Quantifying Chance
Part 1: Sampling Variability

INFO-1301, Quantitative Reasoning 1

University of Colorado Boulder

March 22-24, 2017

Prof. Michael Paul

Estimating Data

We've discussed measurement error in this class

Common source of error: randomness

- What if the value or result was due to chance?

Common source of randomness: sampling

- How reliable is your estimate from a sample?

Estimating Data

Population statistics vs sample statistics

- e.g. **population mean** vs **sample mean**

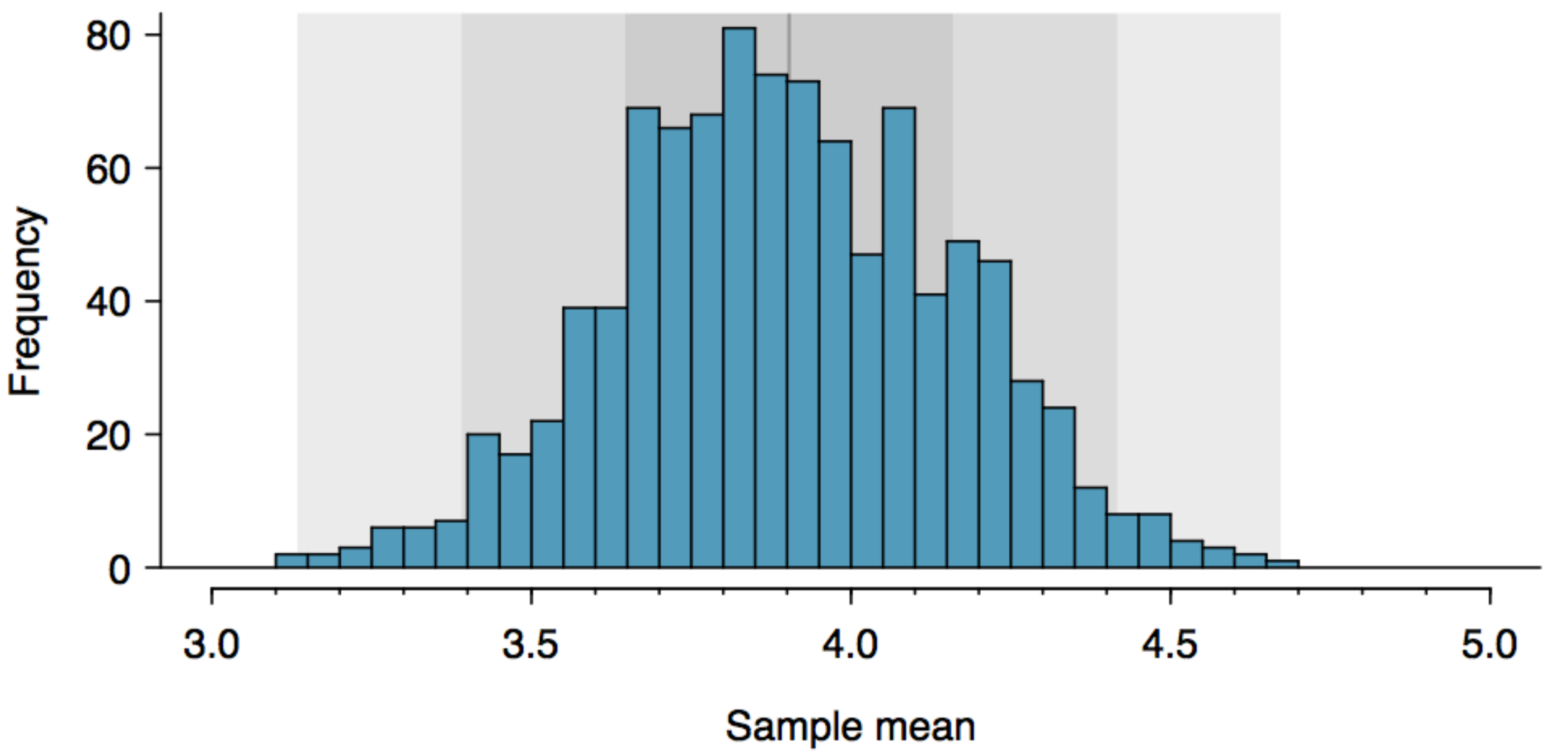
Population statistics have one true value, but you might not be able to measure it

Sample statistics are *estimates*

- You will get different estimates from different samples
- Any one estimate is called a **point estimate**

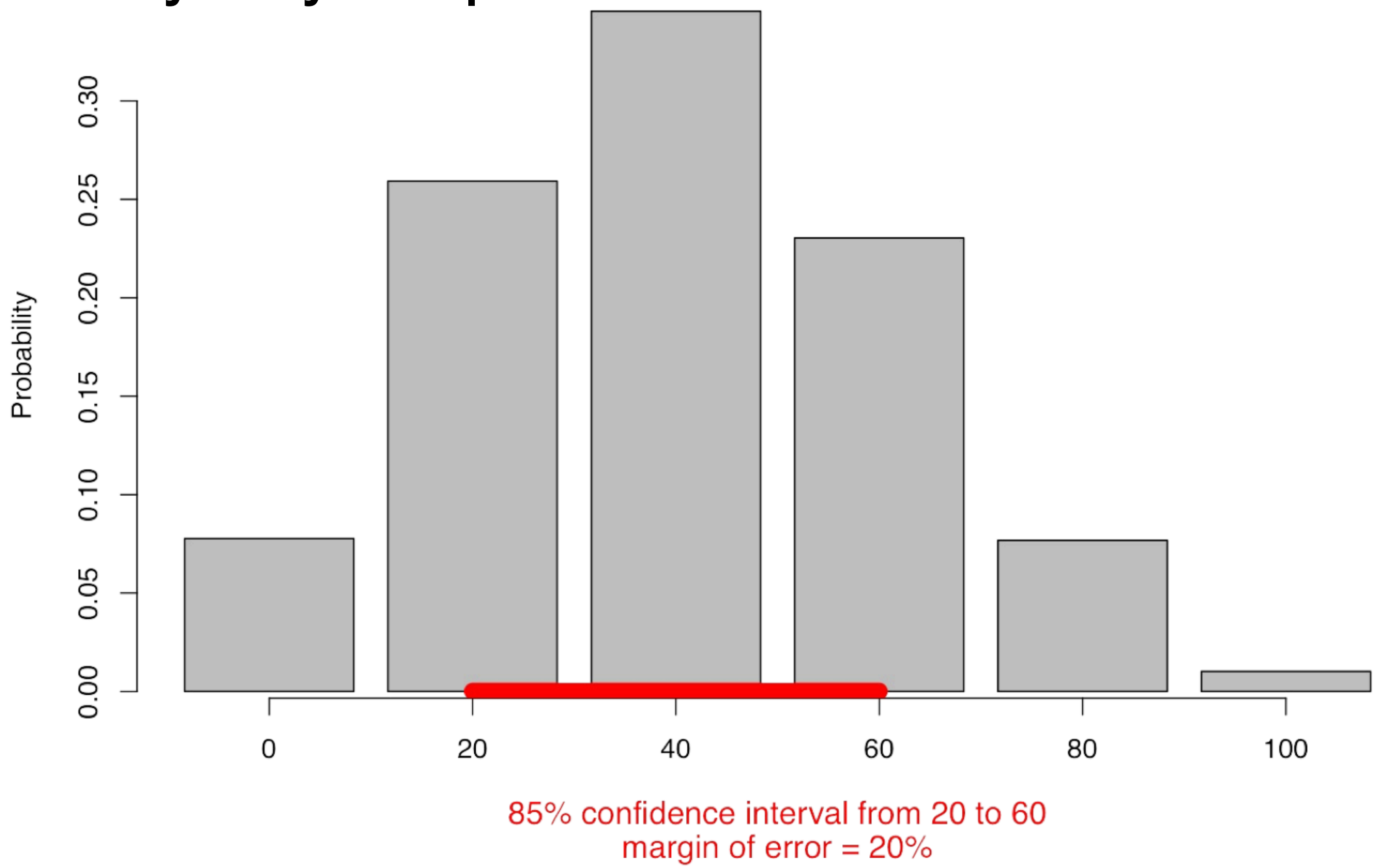
Estimating Data

The **sampling distribution** is the distribution of all point estimates you would get from the different possible samples



Estimating Data

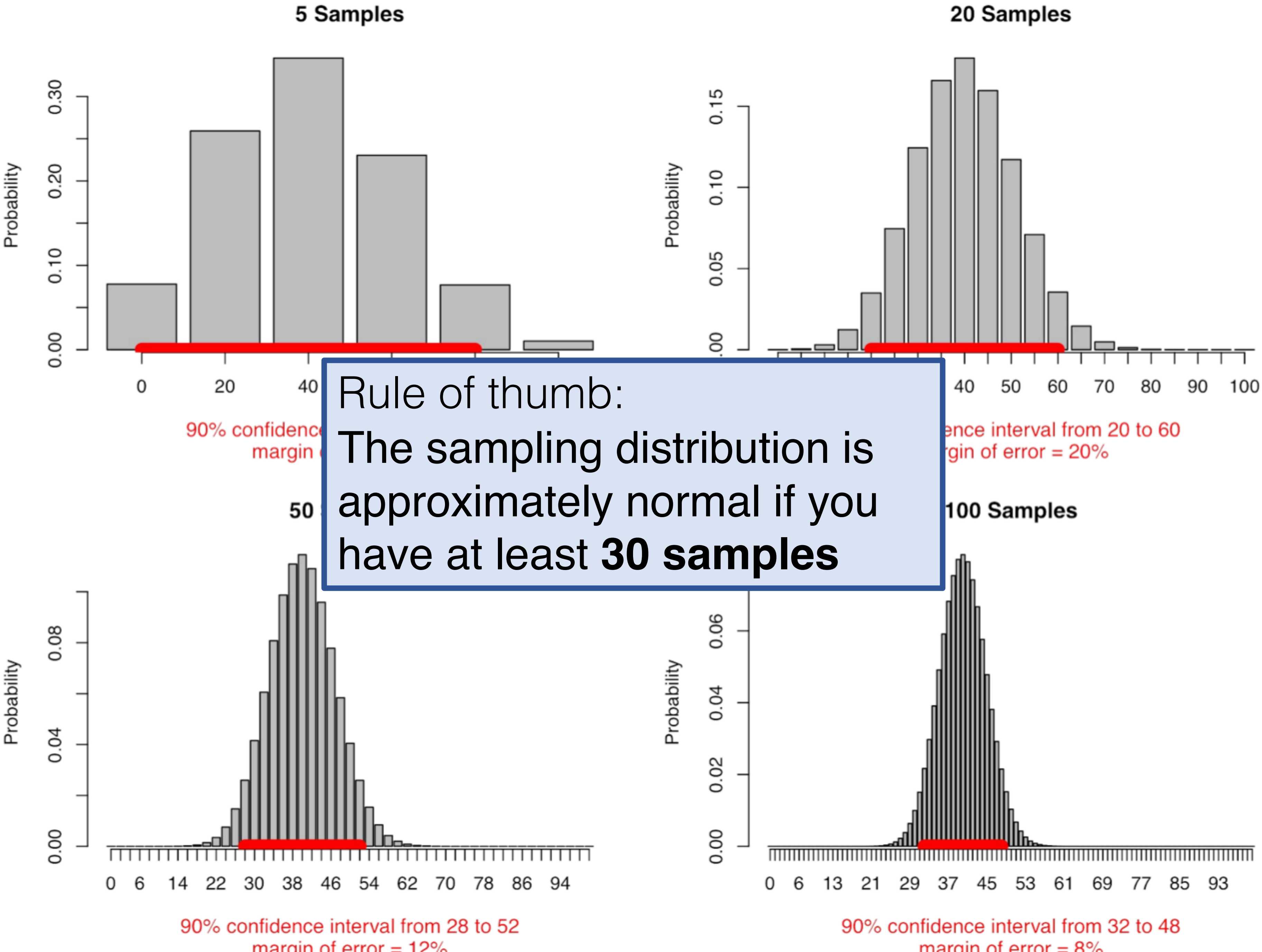
The **sampling distribution** tells you about the variability of your point estimates.



Estimating Data

But how do we get the sampling distribution?

1. Get point estimates from all possible combinations of samples
    - Not even a little practical
2. Take multiple samples to get an approximate distribution
    - For example, 100 different samples of the same size
    - Not common though – defeats the purpose of sampling
3. Normal approximation
    - Turns out the sampling distribution is a normal curve!



Rule of thumb:
The sampling distribution is approximately normal if you have at least **30 samples**

Sampling Distribution

The sampling distribution is approximately normal

- The mean is the true population mean
- The standard deviation is called the **standard error (SE)**

$$SE = \frac{\sigma}{\sqrt{n}}$$

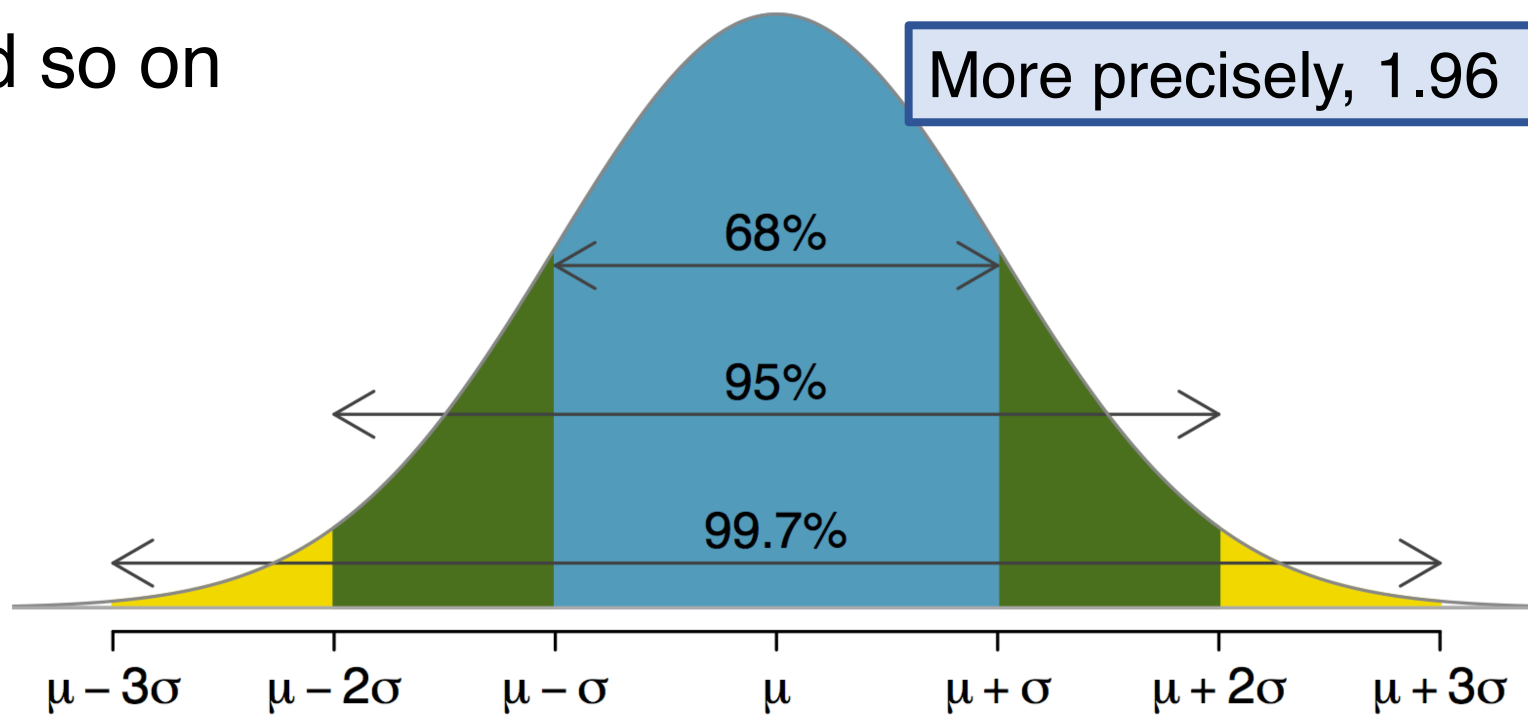
This is known as the **Central Limit Theorem**

- **σ** is the standard deviation of your data (unknown – so use the standard deviation from your sample)
- **n** is the size of your sample
  - Larger n → smaller standard error
    (sample mean is more likely to be close to population mean)

What can we do with this?

- 68% of sample statistics will be correct within 1 SE of the true mean
- 95% of samples will be will be within 2 SEs
- And so on

More precisely, 1.96



What can we do with this?

Suppose you measure the length of 100 randomly sampled lizards, and find a mean of 14cm and a standard deviation of 3cm

Standard error = $3 / \sqrt{100} = 0.3$

$2 * SE = 0.6$

There is a 95% chance that our estimate of 14cm is within 0.6cm of the true average lizard length

What can we do with this?

Suppose you measure the length of 100 randomly sampled lizards, and find a mean of 14cm and a standard deviation of 3cm

Standard error = $3 / \sqrt{100} = 0.3$

2*SE = 0.6

The **margin of error** is 0.6
(at the 95% confidence level)

What can we do with this?

Suppose you measure the length of 100 randomly sampled lizards, and find a mean of 14cm and a standard deviation of 3cm

Standard error = $3 / \sqrt{100} = 0.3$

$2 * SE = 0.6$

The **95% confidence interval** is

$(14 - 0.6, 14 + 0.6) = (13.4, 14.6)$

or: 14 ± 0.6

Confidence

Confidence interval: **μ** ± Z*SE

Margin of error: Z*SE

- Where Z=2 (or 1.96) for 95% confidence level

For other confidence levels, solve for Z. (Find Z such that the middle area under the normal curve equals the confidence percentage.)

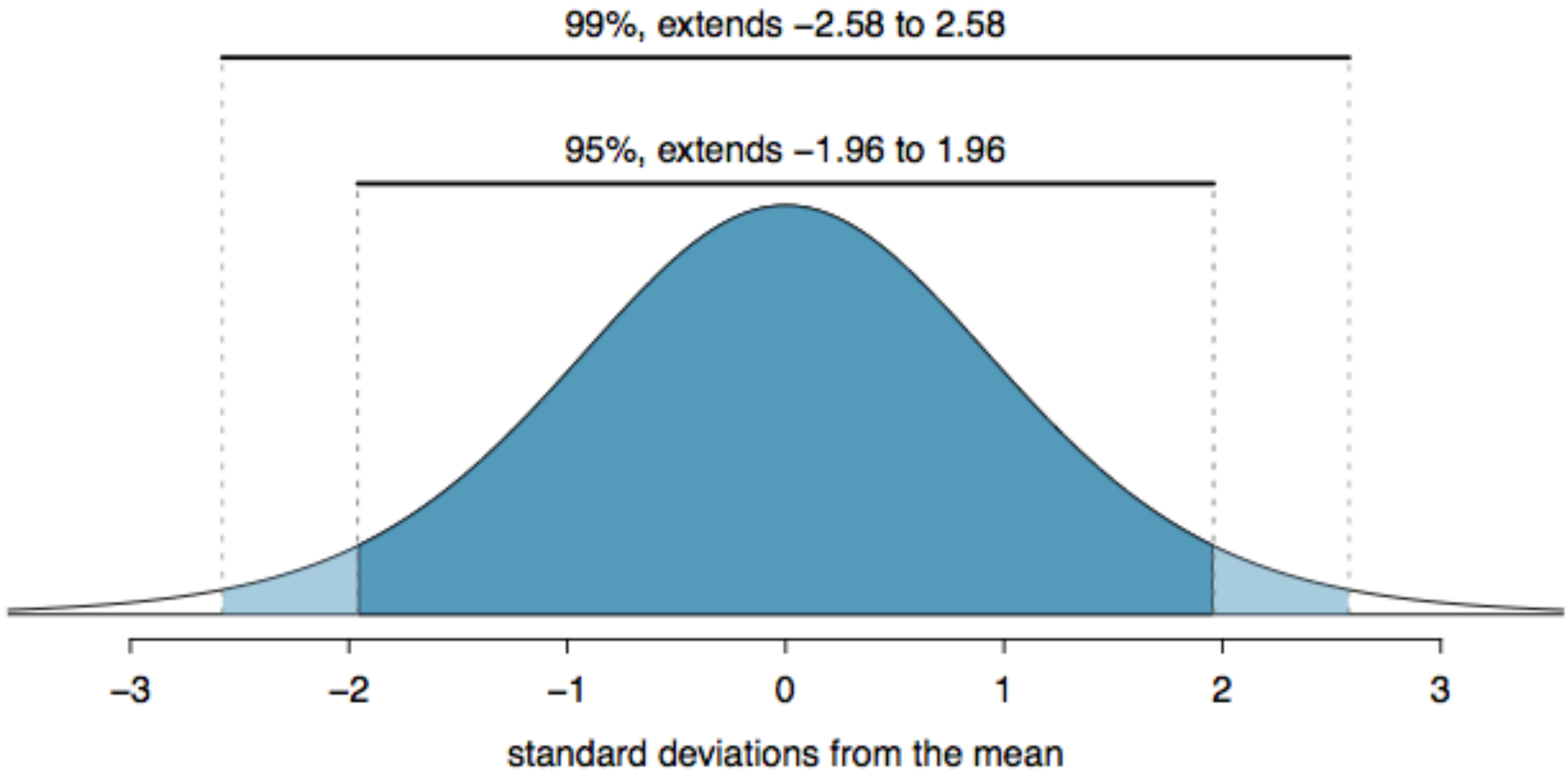


Figure 4.10: The area between -$z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

Confidence

Steps for identifying Z for a confidence level, C:

1. Calculate $X = 100 – C$
2. Calculate $P = 100 – X/2$
3. Find the cell in the Z-table that is closest to P

Example: 80% confidence level

$X = 20$

$X/2 = 10$

$P = (100 – 10) = 90$

Confidence

P = (100 – 20/2) = 90

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Z = 1.28

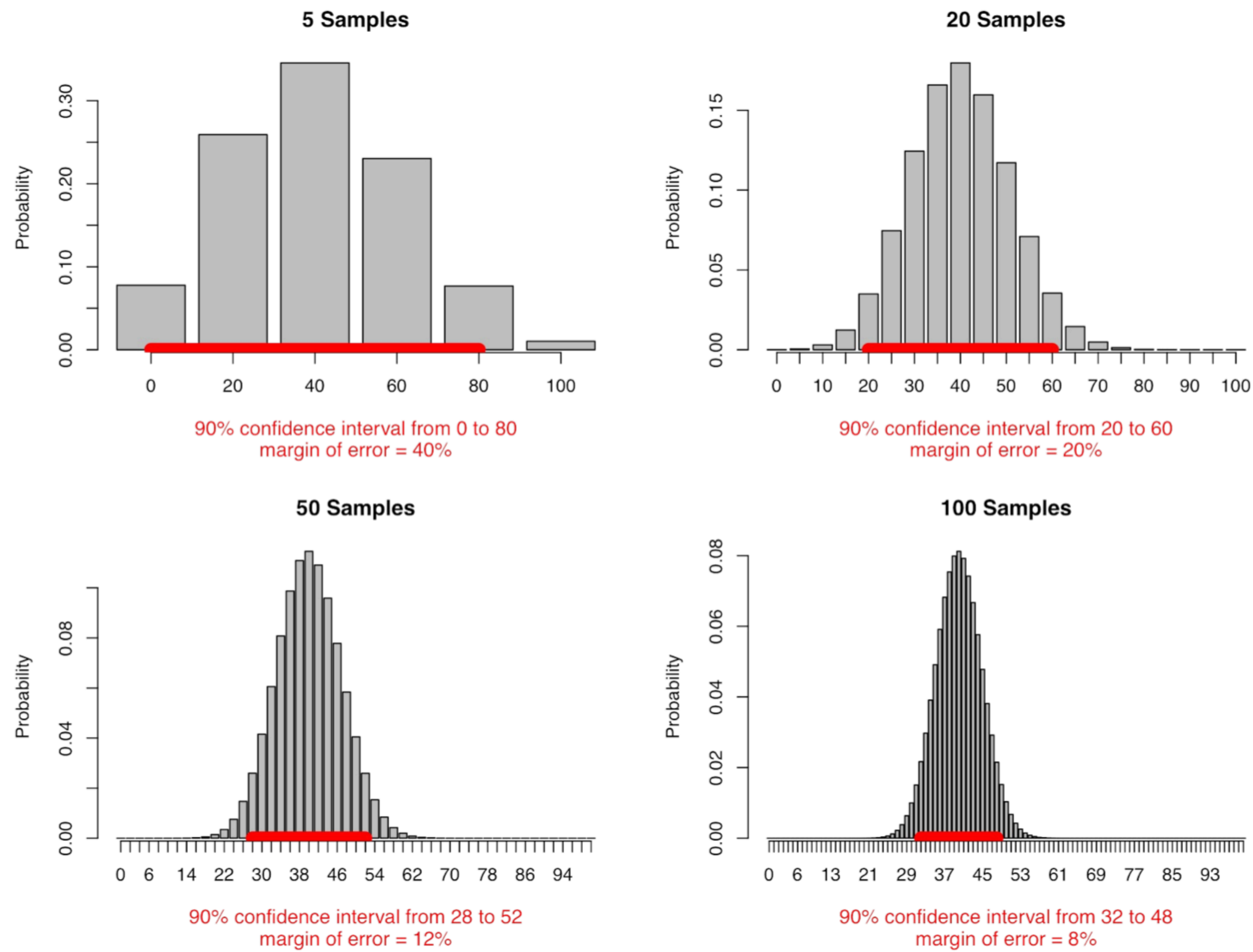
Confidence

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

1. The variability in your data
   - Higher variance of your data → smaller standard error
2. The size of your sample
   - Larger sample → smaller standard error

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

3. The confidence level
   - Higher confidence level → wider confidence interval (larger area under the normal curve)


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%

Practice 1

In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions". However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used in this setting. Create a 95% confidence interval for the proportion of U.S. adults who live with one or more chronic conditions.

45 ± 2.4

Practice 2(a)

The 2010 General Social Survey asked the question: “After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?” to a random sample of 1,155 Americans. A 95% confidence interval for the mean number of hours spent relaxing or pursuing activities they enjoy was (1.38, 1.92).

Interpret this interval in context of the data.

There is a 95% chance that the true mean is within this interval.

Practice 2(b)

The 2010 General Social Survey asked the question: "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans. A 95% confidence interval for the mean number of hours spent relaxing or pursuing activities they enjoy was (1.38, 1.92).

Suppose another set of researchers reported a confidence interval with a larger margin of error based on the same sample of 1,155 Americans. How does their confidence level compare to the confidence level of the interval stated above?

Higher confidence level

Practice 2(c)

The 2010 General Social Survey asked the question: “After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?” to a random sample of 1,155 Americans. A 95% confidence interval for the mean number of hours spent relaxing or pursuing activities they enjoy was (1.38, 1.92).

Suppose next year a new survey asking the same question is conducted, and this time the sample size is 2,500. How will the margin of error of the 95% confidence interval constructed based on data from the new survey compare to the margin of error of the interval stated above?

Smaller margin of error

Practice 3

Suppose your sample mean is 30, your sample standard deviation is 5, and your sample size is 100.

The standard error is 5/10 = 0.5.

The 95% margin of error therefore 2*0.5 = 1.

What is the 90% margin of error?

Find Z such that 90% of the area is covered.

When Z=1.65, the percentile is about 95%.

90% margin of error = 1.65*0.5 = .825