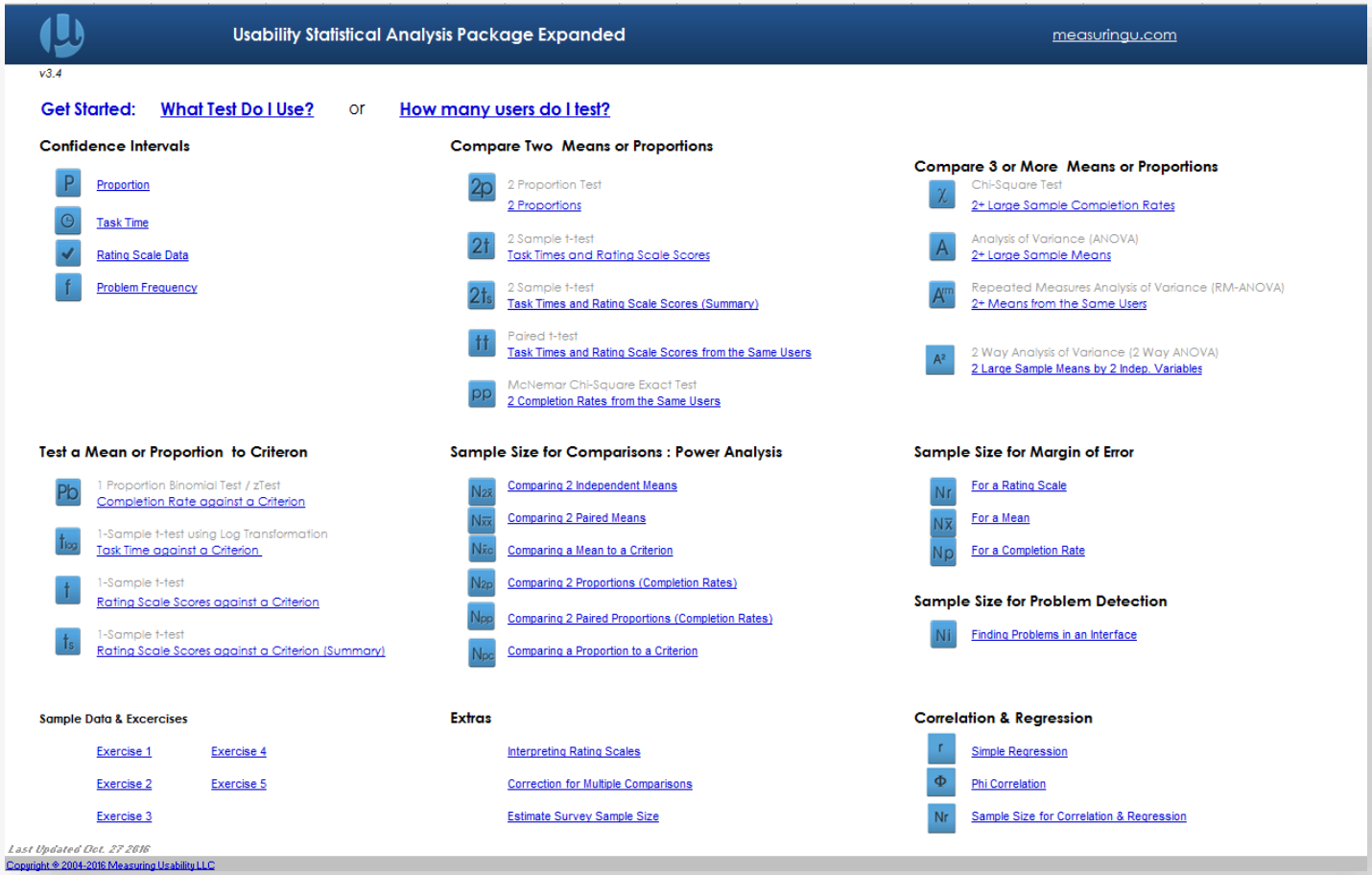


How Precise are our estimates? Confidence Intervals

Jeff Sauro, PhD



Calculator: Try it Yourself



Usability Statistical Analysis Package Expanded — measuringu.com

v3.4

Get Started: What Test Do I Use? or How many users do I test?

Confidence Intervals

- Proportion
- Task Time
- Rating Scale Data
- Problem Frequency

Compare Two Means or Proportions

- 2 Proportion Test — 2 Proportions
- 2 Sample t-test — Task Times and Rating Scale Scores
- 2 Sample t-test — Task Times and Rating Scale Scores (Summary)
- Paired t-test — Task Times and Rating Scale Scores from the Same Users
- McNemar Chi-Square Exact Test — 2 Completion Rates from the Same Users

Compare 3 or More Means or Proportions

- Chi-Square Test — 2+ Large Sample Completion Rates
- Analysis of Variance (ANOVA) — 2+ Large Sample Means
- Repeated Measures Analysis of Variance (RM-ANOVA) — 2+ Means from the Same Users
- 2 Way Analysis of Variance (2 Way ANOVA) — 2 Large Sample Means by 2 Indep. Variables

Test a Mean or Proportion to Criteron

- 1 Proportion Binomial Test / zTest — Completion Rate against a Criterion
- 1-Sample t-test using Log Transformation — Task Time against a Criterion
- 1-Sample t-test — Rating Scale Scores against a Criterion
- 1-Sample t-test — Rating Scale Scores against a Criterion (Summary)

Sample Size for Comparisons : Power Analysis

- Comparing 2 Independent Means
- Comparing 2 Paired Means
- Comparing a Mean to a Criterion
- Comparing 2 Proportions (Completion Rates)
- Comparing 2 Paired Proportions (Completion Rates)
- Comparing a Proportion to a Criterion

Sample Size for Margin of Error

- For a Rating Scale
- For a Mean
- For a Completion Rate

Sample Size for Problem Detection

- Finding Problems in an Interface

Sample Data & Excercises

Exercise 1, Exercise 2, Exercise 3, Exercise 4, Exercise 5

Extras

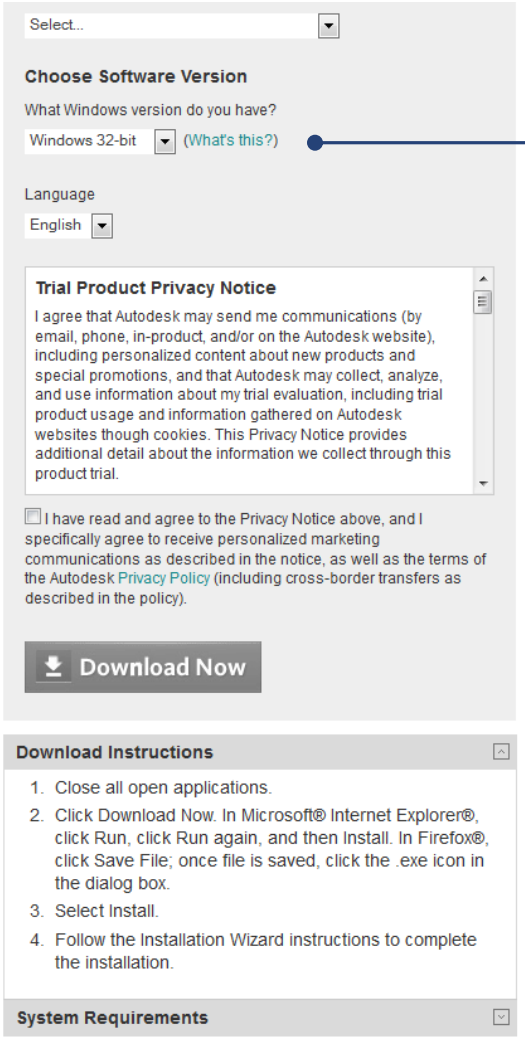
- Interpreting Rating Scales
- Correction for Multiple Comparisons
- Estimate Survey Sample Size

Correlation & Regression

- Simple Regression
- Phi Correlation
- Sample Size for Correlation & Regression

Last Updated Oct. 27 2016

What 3 users can tell you that 3000 cannot



95% Confident

between **9% and 57% of users**

would have the problem

Would an errata sheet guide Users to avoid a mistake in documentation?

Important Installation Information

DO THIS FIRST!

Your system programs *must* be updated to avoid a system malfunction.

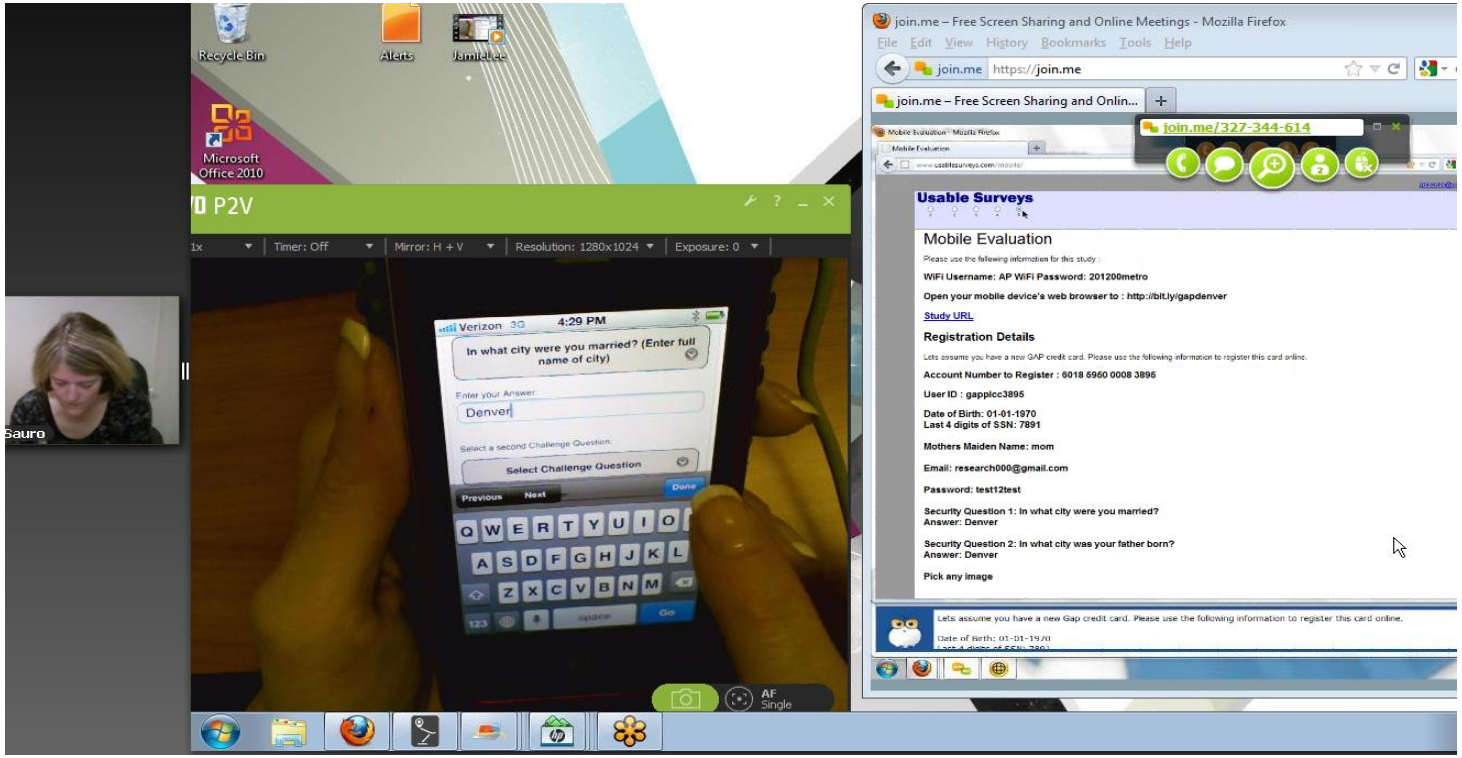
If you have an IBM* Personal System/2* Model XX, *immediately* place this sheet in your Quick Reference between pages 18 and 19, and perform the following instructions before starting step 22 on page 19.

If you have an IBM* Personal System/2* Model YY, *immediately* place this sheet in your Quick Reference between pages 10 and 11, and perform the following instructions before starting step 8 on page 10.

The Oatmeal.com

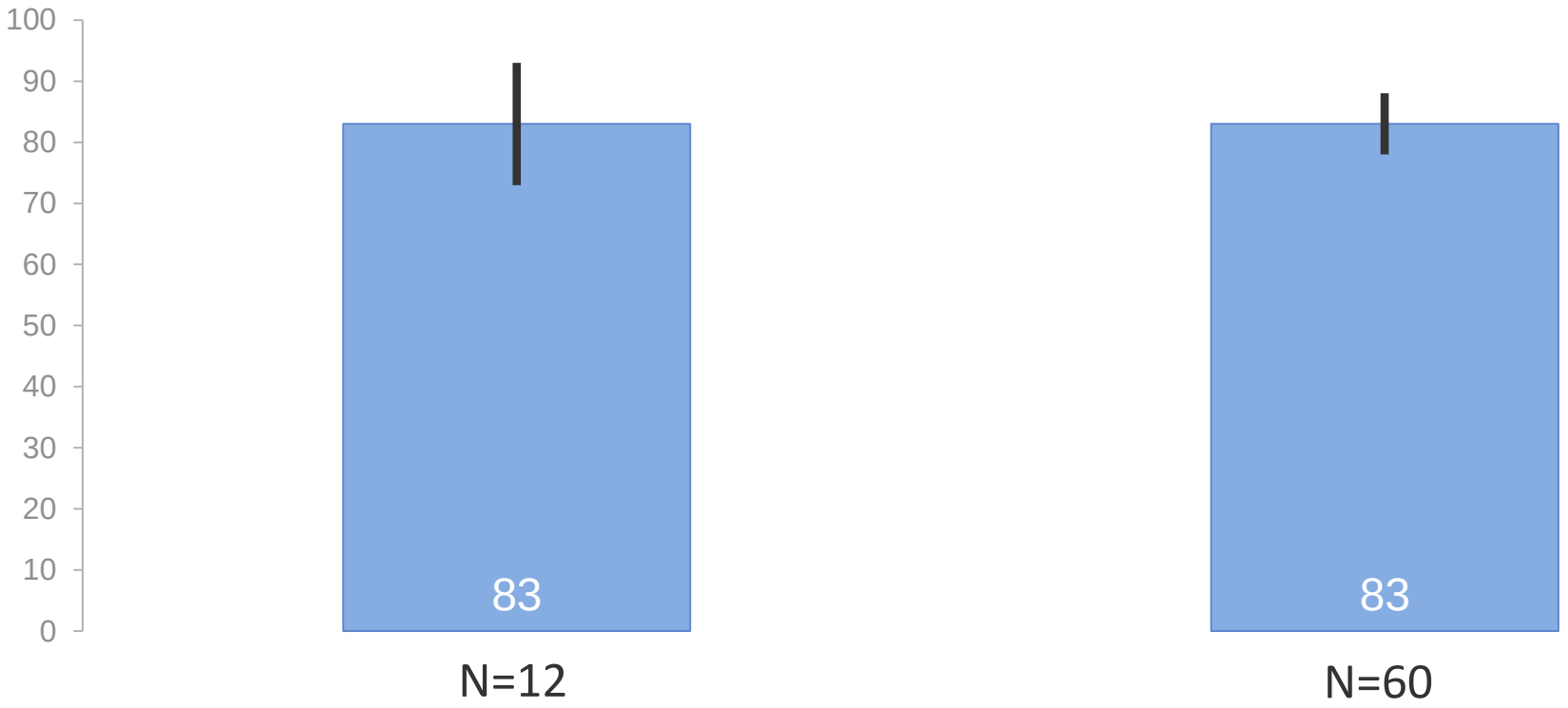
6 out of 8: 40% to 94%

Is fear of stolen data a major inhibitor from paying a bill on a mobile website?



5 out of 16: 14% to 56%

Confidence Intervals: Precision & Location

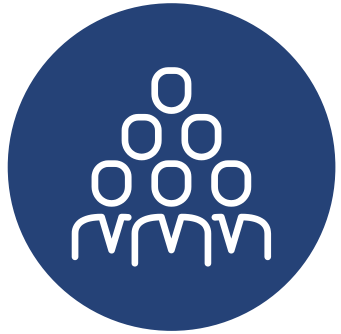


95% Confidence Intervals around SUS Scores

A confidence Interval has 3 ingredients

Confidence Level

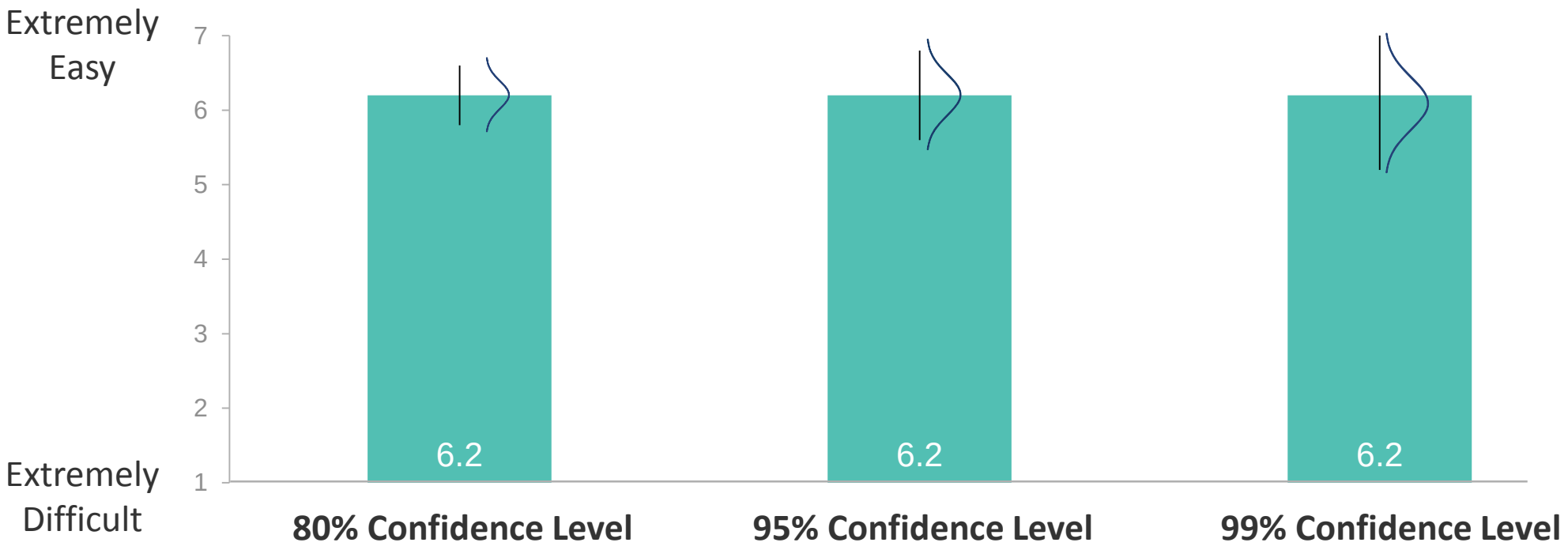
Variability



Sample Size

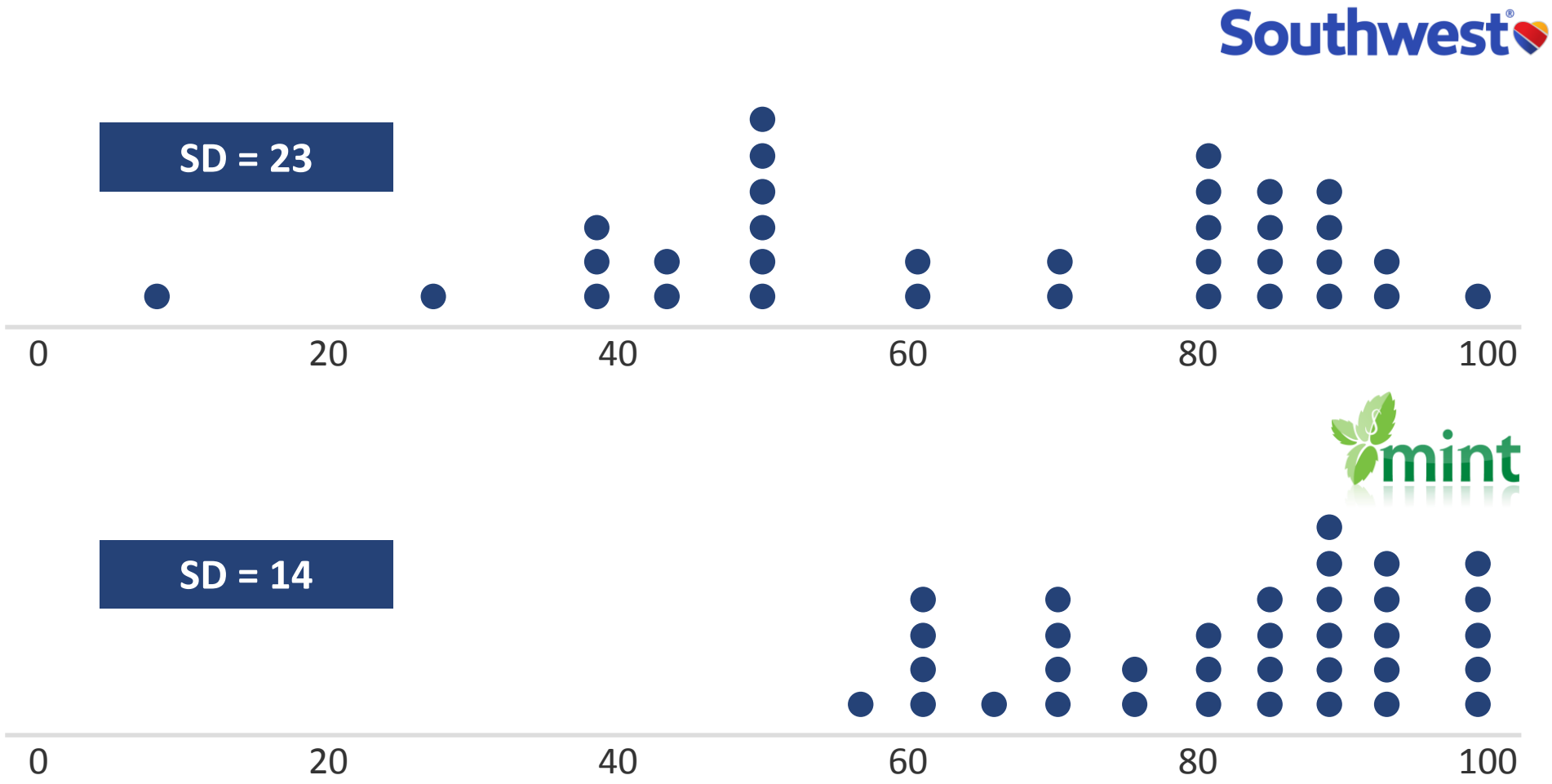
The most likely range of the unknown population mean

Different confidence levels for Task Difficulty

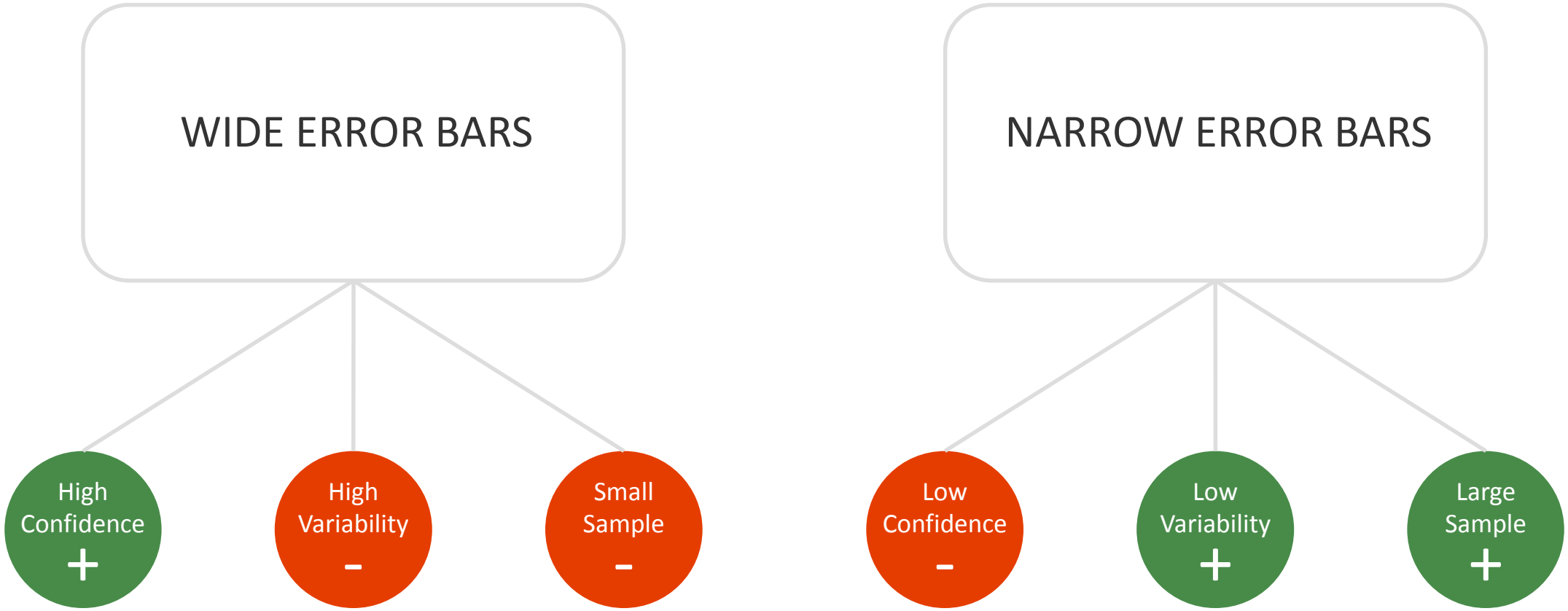


Intervals get wider with higher confidence

Variability

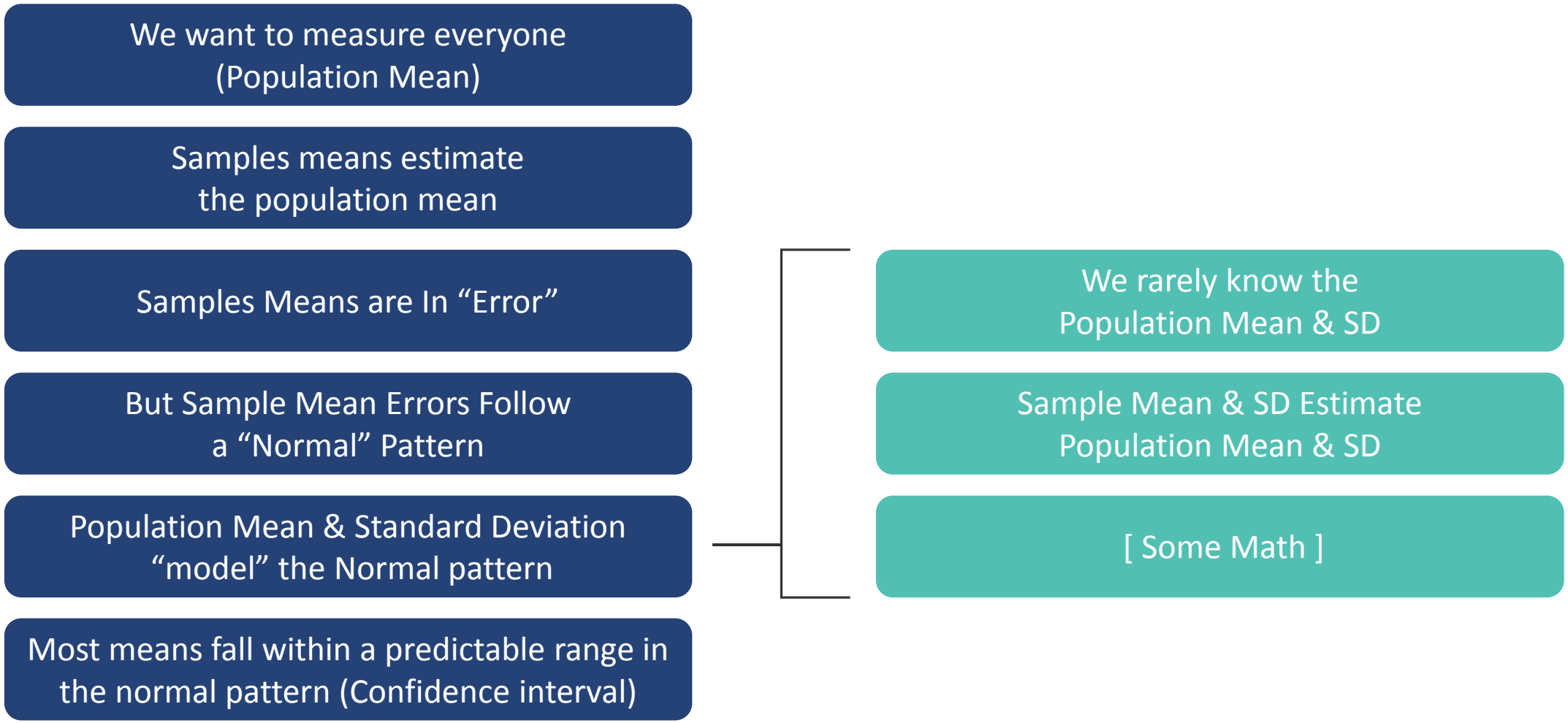


Confidence Interval Width



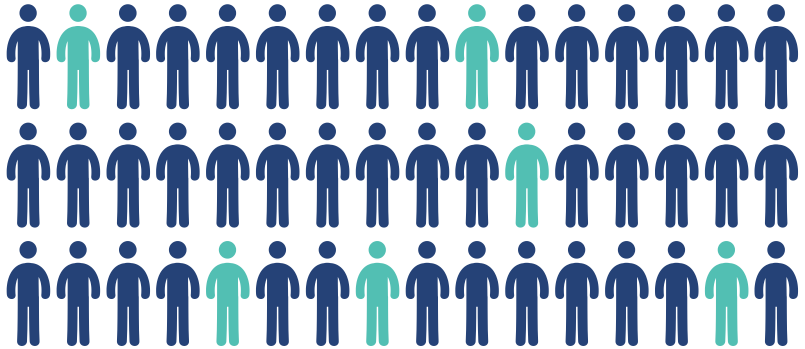
The Method & Math Behind the Confidence Interval

Confidence Interval Learning Map



A population is defined by its mean

μ σ



[PARAMETERS]

SURVEY SAMPLE

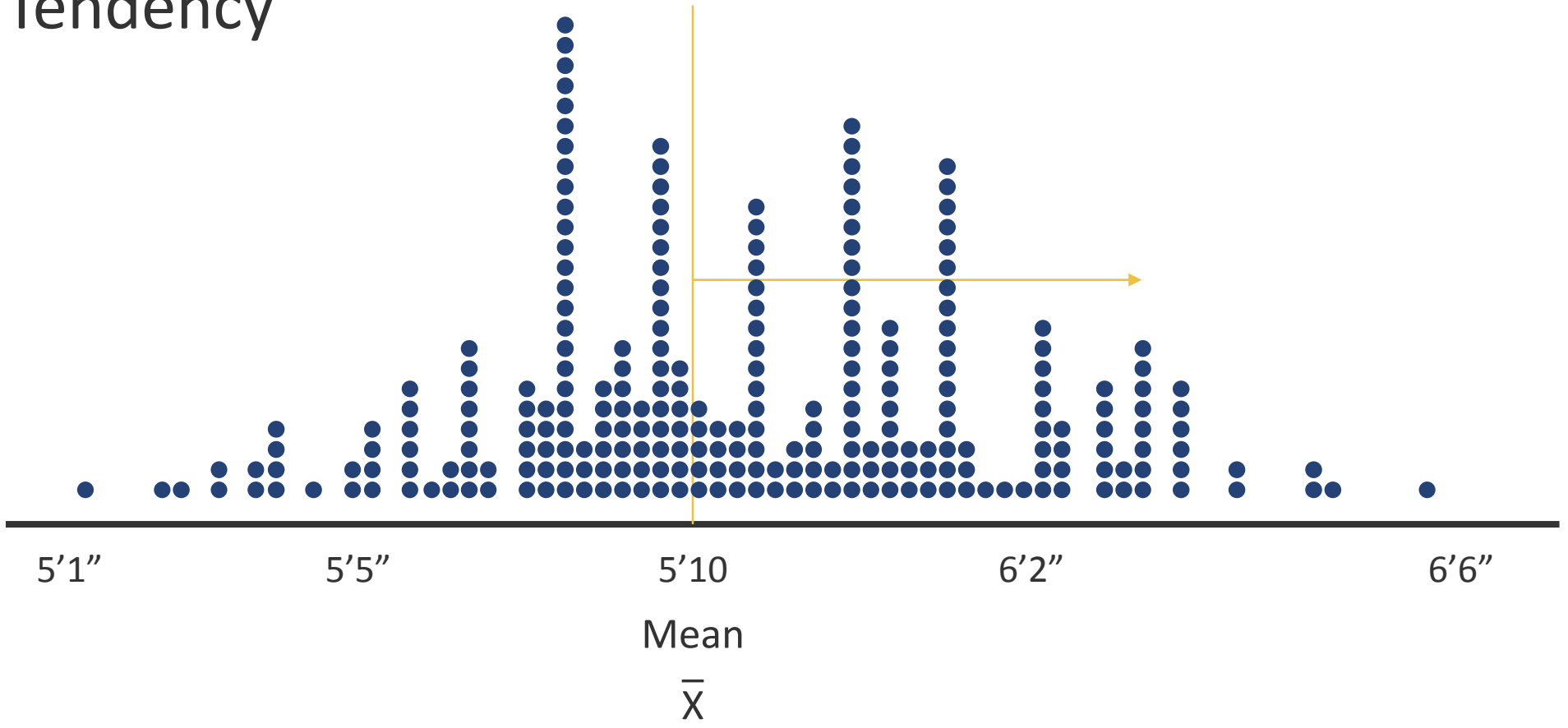
$\overline{x}$ s

[STATISTICS]

Mean & Standard Deviation

Mean: Typical Value or Midpoint

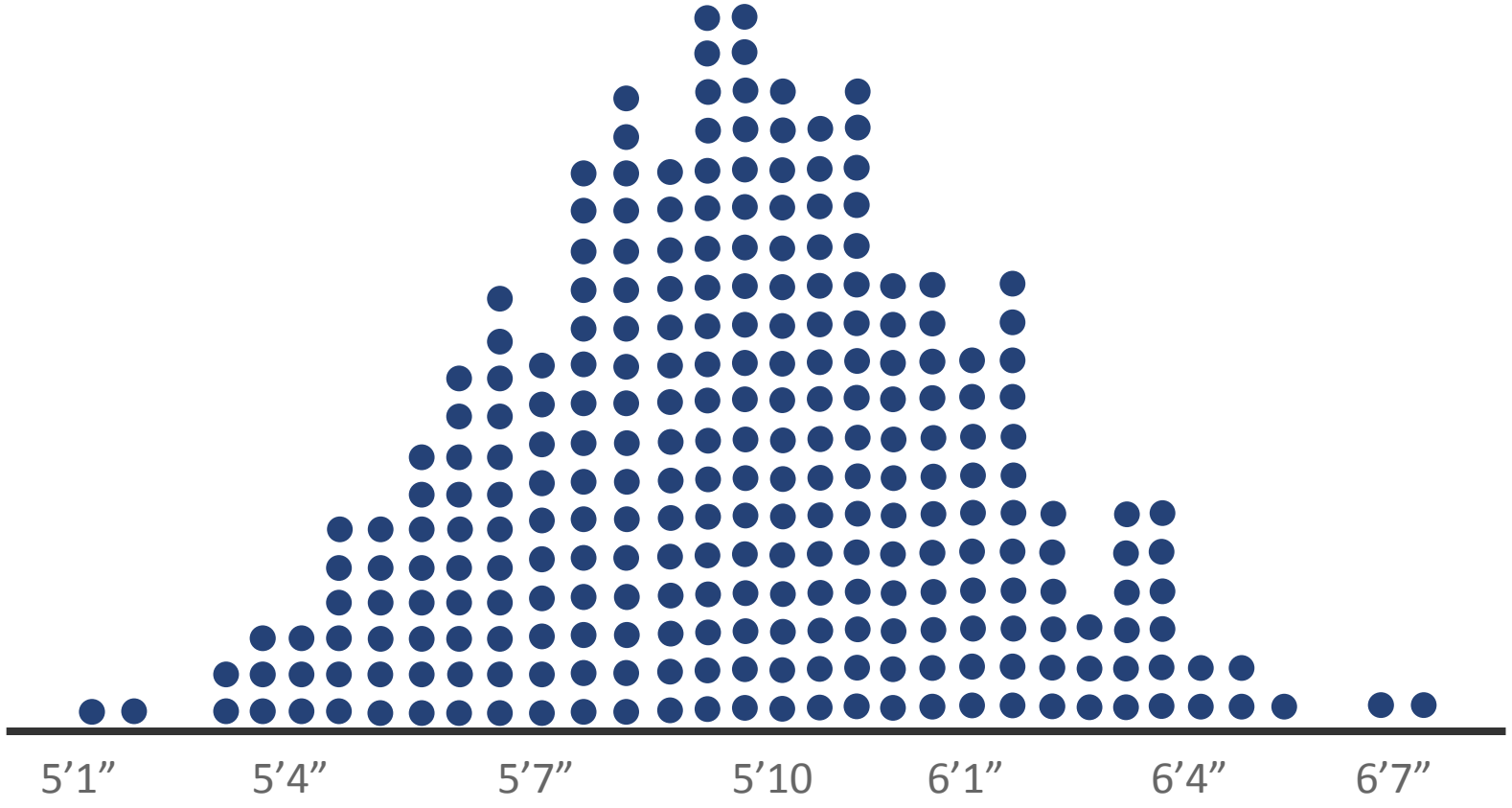
Central Tendency



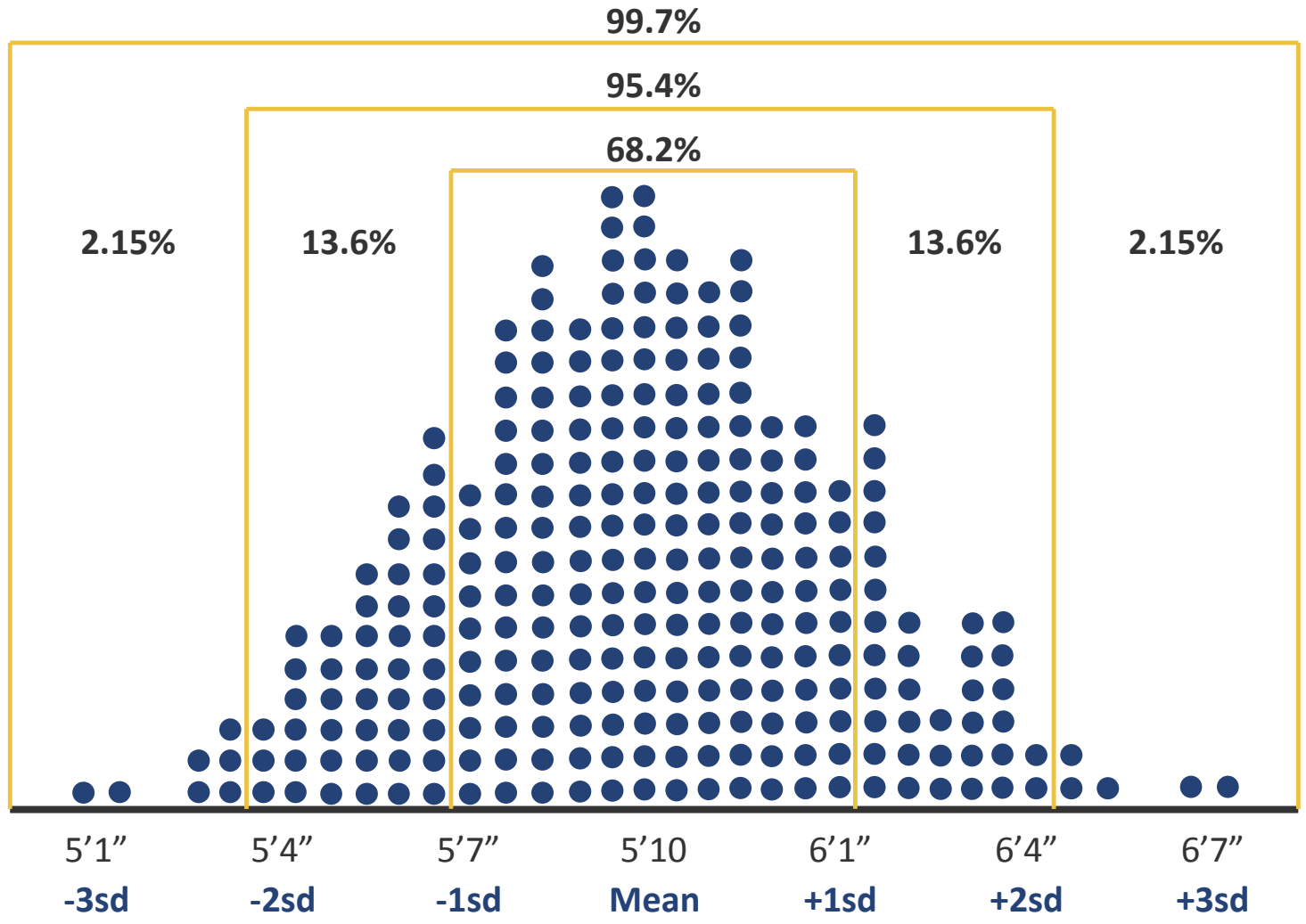
Sample of 85 Men’s heights

Standard Deviation

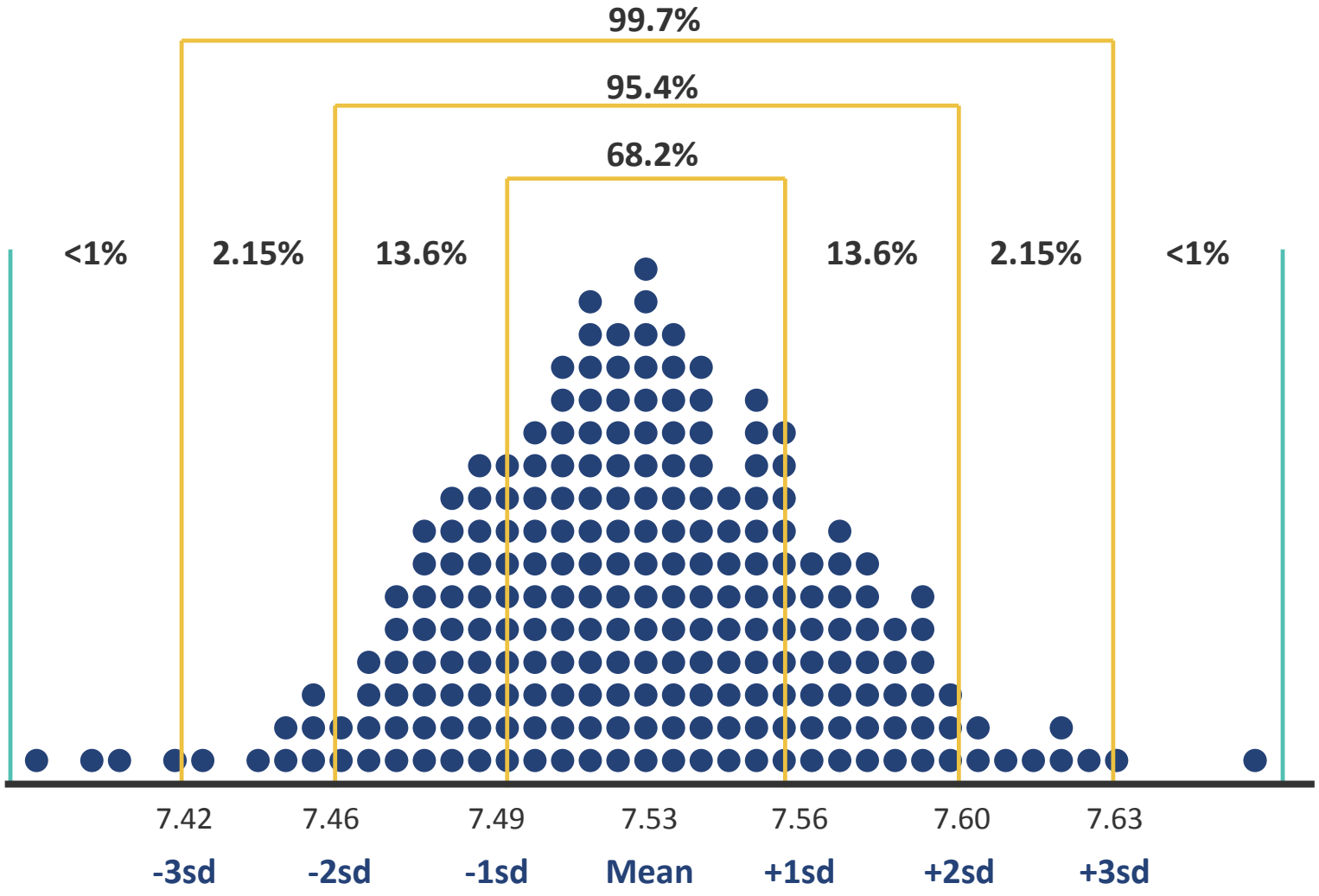
500 Heights of North American Men



500 Heights of North American Men

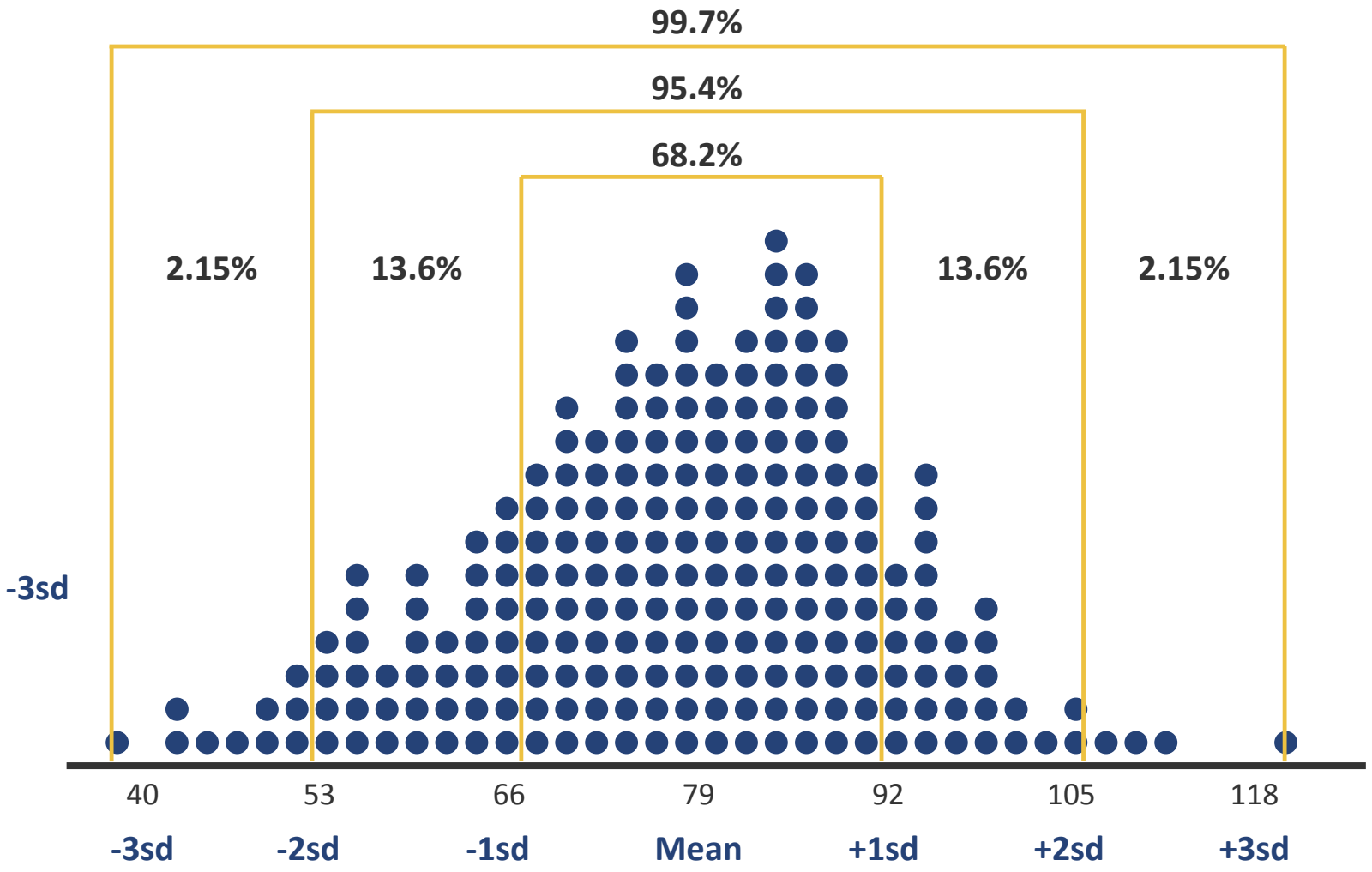


Weight of 2000 Euro Coins

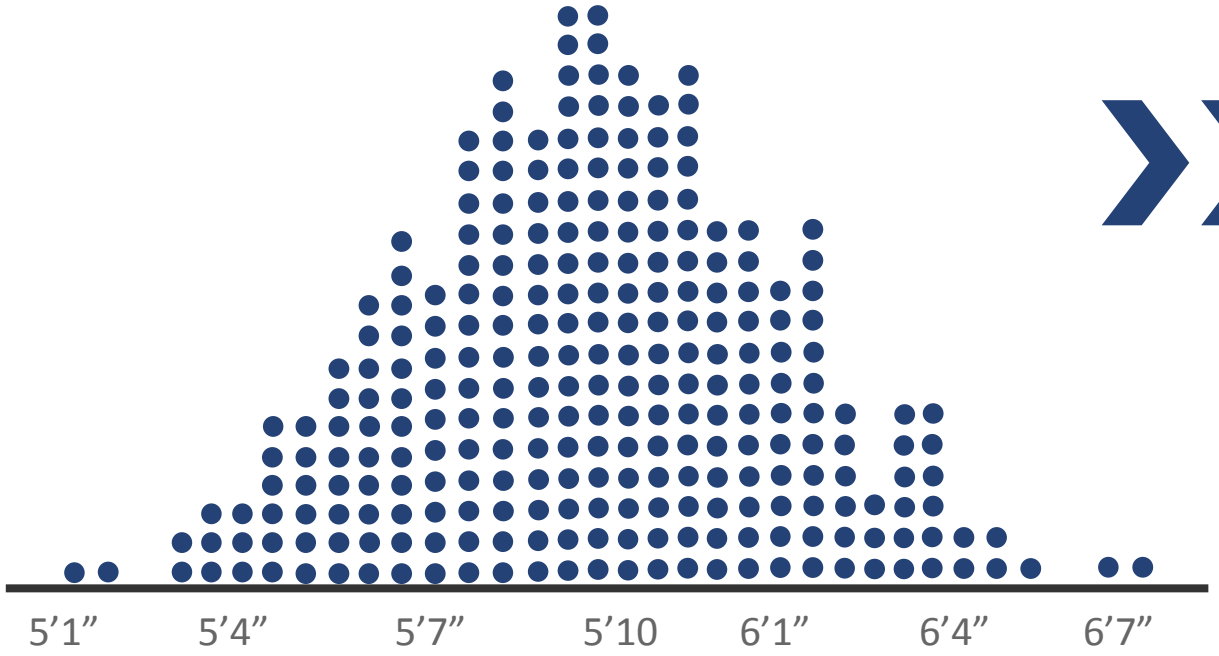


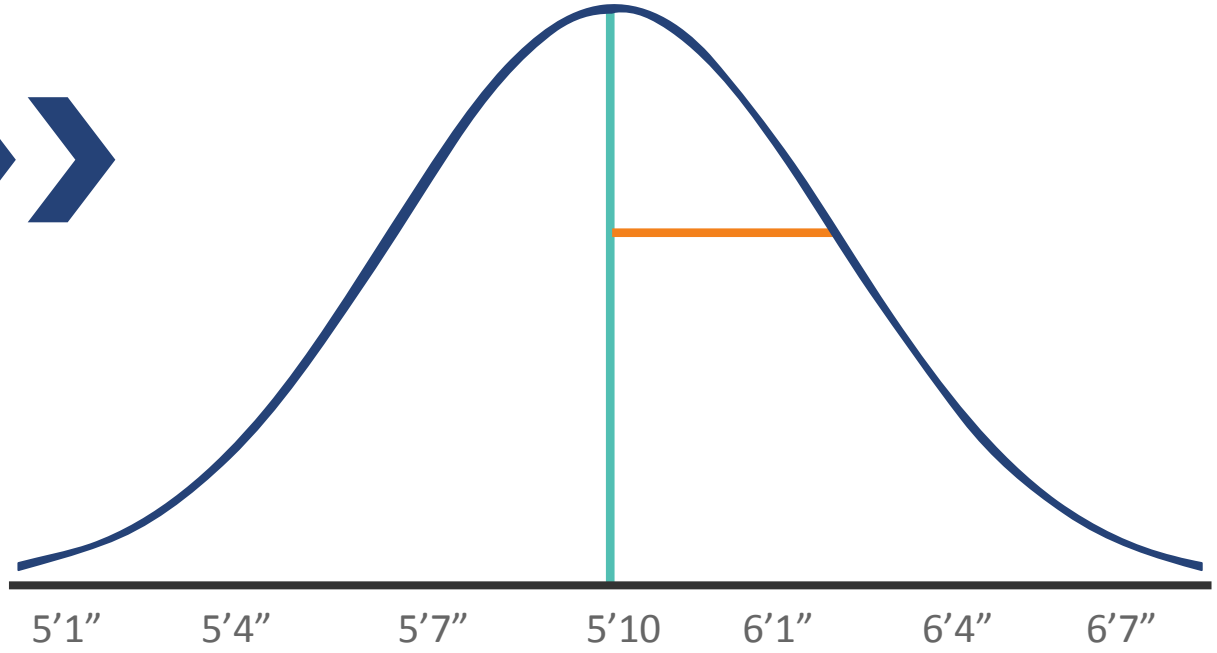

TACO
BELL®

MLB Games Won by Team for 31 seasons

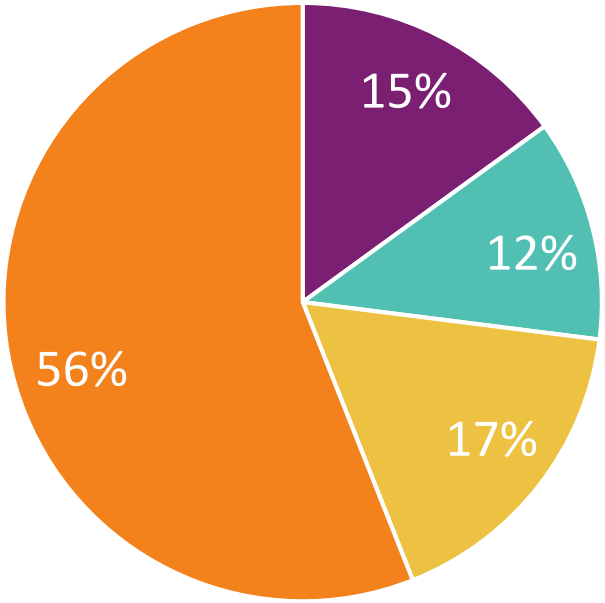


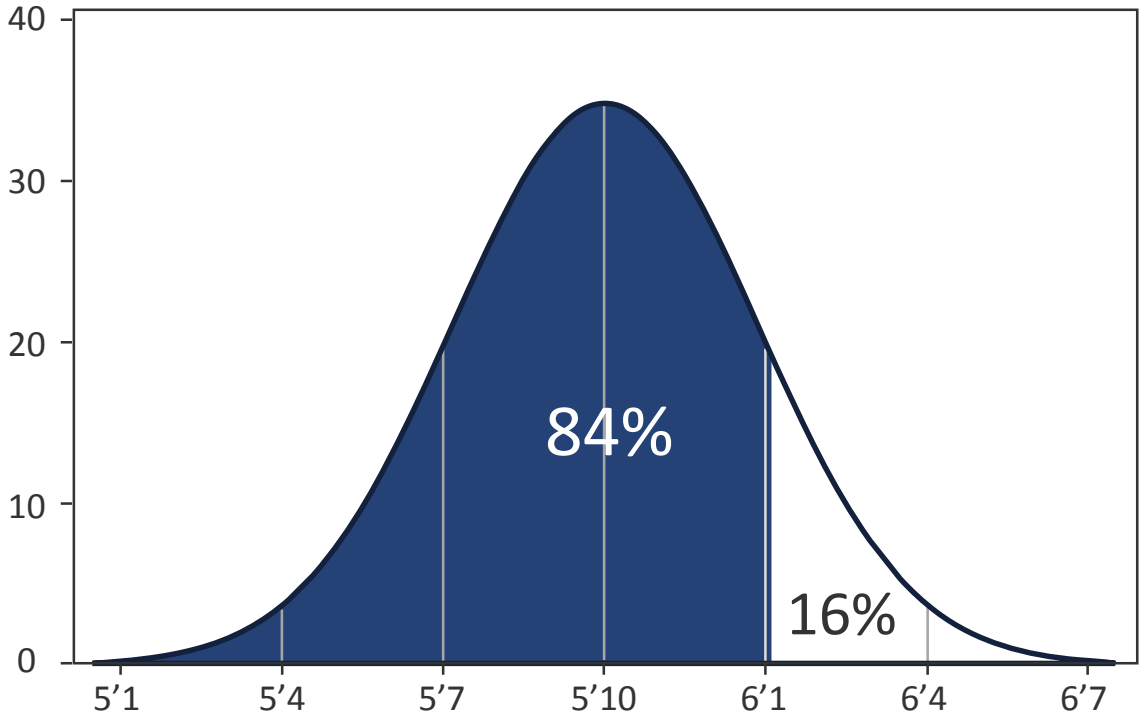
500 Heights of North American Men



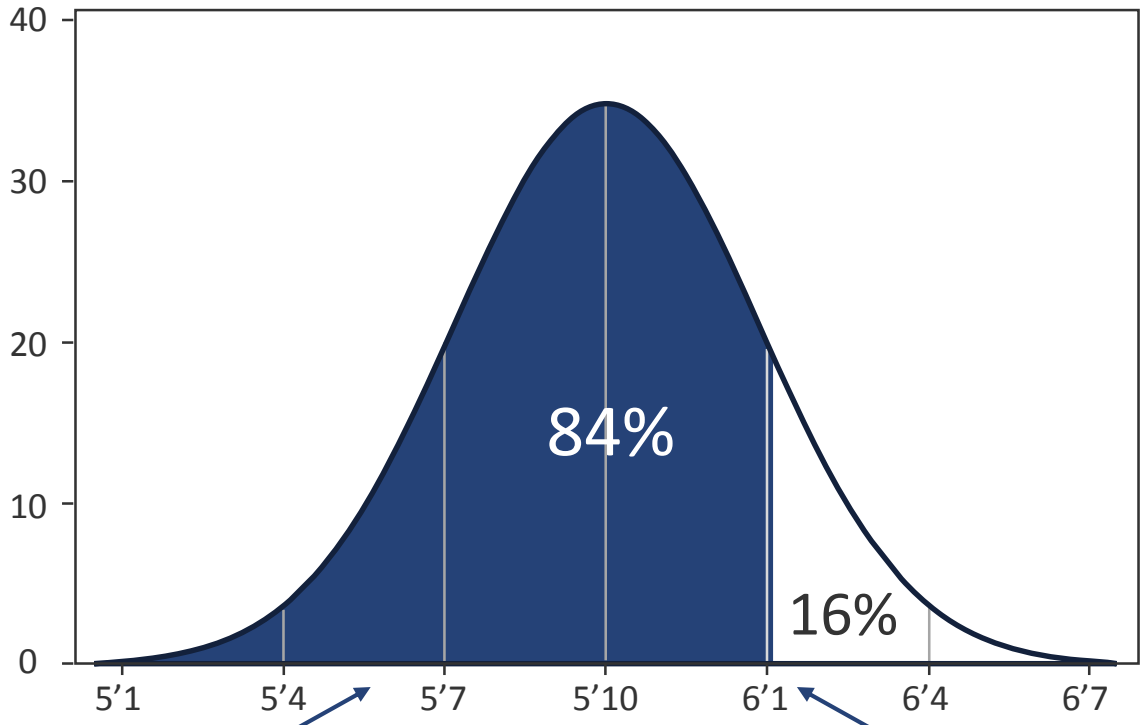


Total Area adds up to 100%





Normal (z) Scores



$$z = \frac{5'5 - 5'10}{3} = \frac{-5}{3} = -1.67$$

$$z = \frac{6'1 - 5'10}{3} = \frac{3}{3} = 1$$

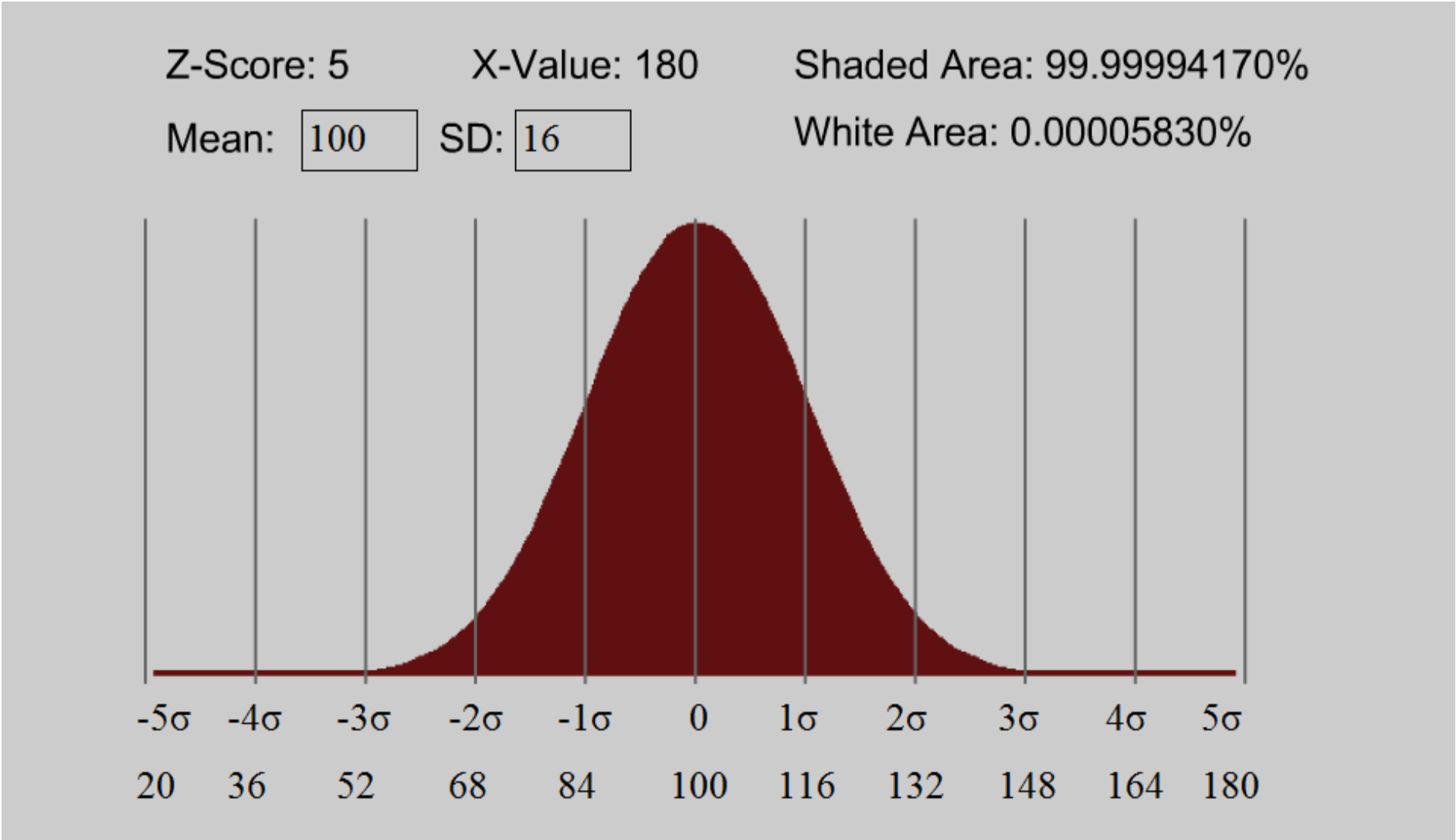
Z-Score is the number of standard deviations above or below the mean

Z-Score Concept

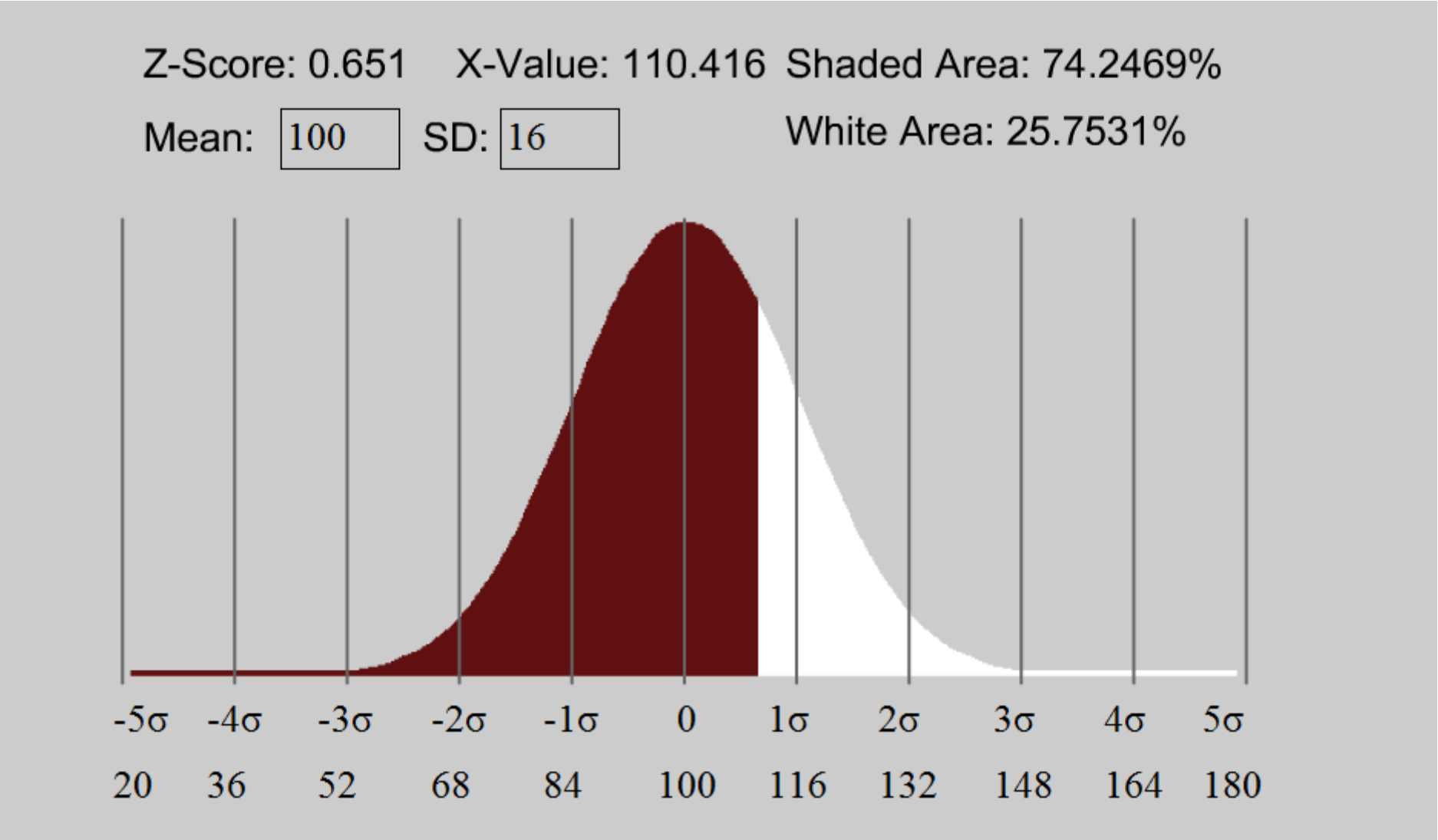
- Z-score is the # of Standard Deviations a value is from the mean

- If you know the z-score, use a table of values to find area:
  - Find how extreme a value is (1 sided area) called percentile rank
  - Find the % of values that fall within that many SD's of the mean (2 sided area)

Normal Distribution (2-Tailed)



Normal Distribution (1-Tailed)



Quick Exercise 1

Find the z-scores if the **industry average** rating of ease is a 5.1 and the standard deviation is a .7 on a 7 point rating scale.

$$z = \frac{(Observed\ Avg - Industry\ Avg)}{Standard\ Deviation}$$

1. An observed average rating of 6
2. An observed average rating of 4.1

$$z = \frac{6 - 5.1}{.7} = \frac{.9}{.7} = 1.29$$

$$z = \frac{4.1 - 5.1}{.7} = \frac{-1}{.7} = -1.42$$

Quick Exercise 2

Use the table to answer the following questions about ease scores and the industry average.

1. What is the percentile rank of a task with an avg. rating of 6 (z-score of 1.29)?
2. What is the percentile rank of a task with an avg. rating of 4.1 (z-score of -1.42)?
3. What percent of values would we expect fall within 1.6 SD of the industry average score of 5.1?

Z-Score	1 Tailed %	2 Tailed %
-2.0	2%	5%
-1.7	4%	9%
-1.4	8%	16%
-1.1	14%	27%
-0.8	21%	42%
-0.5	31%	62%
-0.2	42%	84%
0.1	54%	8%
0.4	66%	31%
0.7	76%	52%
1	84%	68%
1.3	90%	81%
1.6	95%	89%
1.9	97%	94%

Quick Quiz

True or False

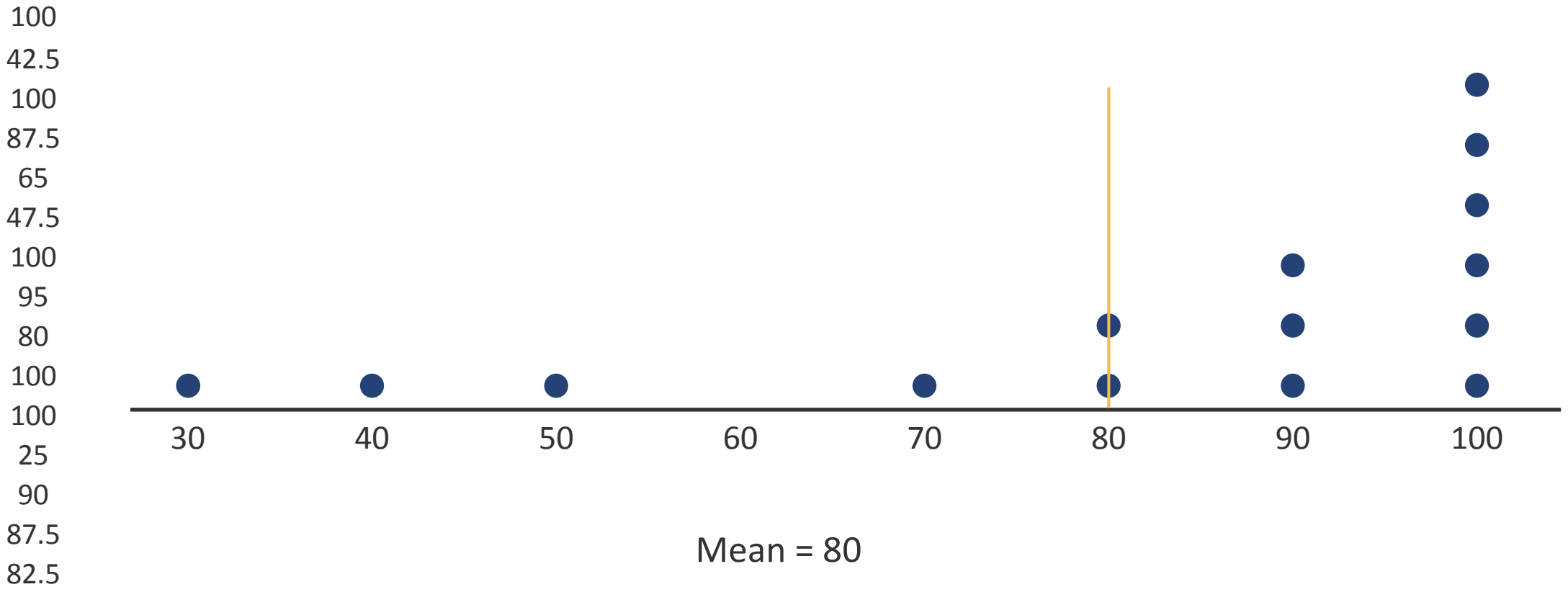
95% of values fall within 2 standard deviations of the mean.

A z-score is measured in standard deviations

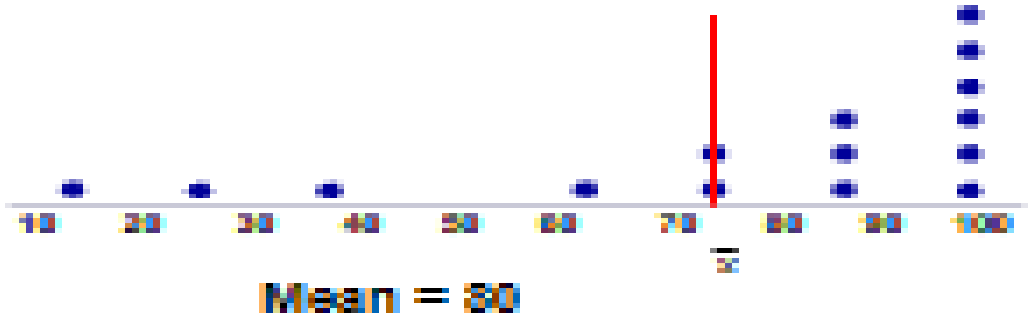
3 Standard deviations above and below the mean account for around 99% of values.

Confidence Interval for Rating Scale Data

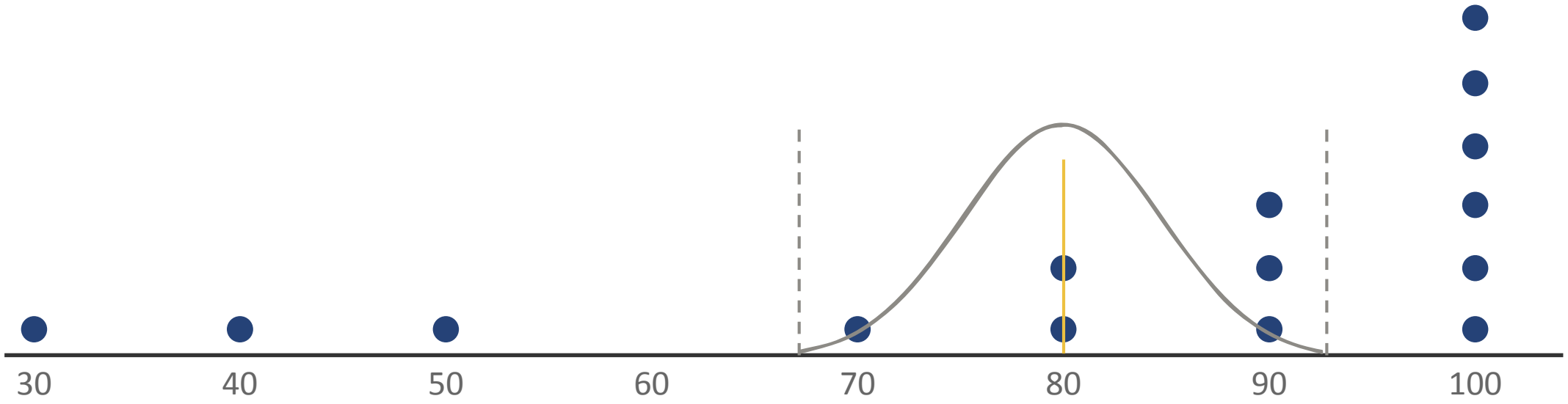
Budget.com 15 SUS Scores



How Much Will the Sample Mean Vary?

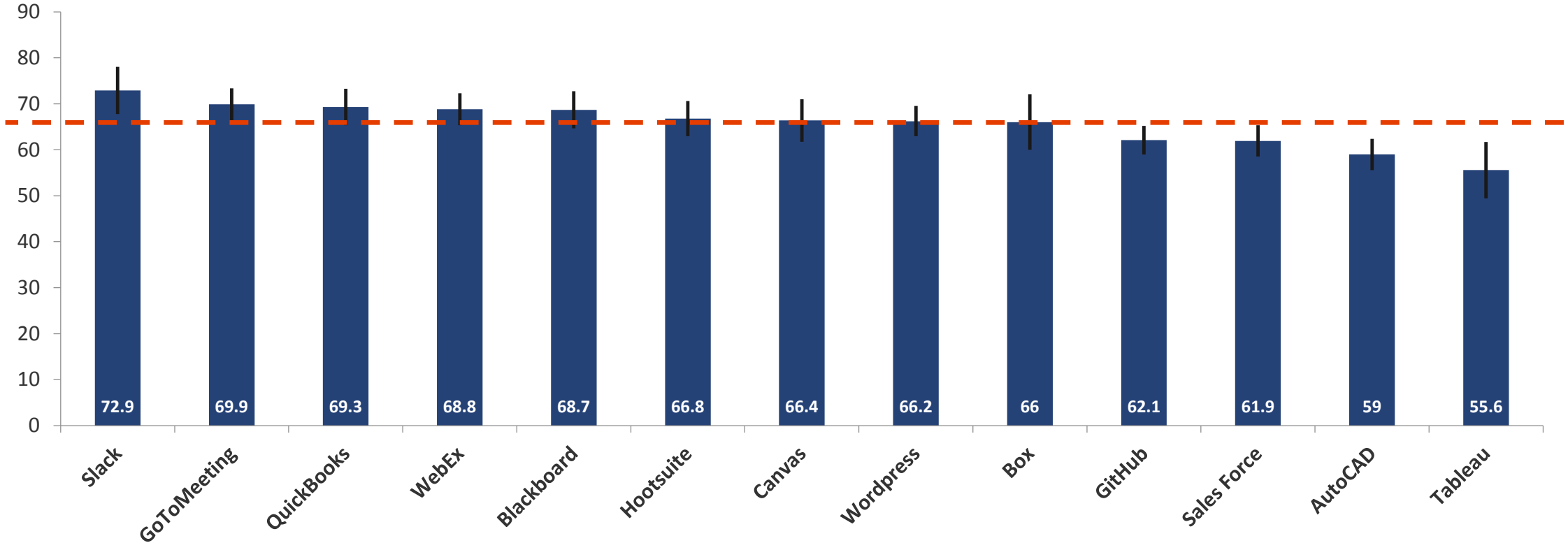


Visualize the confidence interval

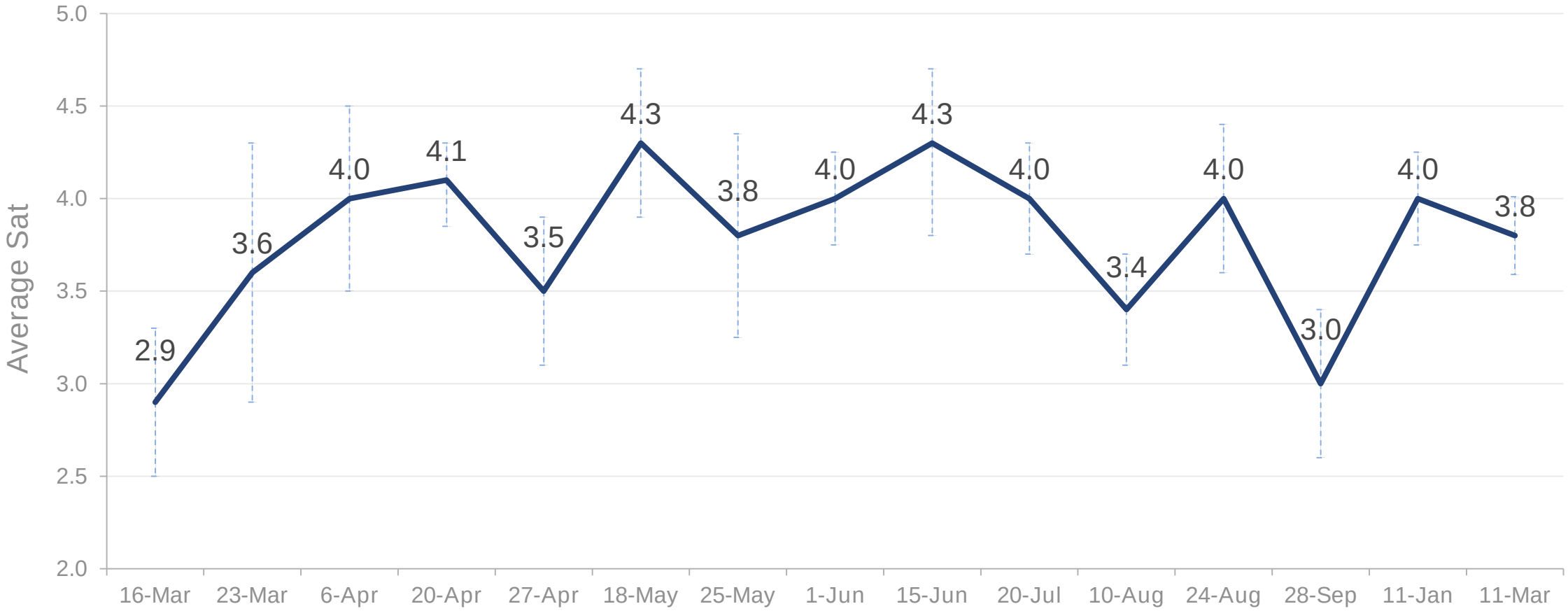


95% CI (67 to 94)

SUS Benchmarks by Product



CIS For Management



QUICK QUIZ

Which has the wider confidence Intervals?

A

Sample Size (n) = 50

95% Confidence Levels

80% Confidence Levels

Standard Deviation = 12

Standard Deviation = 12

B

Sample Size (n) = 10

90% Confidence Levels

99% Confidence Levels

Standard Deviation = 82

Variance = 144

Confidence Interval for Completion Rates

Task Completion Rate CI

Wald Confidence Interval

What is taught in school

Z score (α changes for confidence level)

$$\hat{p} \pm z_{(1-\frac{\alpha}{2})} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Proportion (#instances/# total responses)

Terrible for Small Samples or when np < 15

Adjusted – WALD [8/14 users = .57]

$$.57 \pm 1.96 \sqrt{\frac{.57\ (1-.57)}{14}}$$

$$\frac{\frac{Z^2}{2}}{Z^2} \sim \begin{matrix} +\ 2 \text{ to Num.} \\ +\ 4 \text{ to Denom.} \end{matrix}$$

$$.56 \pm 1.96 \sqrt{\frac{.56\ (1-.56)}{18}}$$

.56 Adjust the proportion by adding 2 to the numerator and 4 to the denominator (10/18)

18 The adjusted denominator (total number of users)

We are 95% confident that the best estimate of the entire customer population is between 33% and 79%

Confidence Interval for Task Time Data

Task Time Data is *Skewed*

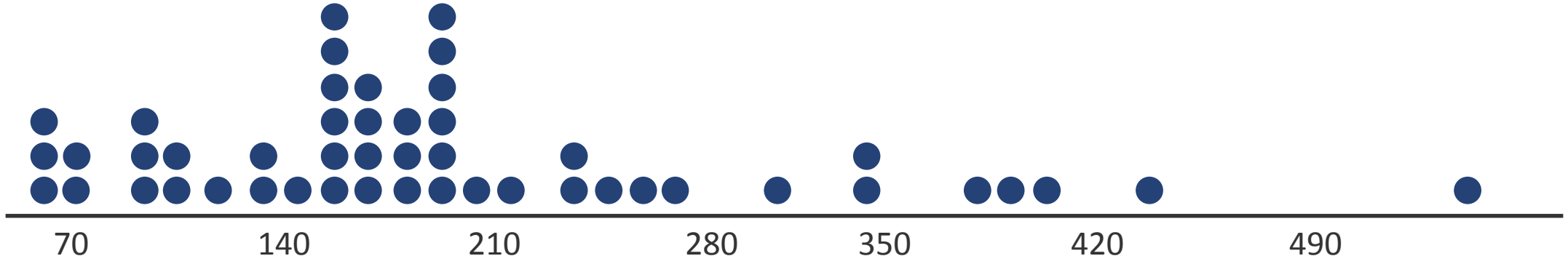

MOONLIGHTING
SEASON FIVE
The Final Season

Logarithm is just an Exponent

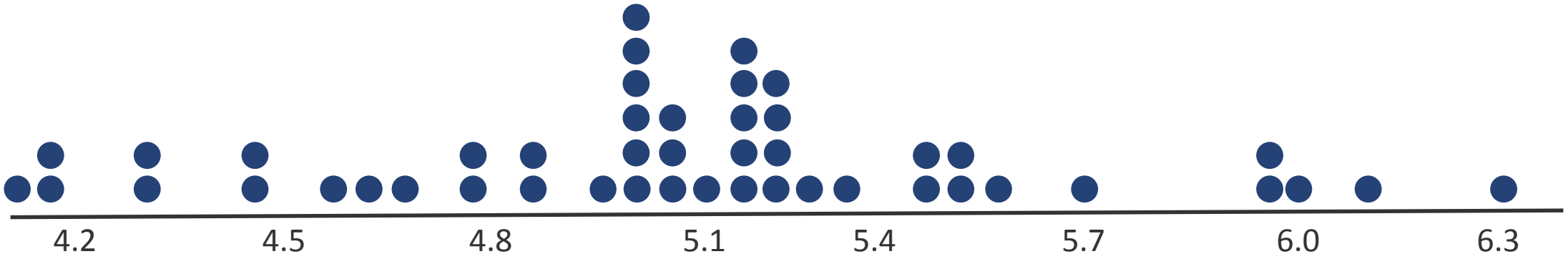
Exponent	Logarithm
$10^1 = 10$	Log(10) =1
$10^2 = 100$	Log(100) = 2
$10^3 = 1000$	Log(1000) = 3

What is Log 10,000?

50 Task Times Transformed

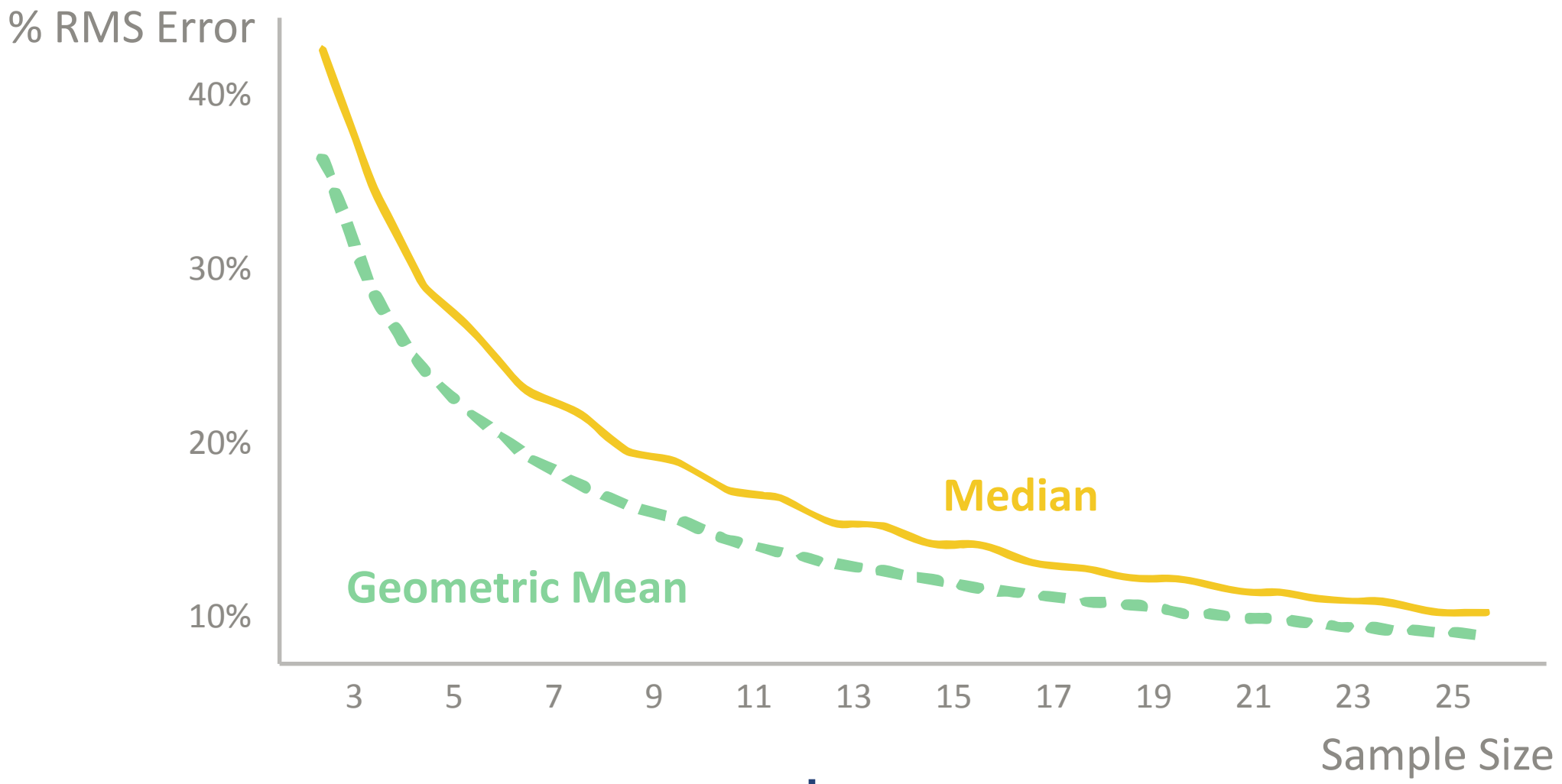


Raw Times



Transformed Times (logs)

Error


% RMS Error
40%
30%
20%
10%
Median
Geometric Mean
3
5
7
9
11
13
15
17
19
21
23
25
Sample Size


ESTD 1759
GUINNESS
GUINNESS

Confidence Interval General Form

Continuous

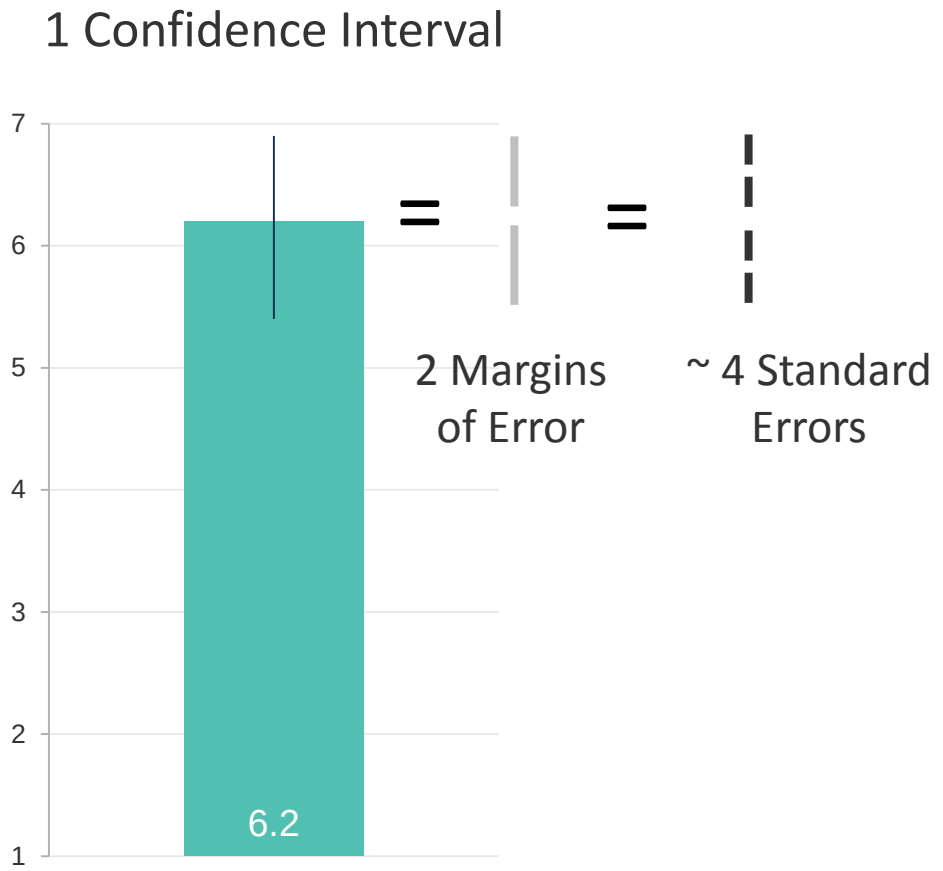
$$\bar{x} \pm t_{(1-\frac{\alpha}{2})} \frac{s}{\sqrt{n}}$$

Binary

$$\hat{p} \pm z_{(1-\frac{\alpha}{2})} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Avg. **Z/t** **Standard Error**

95% Confidence Interval

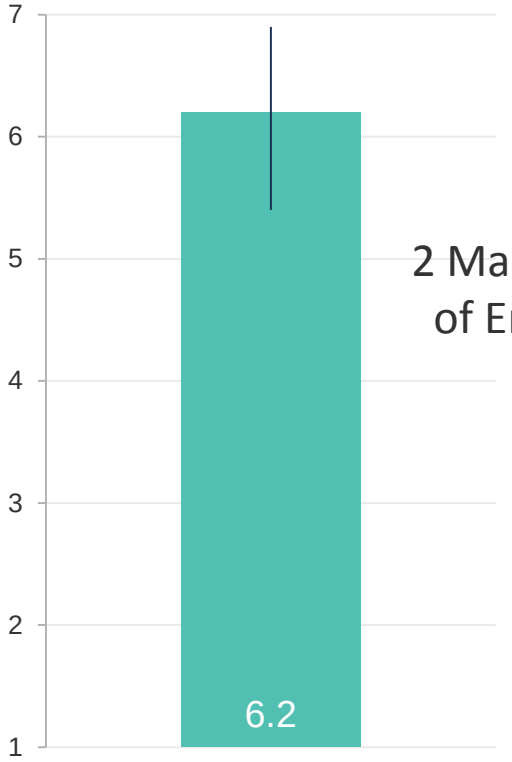


$$\frac{s}{\sqrt{n}}$$
$$\frac{s}{\sqrt{n}}$$
$$\bar{x} = \bar{x} \pm 2 * \frac{s}{\sqrt{n}}$$
$$\frac{s}{\sqrt{n}}$$
$$\frac{s}{\sqrt{n}}$$

Margin of Error (d)

95% Confidence Interval Exercise

1 Confidence Interval **Mean: 6.2** **SD: 1.18** **N:50**



$$= \quad \bar{x} \pm 2\left[\frac{s}{\sqrt{n}}\right] \quad = \quad \overline{\boldsymbol{x}} \pm 2 * \frac{\mathrm{s}}{\sqrt{\mathrm{n}}}$$

Margin of Error (d)

$$= \quad 6.2 \pm 2\left[\frac{1.18}{\sqrt{50}}\right]$$

$$= \quad 6.2 \pm .33$$

= 95% Confident Mean is between 5.8 & 6.5

10 Things to Know about Confidence Intervals

1. Most likely range of the unknown population average or percentage
2. Location and precision of a measure
3. Three things impact the width of a confidence interval
    1. Confidence Level
    2. Variability
    3. Sample Size
4. The confidence interval is equal to two margins or errors
5. Use the t-confidence interval for continuous data
6. For task times, log transform then use t
7. Use the adjusted Wald for binary data
8. Compute on sample sizes as small as 2
9. We are 95% confident in the method, not in any given interval
10. Overlap in intervals are a quick way to check for statistical significance

Confidence Interval Exercises

About MeasuringU

MeasuringU is a quantitative research firm based in Denver, Colorado focusing on quantifying the user experience.

MUIQ

Remote UX Testing Platform (Desktop & Mobile)

UX Research

SUPR-Q

Measurement & Statistical Analysis

Eye Tracking & Lab Based Testing