Data Uncertainty

INFO-1301, Quantitative Reasoning 1

University of Colorado Boulder

March 13, 2017

Prof. Michael Paul

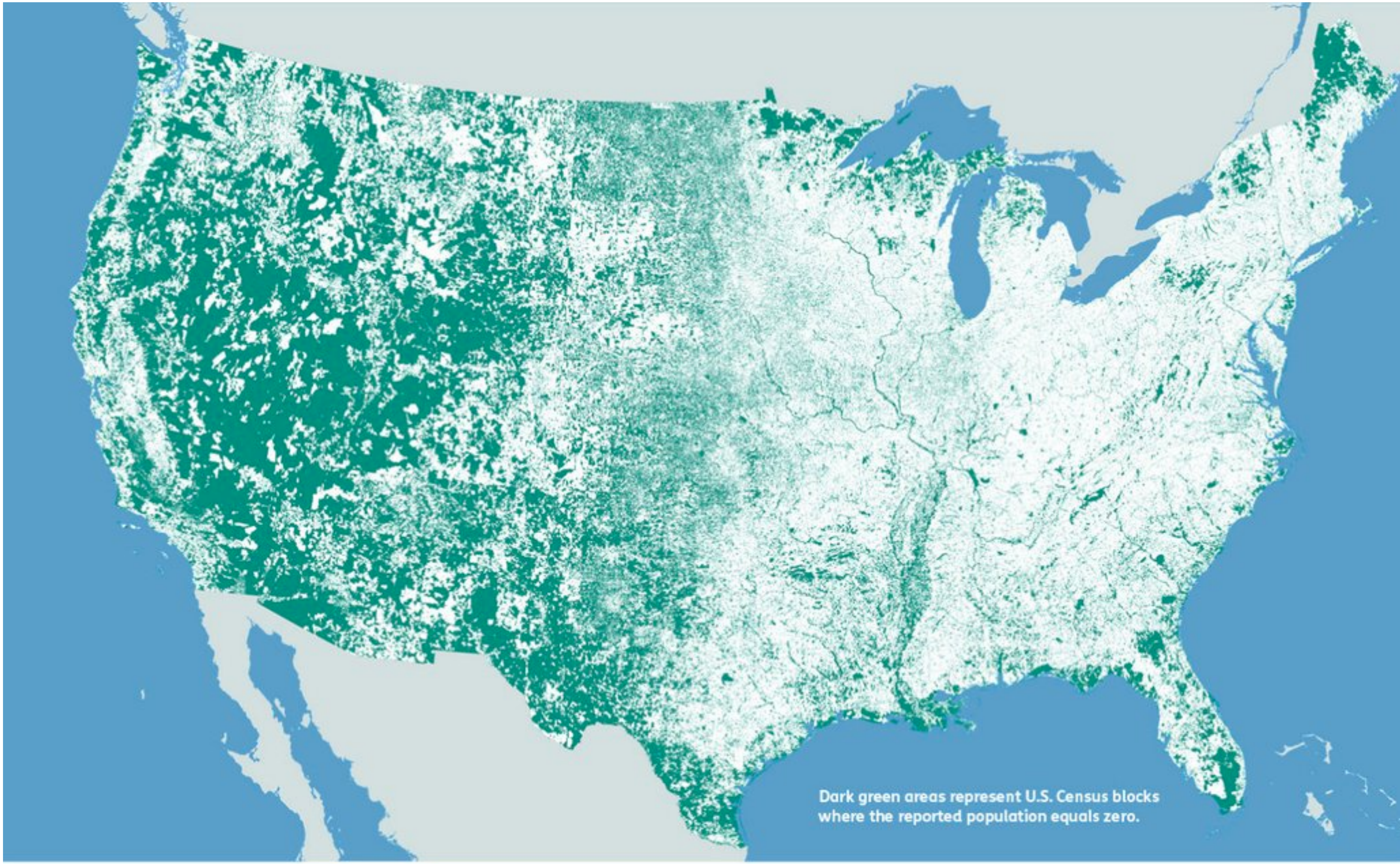
Measurement Error

When we construct data, how accurate are the values? How do we know?

- I am 72 inches tall (to the nearest 1/8 inch)
- My Subaru gets 22 miles per gallon on average (which is a 5% overestimate, on average!)
- I do not have strep throat (according to a test that has a false negative rate of 10-20%)

Measurement Error

"Even simple counts break down when you have to count a lot of things. We've all sensed that population figures are somewhat fictitious. Are there really 536,348 people in your hometown, as the number on the 'Welcome to …' sign suggests? If the sign said 540,000 we would know to treat it as a rough figure, yet far too often we're willing to imagine that every last digit is accurate."


Dark green areas represent U.S. Census blocks
where the reported population equals zero.

Measurement Error

Common sources of error:

- Limitations of measurement device (lab test, sensor)
- Reliability of human feedback
  - Misrepresentation (intentional or unintentional)
  - Memory limitations
  - Misunderstanding or skipping a prompt
- Sampling from a population
  - Size of sample
  - Bias in sample

Measurement Error

Often one kind of error is more harmful than the other

Two types of errors:

- **False positive**
  - You incorrectly identified something as something
  - Ex: Test says you have strep, but you don't
- **False negative**
  - You incorrectly identified something as *not* something
  - Ex: Test says you don't have strep, but you do

Need to specify what is considered "positive"

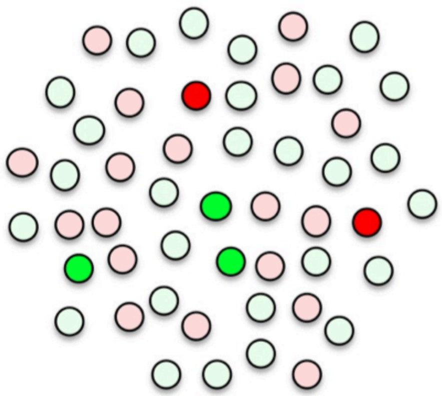
- Diagnostic tests
- Screenings (e.g., security, quality control)

Quantifying Errors

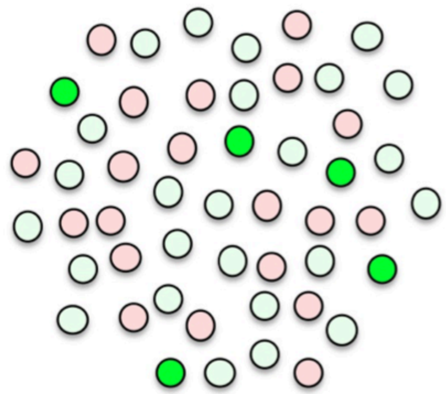
We can often assign a probability to different errors, and use this probability to communicate our certainty about how accurate a value is

- For binary values (e.g., positive vs negative), probability of having a false positive and/or false negative
- For continuous values (e.g., unemployment rate), probability of making an overestimate or underestimate of a certain value
  - Errors often form a bell curve, which we can use to quantify the probability of different sizes of errors

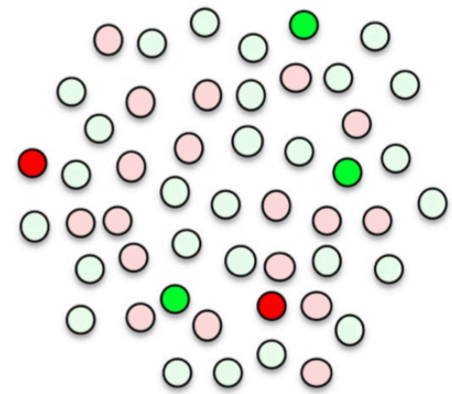
Sampling Error



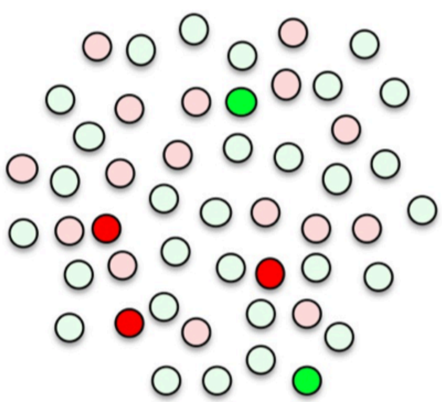
40%



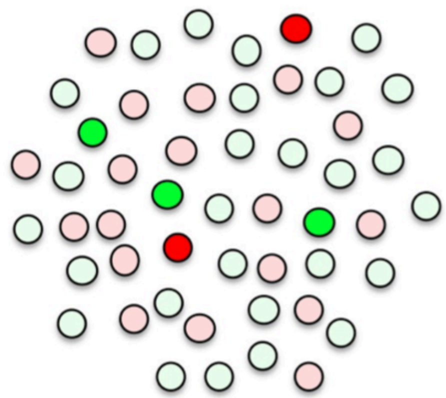
0%



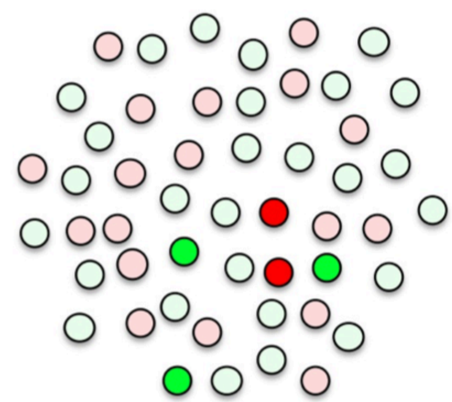
40%



60%

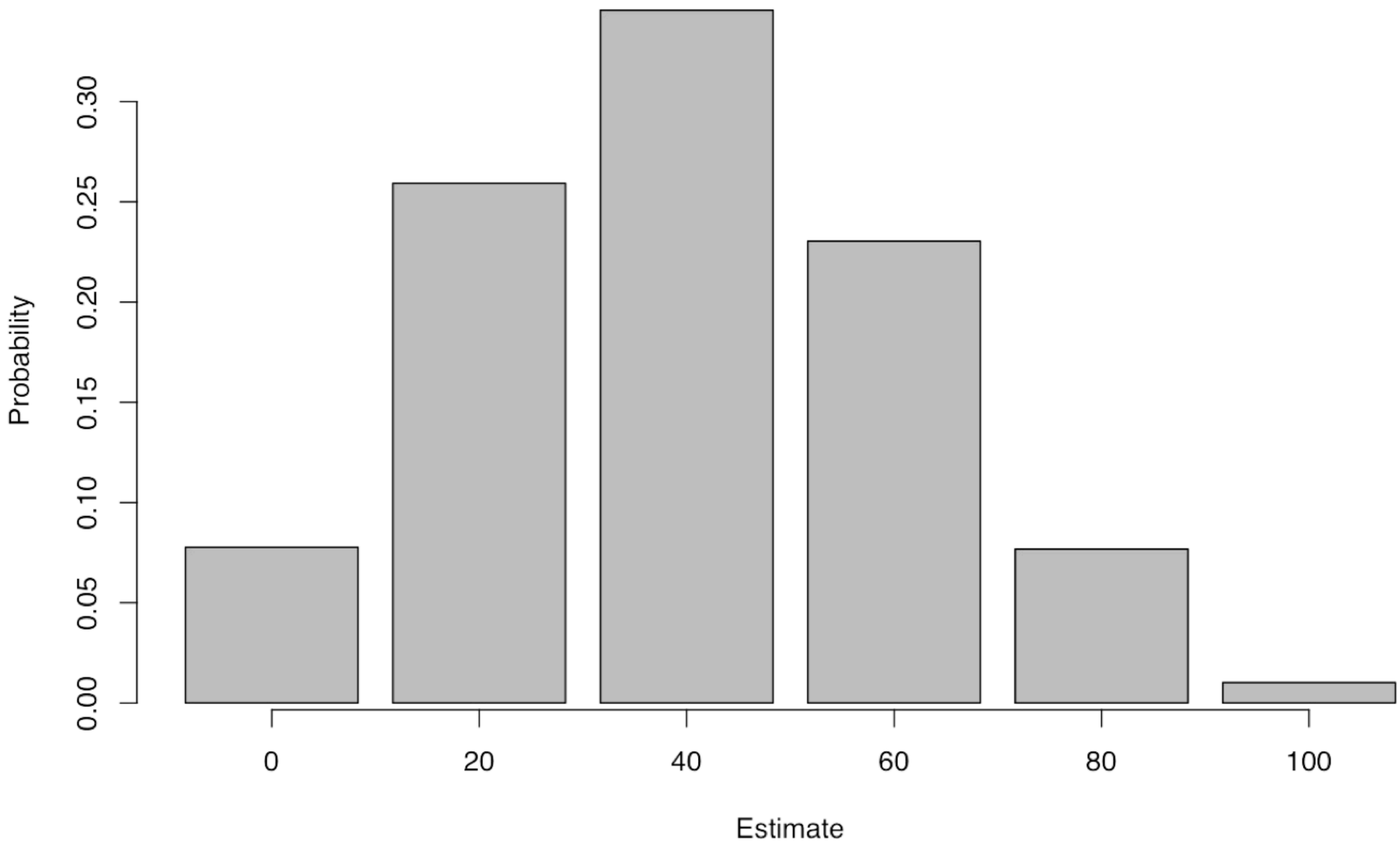


40%

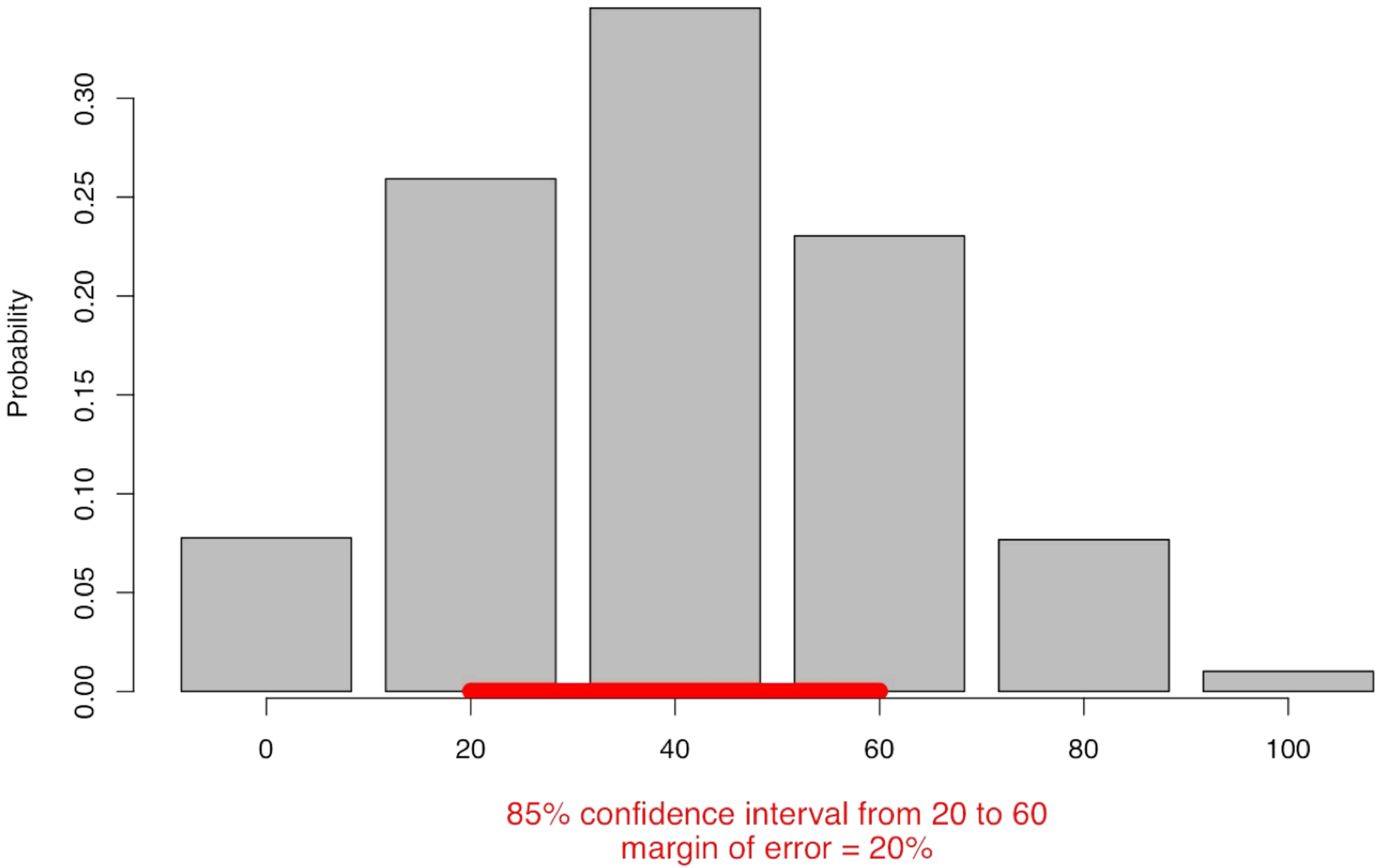


20%

Sampling Error



Sampling Error



Quantifying Errors

A **confidence interval (CI)** is a range of values that the true value is most likely to be between

The **margin of error** is the distance from the estimated value to boundaries of the confidence interval

- Always half the width of the confidence interval (if the CI is symmetric)

Ex: we are confidence that the true value is:

- inside the interval $\boxed{(20, 60)}$ ← **This is the CI; Width of CI = 40**
- $40 \pm \boxed{20}$ ← **Margin of error**

Quantifying Errors

Confidence intervals and margin of error tell you the *size* of the error

We also need to express the *probability* that the error will be of that size

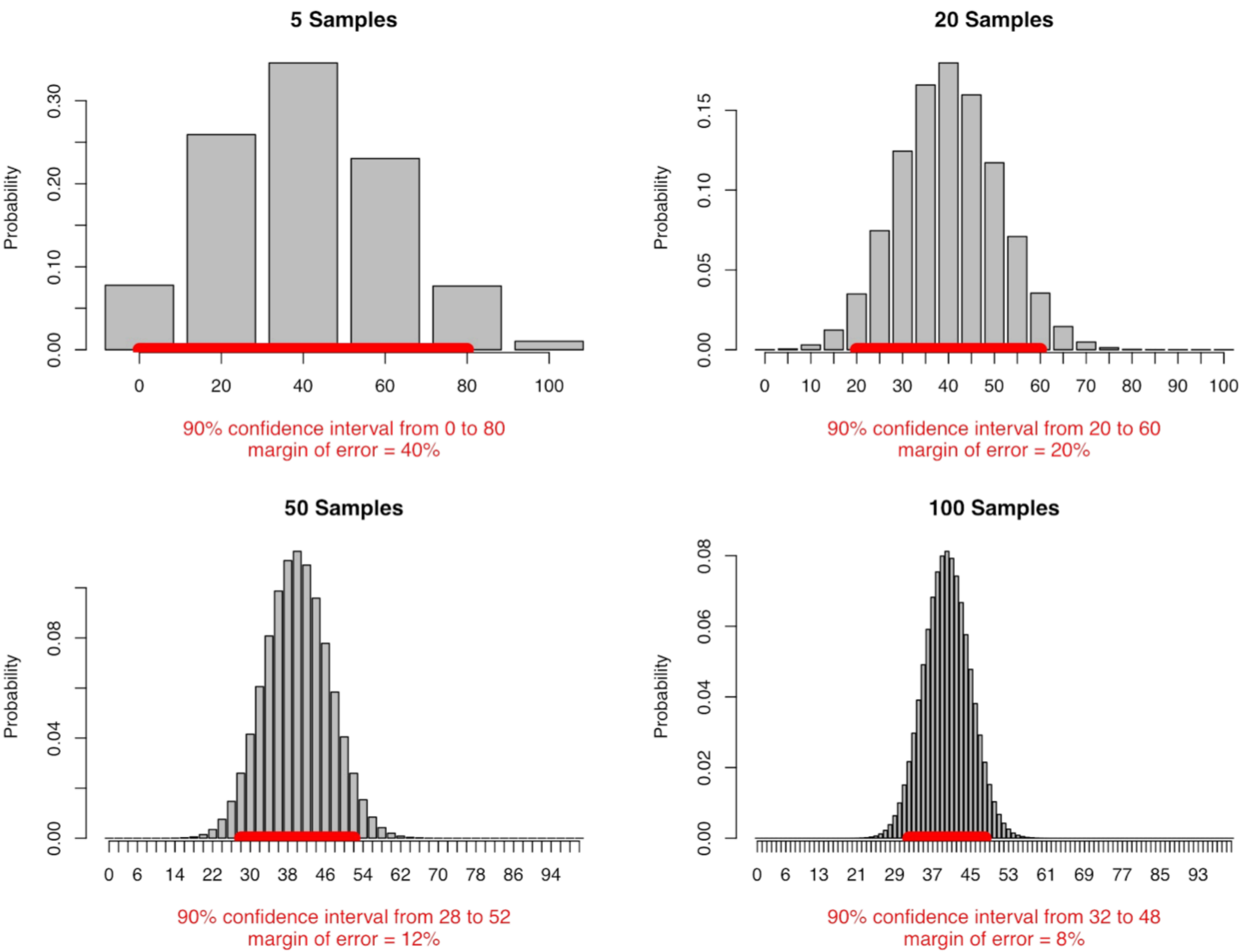
The **confidence level** is the probability that:

- the true value is within the confidence interval
  “the 85% confidence interval is (20, 60)”
- the error in measuring the true value is at least as small as the margin of error
  “the margin of error is 20 at 85% confidence”

Confidence

The size/width of a confidence interval depends on three factors:

1. The variability in your data
   - Don't worry about this for now – we'll see this mathematically later this month
2. The size of your sample
   - Larger sample → smaller confidence interval
   - Size of population does *not* affect your confidence interval (unless the size of population affects variability in #1)
3. The confidence level
   - Higher confidence level → wider confidence interval


5 Samples
Probability
0.00
0.10
0.20
0.30
0 20 40 60 80 100
90% confidence interval from 0 to 80
margin of error = 40%
20 Samples
Probability
0.00
0.05
0.10
0.15
0 10 20 30 40 50 60 70 80 90 100
90% confidence interval from 20 to 60
margin of error = 20%
50 Samples
Probability
0.00
0.04
0.08
0 6 14 22 30 38 46 54 62 70 78 86 94
90% confidence interval from 28 to 52
margin of error = 12%
100 Samples
Probability
0.00
0.02
0.04
0.06
0.08
0 6 13 21 29 37 45 53 61 69 77 85 93
90% confidence interval from 32 to 48
margin of error = 8%