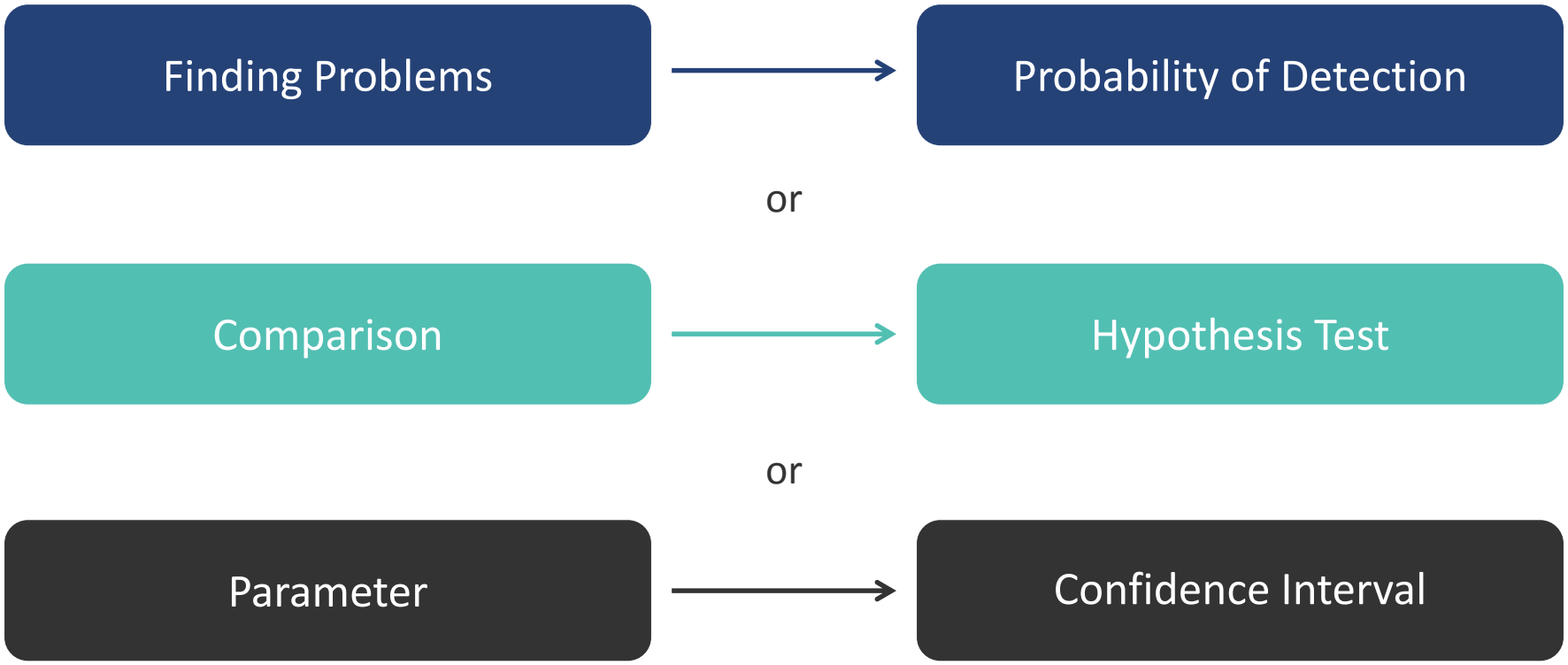
MeasuringU
Usability, Customer Experience & Statistics

Denver UX
BOOTCAMP

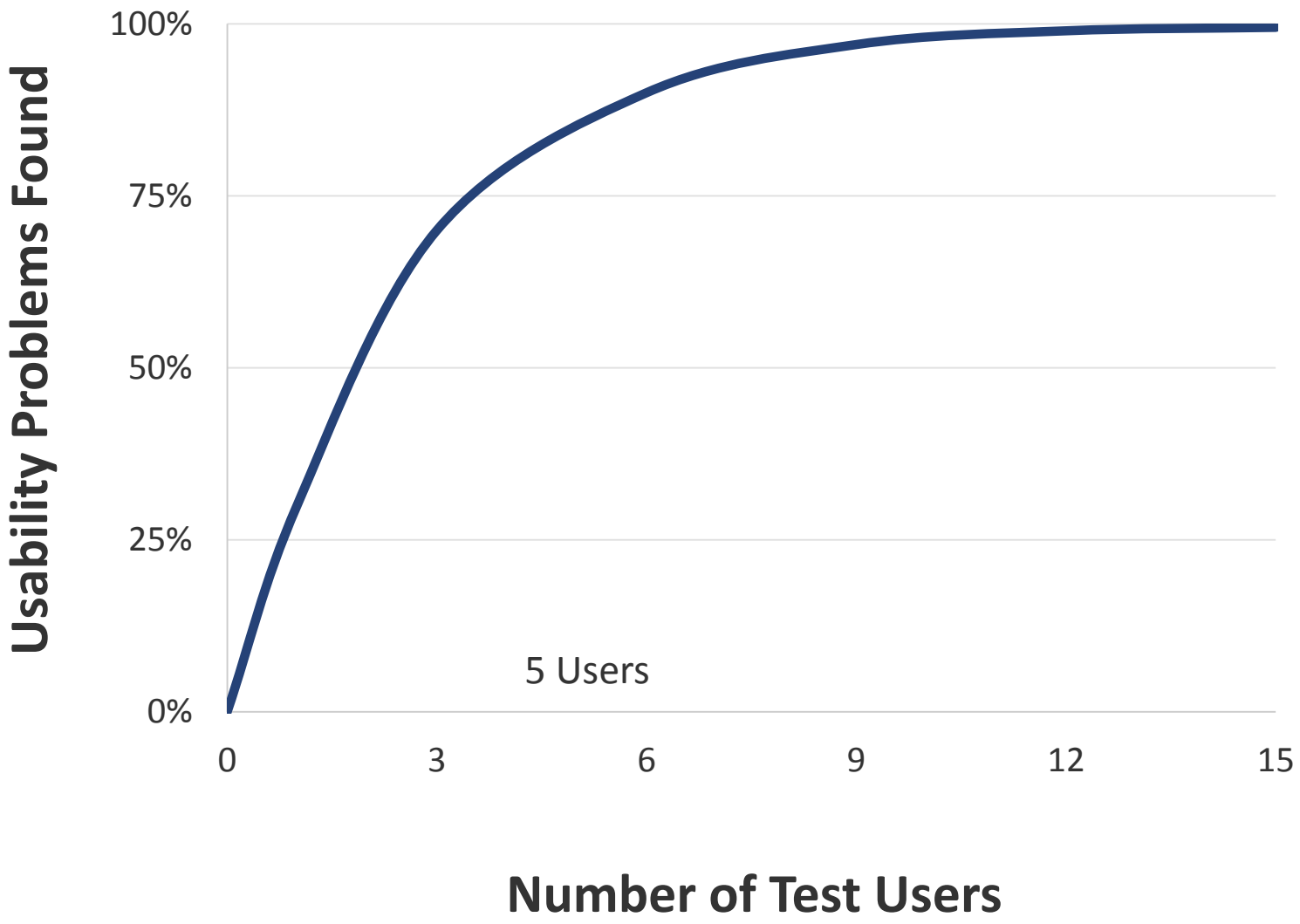
Sample Size & Power

Jeff Sauro, PhD

What sample size Do I need?



Problem Detection Sample sizes


Usability Problems Found
100%
75%
50%
25%
0%
5 Users
0
3
6
9
12
15
Number of Test Users

85% of the Time, 3 Tosses will show tails
50% Probability of tails

Toss

Sample Size 1

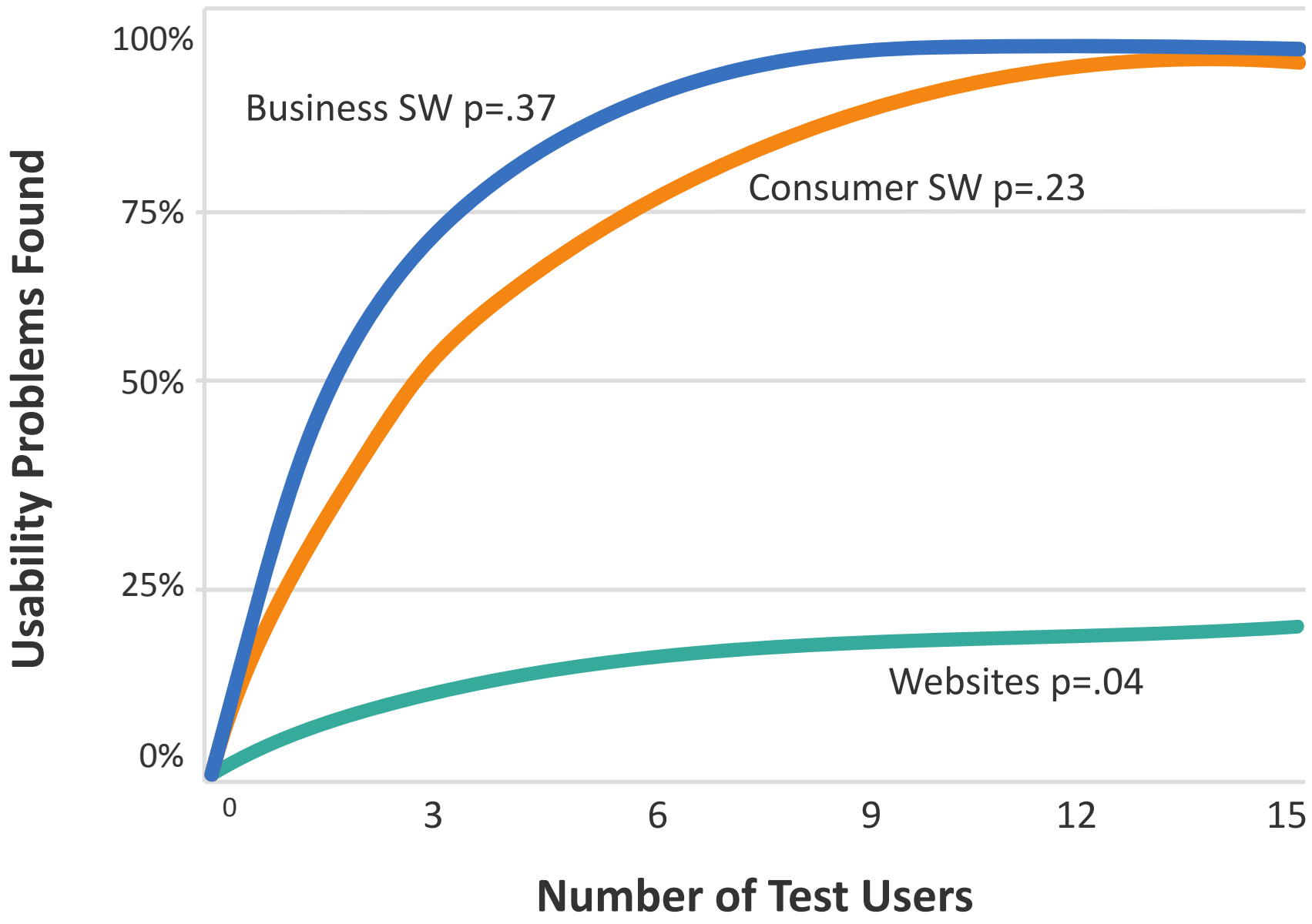
1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 4, 1

% Samples <= 3 85% Trials 12

85% of the Time, 10 Tosses will show a 1
16.7% Probability of a 1

85% of the Time, 5 Users will encounter a UI Problem
31% probability of users encountering a UI problem

Sample Size	% Samples <= 5	Test	Usability Tests
4	100%		1


Usability Problems Found
100%
75%
50%
25%
0%
Business SW p=.37
Consumer SW p=.23
Websites p=.04
0
3
6
9
12
15
Number of Test Users

Derivation of the binomial probability Formula

$$\frac{\text{Log (1- Chance of Detecting)}}{\text{Log (1- Probability of Occurring)}}$$

$$\frac{\text{Log (1-.85)}}{\text{Log (1-.31)}} = 5.11$$

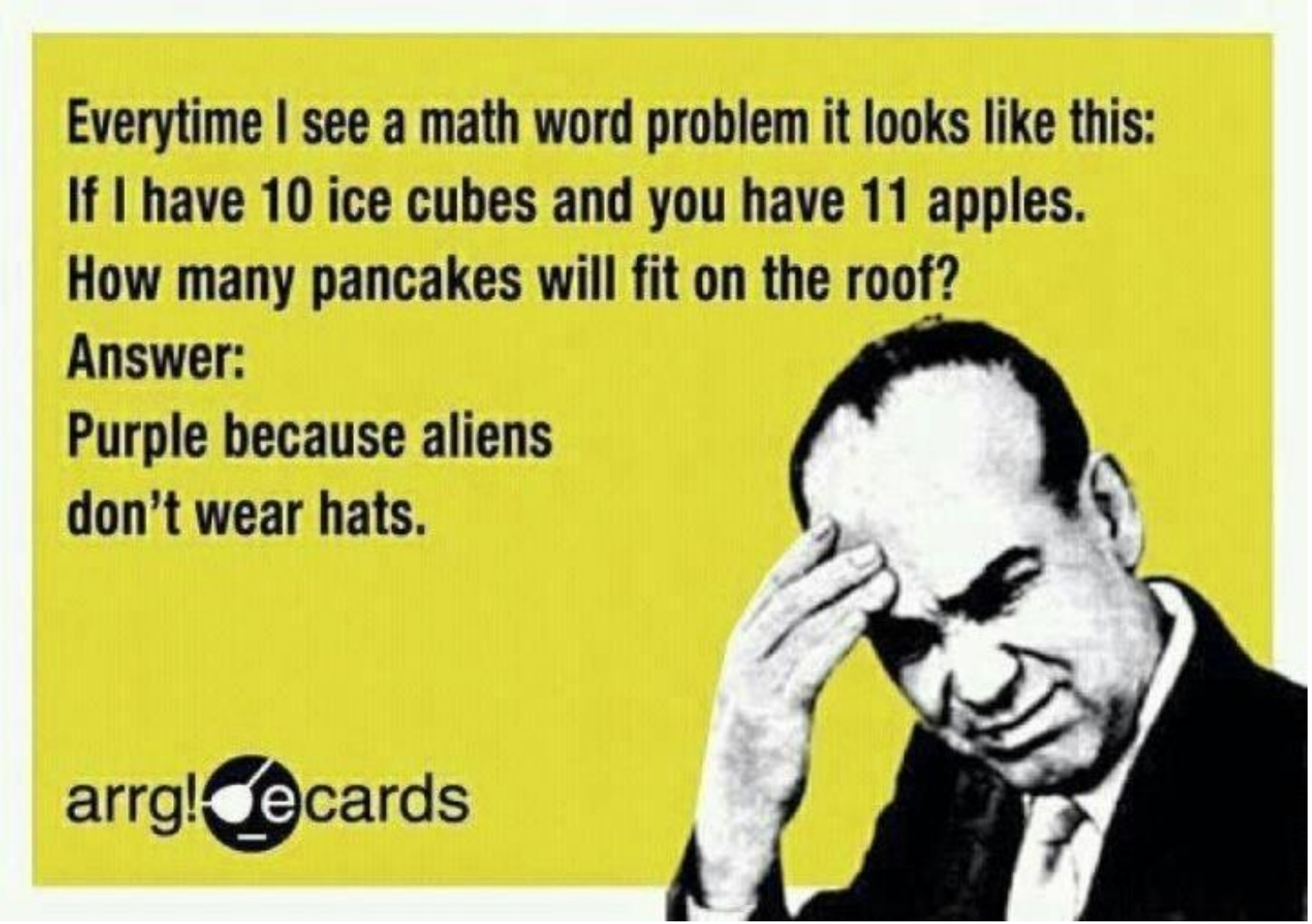
$$1-((1-.31)^{13}) = 99\%$$

$$1-((1-.10)^{13}) = 75\%$$

Planning Formative Sample Sizes

Problem Detection Probability	Likelihood of Detecting the Problem at Least Once					
	.50	.75	.85	.90	.95	.99
.01	69	138	189	230	299	459
.05	14	28	37	45	59	90
.10	7	14	19	22	29	44
.15	5	9	12	15	19	29
.25	3	5	7	9	11	17
.50	1	2	3	4	5	7
.90	1	1	1	1	2	2

For example, if n=6, the likelihood of observing very rare problems is low, for common problems is high, and for intermediate problems is decent. If all you can afford to study are six (or even three) participants, it will generally be worth it.


Everytime I see a math word problem it looks like this:
If I have 10 ice cubes and you have 11 apples.
How many pancakes will fit on the roof?
Answer:
Purple because aliens
don't wear hats.
arrg!ecards

Sample Size for Precision

Sample Size for Precision

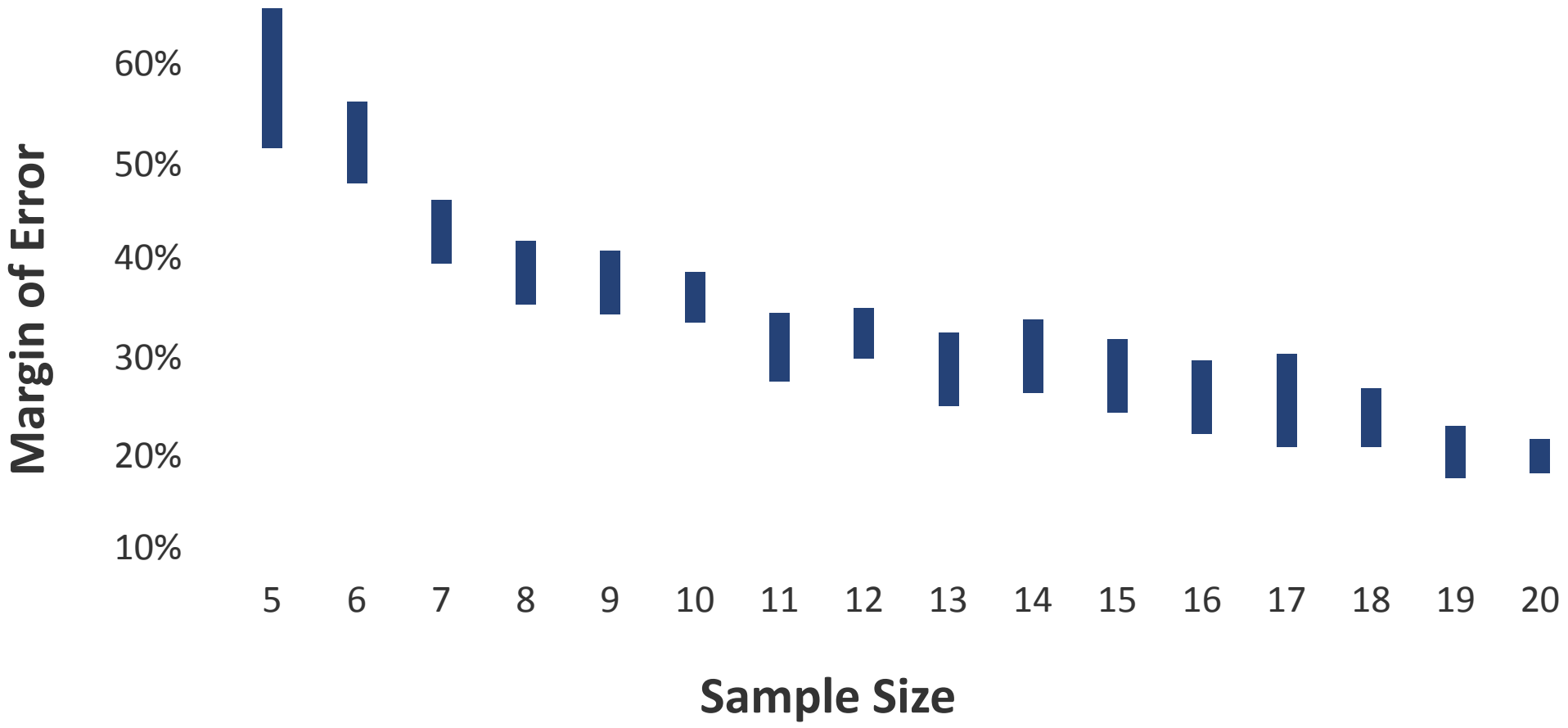
How Precise do we need to be?

- Within 1%?
- Within 5%?
- Within 10%?
- Within 20%?

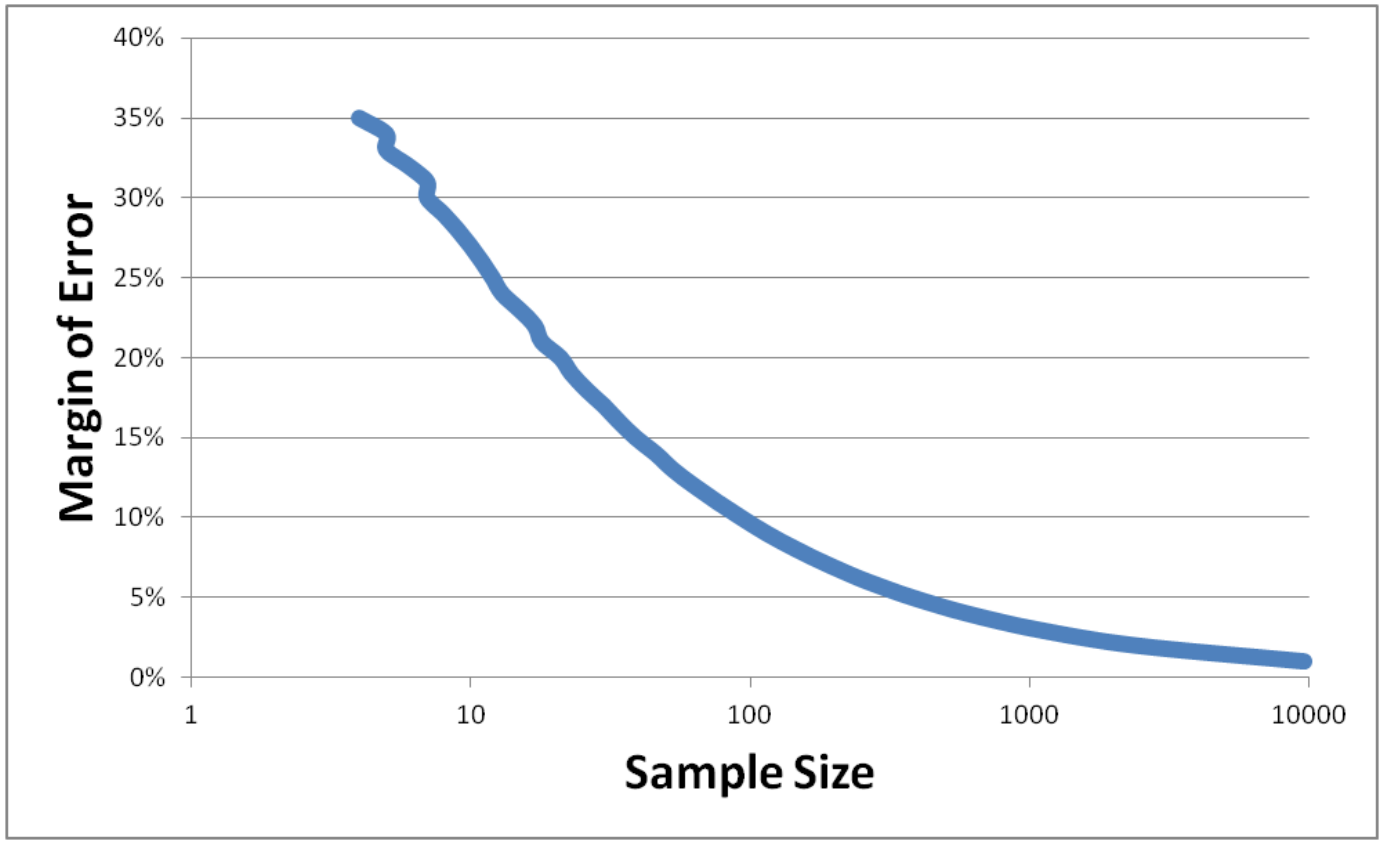
People prefer precise to imprecise measurement, but other things being equal:

- The more precise a measurement is, the more it will cost
- Running more participants than necessary is a waste of resources

20/20 Rule of Precision



To Cut margin of error in half – quadruple sample size



Sample Size Needed for 95% Level of Confidence

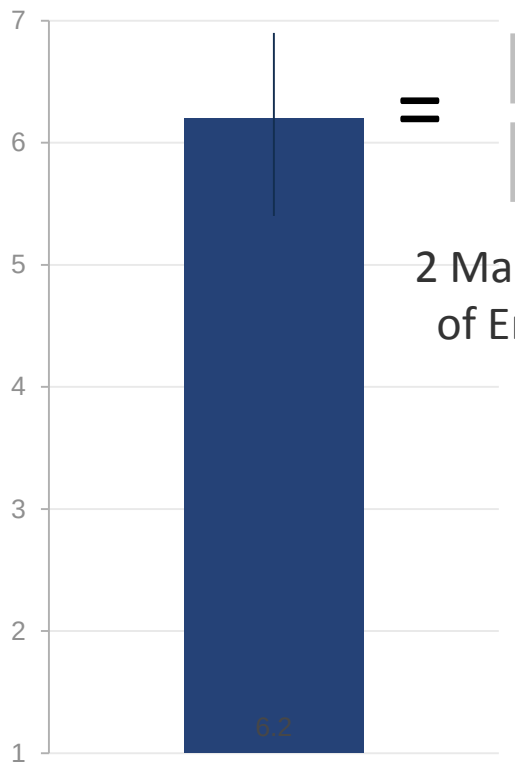
Sample size for precision table

Margin of Error	90% Confidence (+/-) Sample Size	95% Confidence (+/-) Sample Size
24%	10	13
20%	15	21
15%	28	39
14%	32	46
13%	38	53
12%	45	63
11%	54	76
10%	65	93
9%	81	115
8%	103	147
7%	136	193
6%	186	263
5%	268	381
4%	421	597
3%	749	1,064

Assumes a 50% Proportion/Completion Rate

95% Confidence Interval

1 Confidence Interval



= 2 Margins of Error

= ~ 4 Standard Errors

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

Margin of Error (d)

Sample Size for a Parameter

CONFIDENCE INTERVAL

$$\text{Margin (d)} = \frac{s}{\sqrt{n}} * t$$

ALGEBRA

$$\frac{d}{t} = \frac{s}{\sqrt{n}}$$

$$d * \sqrt{n} = t * s$$

$$\sqrt{n} = \frac{t * s}{d}$$

SAMPLE SIZE

$$n = \frac{t^2 * s^2}{d^2}$$

4 Steps to Sample Size Nirvana

1. How are you asking the questions?

   Rating Scale, Time, Binary?

2. How precise do you need to be?

   +/- 5%? +/-3%?

3. Estimate the Standard Deviation

   Use .5 binary or historical for scales

4. How confident do we need to be?

   Start with 90% and go up or down as needed

Sample Size for Comparisons

It’s all about the **SIZE** of the difference

Difference in Average Height?

Difference in Average Height?

What is the critical difference to detect?

YOUR DECISION	Is a Difference *Guilty*	No Difference *Innocent*
Difference!! *Convict*	✚ **Power 1- Beta = .80**	Type II ✕✕ Alpha = .10
No Difference! *Acquit*	Type II ✕ Beta= .20	✚

REALITY

4 Ingredients for Computing Sample Size

1. Confidence Level
2. Difference you want to Detect (d)
3. Variability of Groups (s)
4. Power (Confidence in Detecting a Difference)

2 sided Confidence *t* + 1 sided Power *t*

Sample Size Formula

$$n = \frac{2s^2 * t^2}{d^2}$$

4 Ingredients for Computing Sample Size

What sample size do you need for a comparative task based study to detect a difference of 10%?

1. Confidence Level: **95%**
2. Difference You Want to Detect (d): **.10**
3. Variability of groups (s): **.5**
4. Power (Confidence in Detecting a Difference): **.80**

2 sided Confidence *t* + 1 sided Power *t*

$$n = \frac{2s^2 * t^2}{d^2} = \frac{2(.5^2) * (1.96 + .84)^2}{10^2} =$$

390 each sample or
780 total

Sample Size for Comparing Means or Proportions

p-value from test statistics

$$t = \frac{d}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}}$$

Simplify for equal variances
and equal sample sizes

$$t = \frac{d}{\sqrt{2\left(\frac{s^2}{n}\right)}}$$

Algebra

$$t^2 = \frac{d^2}{2\left(\frac{s^2}{n}\right)}$$

Sample Size

$$2t^2 = \frac{d^2}{\frac{s^2}{n}} \qquad \frac{2t^2s^2}{n} = d^2 \qquad \frac{2t^2s^2}{d^2} = n$$

Power is the Probability We Can Detect a Difference

- We set alpha typically to .05
- We set beta typically to .20

OR

- We set our confidence level to .95
- We set out power to .80

Power is the Probability We Can Detect a Difference

A	B
95% Confidence Levels	90% Confidence Level
80% Power	70% Power
10% Point Difference	1% Point Difference
Standard Deviation = 50	Standard Deviation = 10
5 second Difference	30 Second Difference

5 Steps to Sample Size Nirvana

1. How are you asking the questions?
   Rating Scale, Binary?
2. What size of a difference do you want to detect?
   .1 point, .5 point? +/-5%? +/-3%
3. Estimate the Standard Deviation
   Use .5 for binary (or other%) or historical for scales
4. How confident do we need to be that this difference isn't due to chance?
   Start with 90% and go up or down as needed
5. How confident do we need to be that if we don’t see a difference this large, it wasn’t due to chance?
   Start with 80% and go up or down as needed

5 Steps to Sample Size Nirvana

What sample size would you need to detect a .5 difference in a satisfaction on a 5 point scale?

1. How are you asking the questions?
   Rating Scale, Binary?
2. What size of a difference do you want to detect?
   .1 point? **.5 point**? +/-5%? +/-3%?
3. Estimate the Standard Deviation
   Use .5 binary (or other %) or **historical 1.3 for 5 point for scales**
4. How confident do we need to be that this difference isn't due to chance?
   Start with **90% confidence** and go up or down as needed
5. How confident do we need to be that if we don’t see a difference this large, it wasn’t due to chance?
   Start with **80% power** and go up or down as needed

You should plan on having 86 responses in each group for a total sample of 172

5 Steps to Sample Size Nirvana

What sample size would you need to detect a 4% point different in recommendation rates, assuming the current recommendation rate is around 70%?

1. How are you asking the questions?
   Rating Scale, **Binary?**
2. What size of a difference do you want to detect?
   .1 point? .5 point? **+/-4%**? +/-3%?
3. Estimate the Standard Deviation
   Use .5 binary (**or other % .7**) or historical 1.3 for 5 point for scales
4. How confident do we need to be that this difference isn't due to chance?
   Start with **90% confidence** and go up or down as needed
5. How confident do we need to be that if we don't see a difference this large, it wasn't due to chance?
   Start with **80% power** and go up or down as needed

You should plan on having 1559 responses in each group for a total sample of 3118

Sample Size Exercises

More Sample Size Exercises

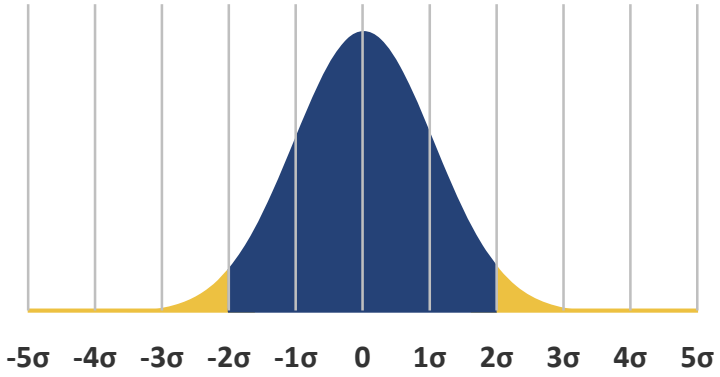
- What sample size would you need to detect a .2 point different in satisfaction on a 5 point scale?
- You launch a survey pretest to 500 randomly selected customers and 47 respond. How many customers do you need to email to have a margin of error of 5%?
- What sample size would you need to detect a 2% point increase in referral rates, assuming your current referral rate is 35%?
- What sample size would you need to detect a .1 point increase in satisfaction on a 5 point scale? Last year the average score was a 4.4 (sd=.8).

Confidence and Power

Z-Score: 2.001
Mean: 100

X-Value: 132.016
SD: 16

Blue Area: 95.4581%
Yellow Area: 4.5419%

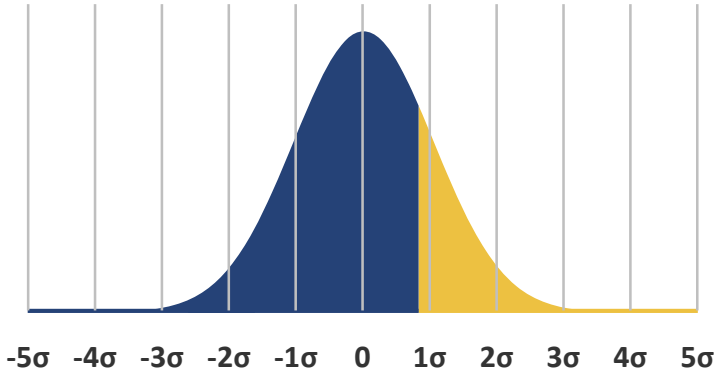


2 Tailed Confidence

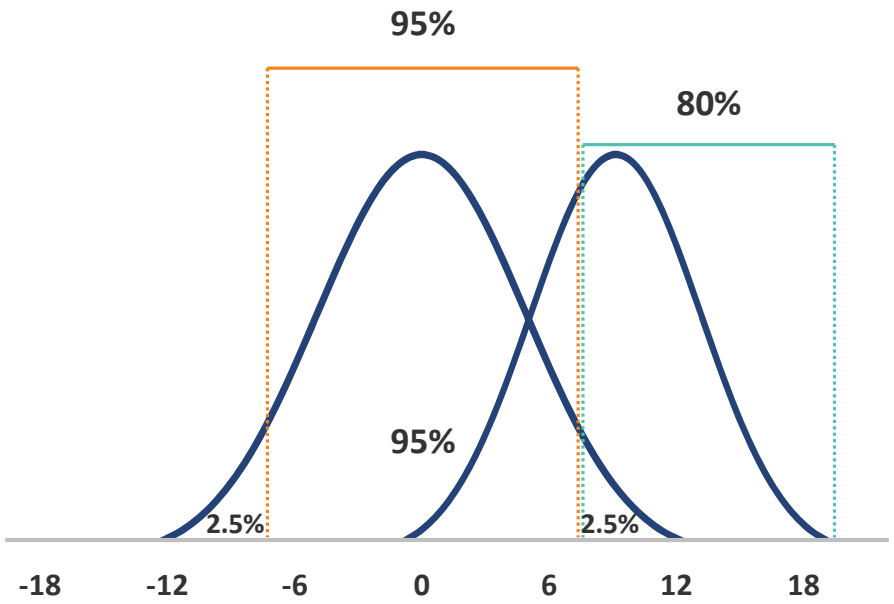
Z-Score: 0.842
Mean: 100

X-Value: 113.472
SD: 16

Blue Area: 80.0096%
Yellow Area: 19.9904%



1 Tailed Power



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

Denver UX BOOTCAMP

UX Boot Camp Aug 16th-18th
denverux.com

MeasuringU.com
@MeasuringU