Statistics - Part 1

Feb. 4, 2025

Statistics - Part 1

Feb. 4, 2025

By the end of this lecture, you will be able to:

1. Give examples of statistical inference
2. Interpret and compute confidence intervals
3. Give examples of statistical bias

Statistical inference and confidence intervals

Statistics: the collection, analysis and interpretation of data

Population: the set of items or events of interest

Statistical inference is the process of using data drawn from parts of the population to infer properties of the whole population.

Example 1: Using polls to predict election results

Statistics: the collection, analysis and interpretation of data

Population: the set of items or events of interest

Statistical inference is the process of using data drawn from parts of the population to infer properties of the whole population.

Example 1: Using polls to predict election results

Example 2: test the efficacy of drugs or vaccines in clinical trials

Example 3: test autonomous driving systems with data collected from different road conditions

Statistics

- Estimation/Prediction

  “Best guess” of targets of interest

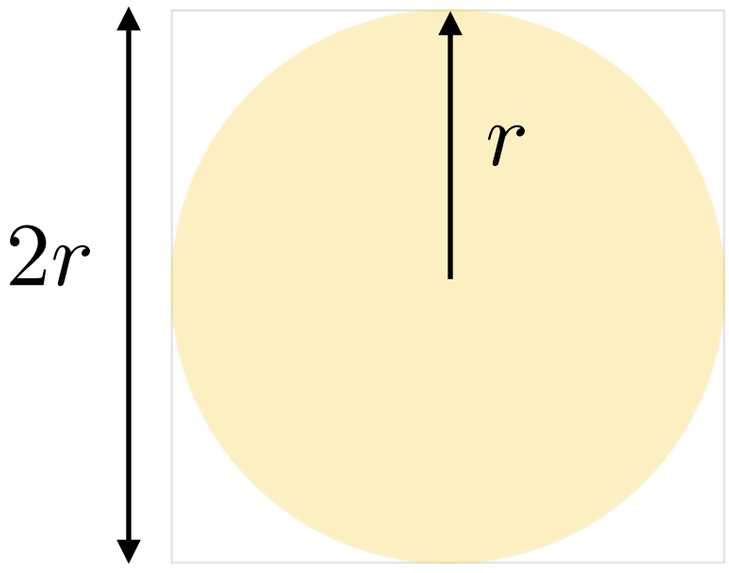
- Inference/Testing

  “Range with a high probability”

- Funds/Resources Constraint
- Sensitive Information
- Impossible/Missing

Point estimate: a single value calculated from the sample data as the “best guess” of the unknown parameter.

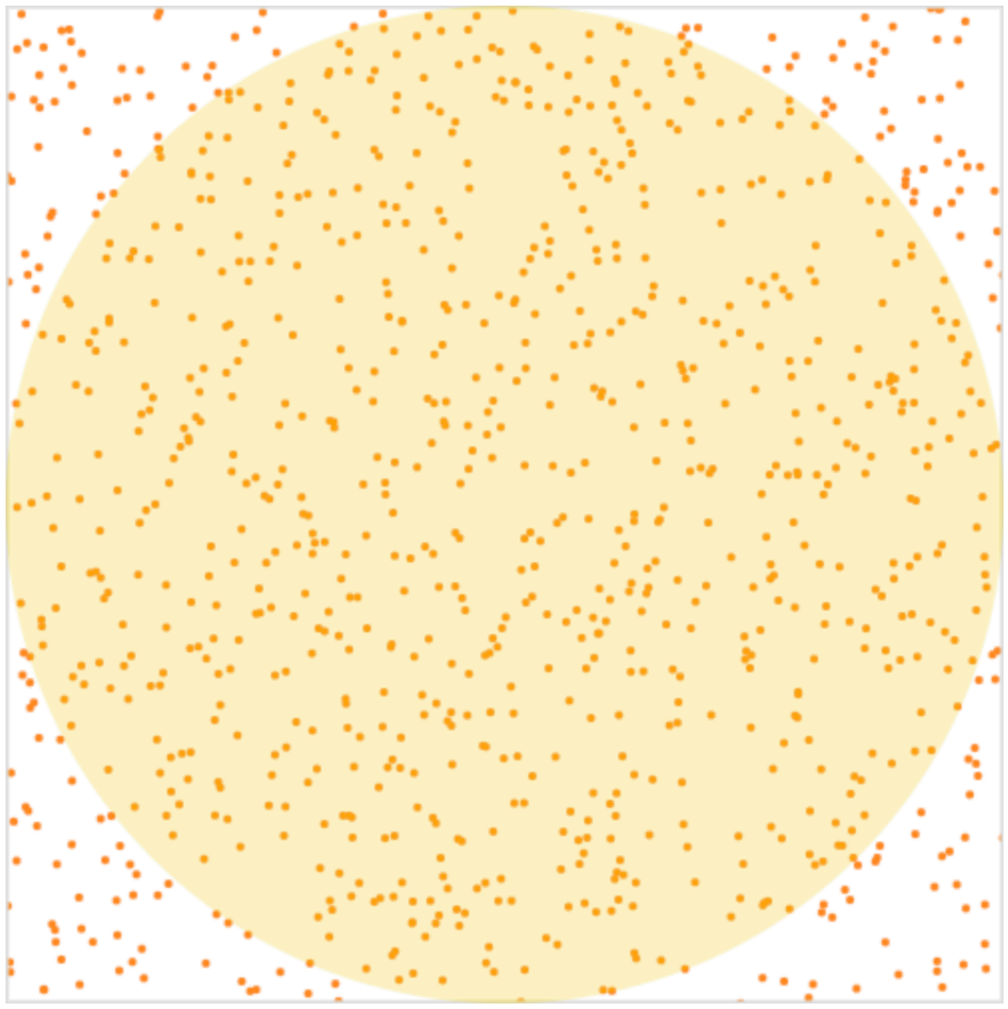
Example: estimate the value of π by uniformly throwing darts on a square containing an inscribed circle.



$$S_{\text{circle}} = \pi r^2$$

$$S_{\text{square}} = 4r^2$$

$$\pi = 4S_{\text{circle}}/S_{\text{square}}$$



m: # of samples in the circle

n: # of total samples

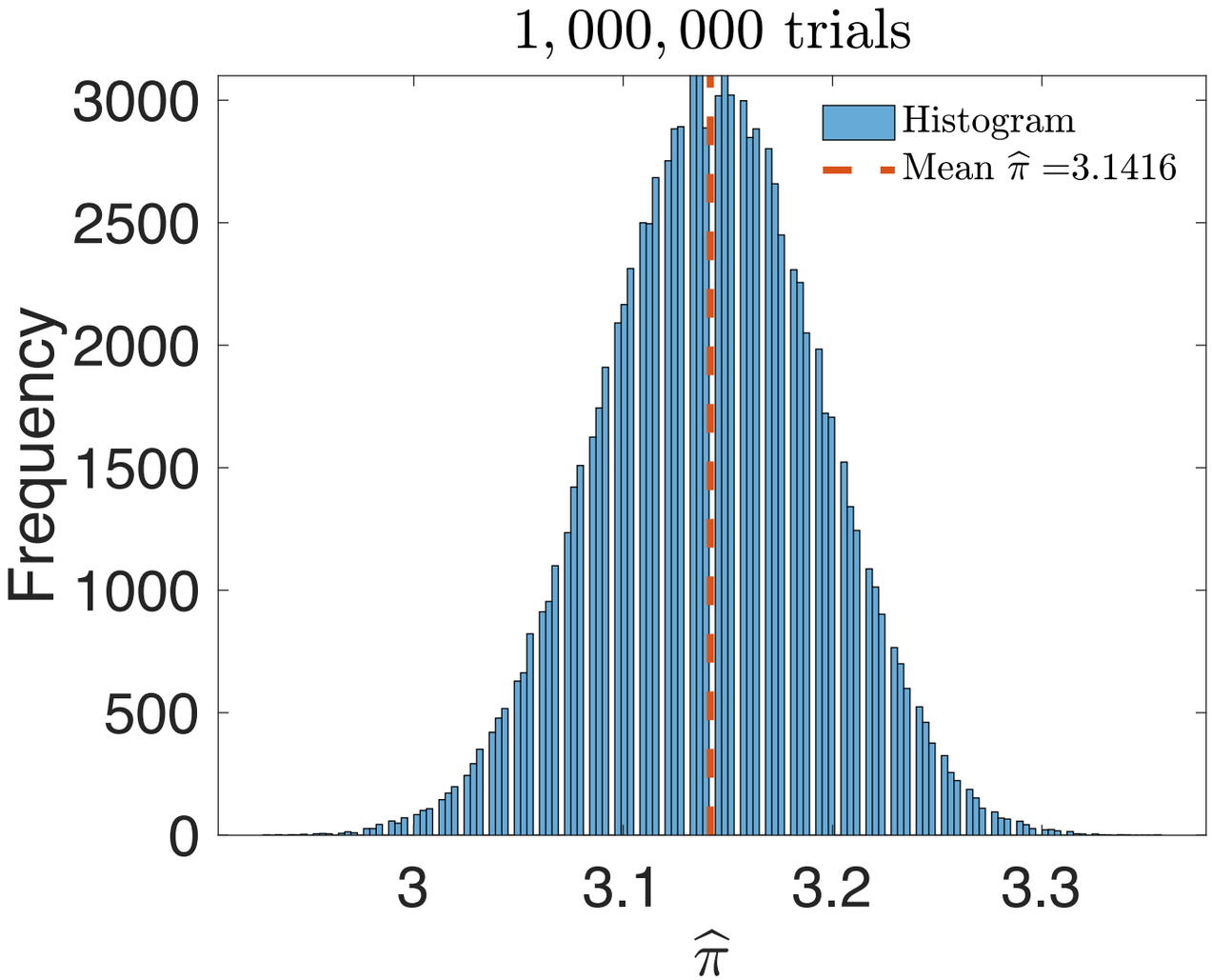
$$\pi = 4S_{\text{circle}}/S_{\text{square}} \approx 4\frac{m}{n}$$

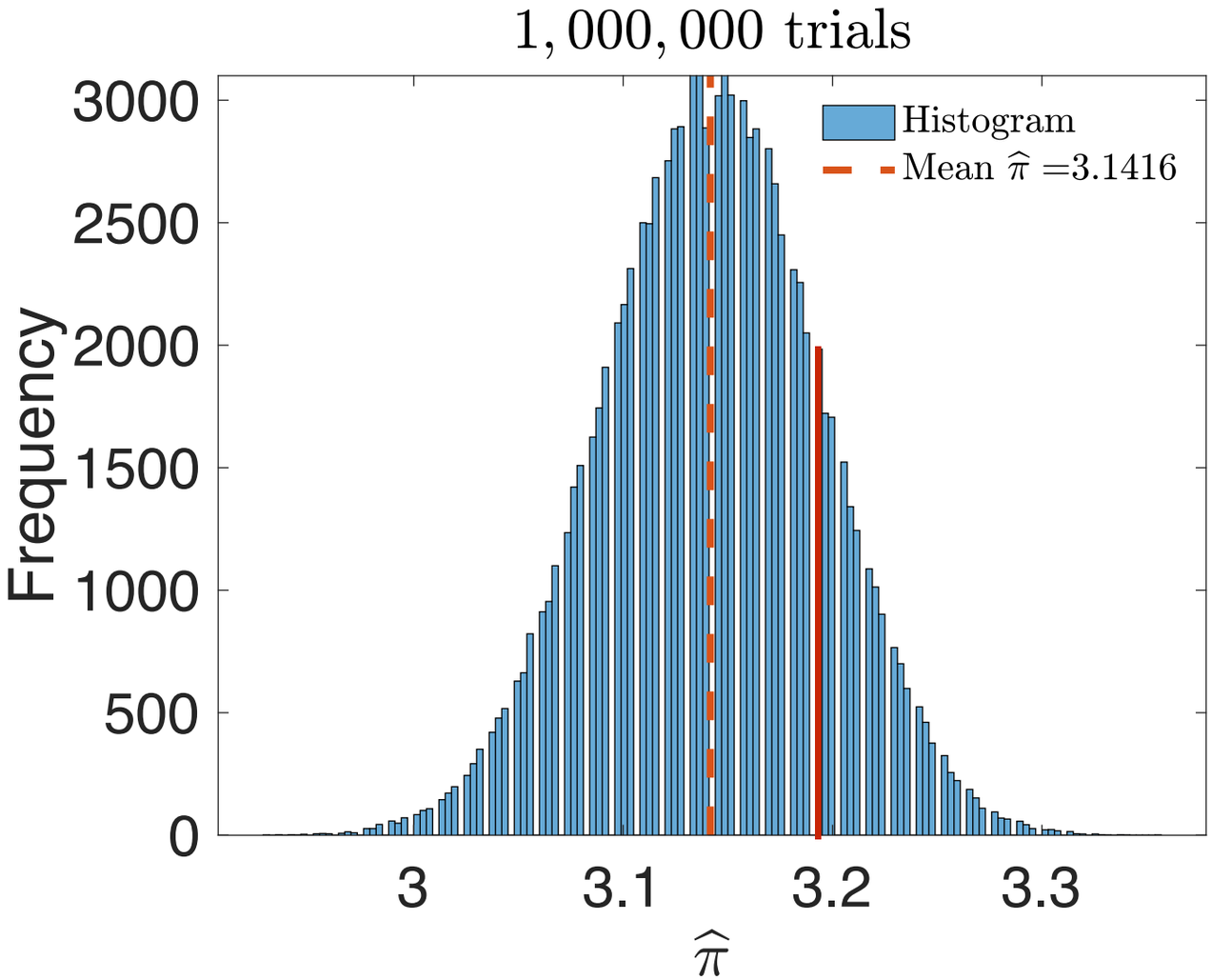
m = 798, n = 1000

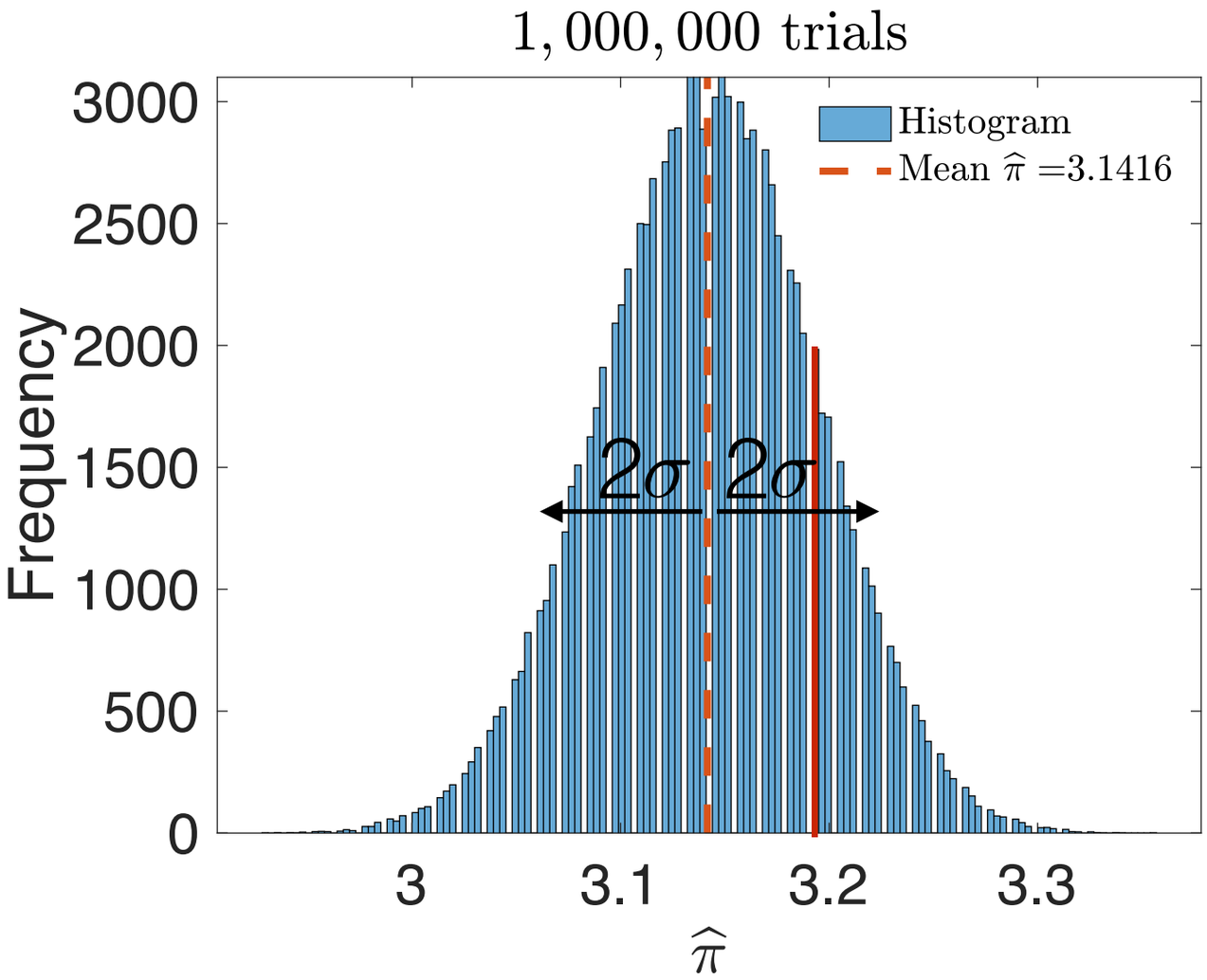
$$\widehat{\pi} = 4\frac{m}{n} = 3.192$$

But how well can we trust the results from experiments with measurement errors?

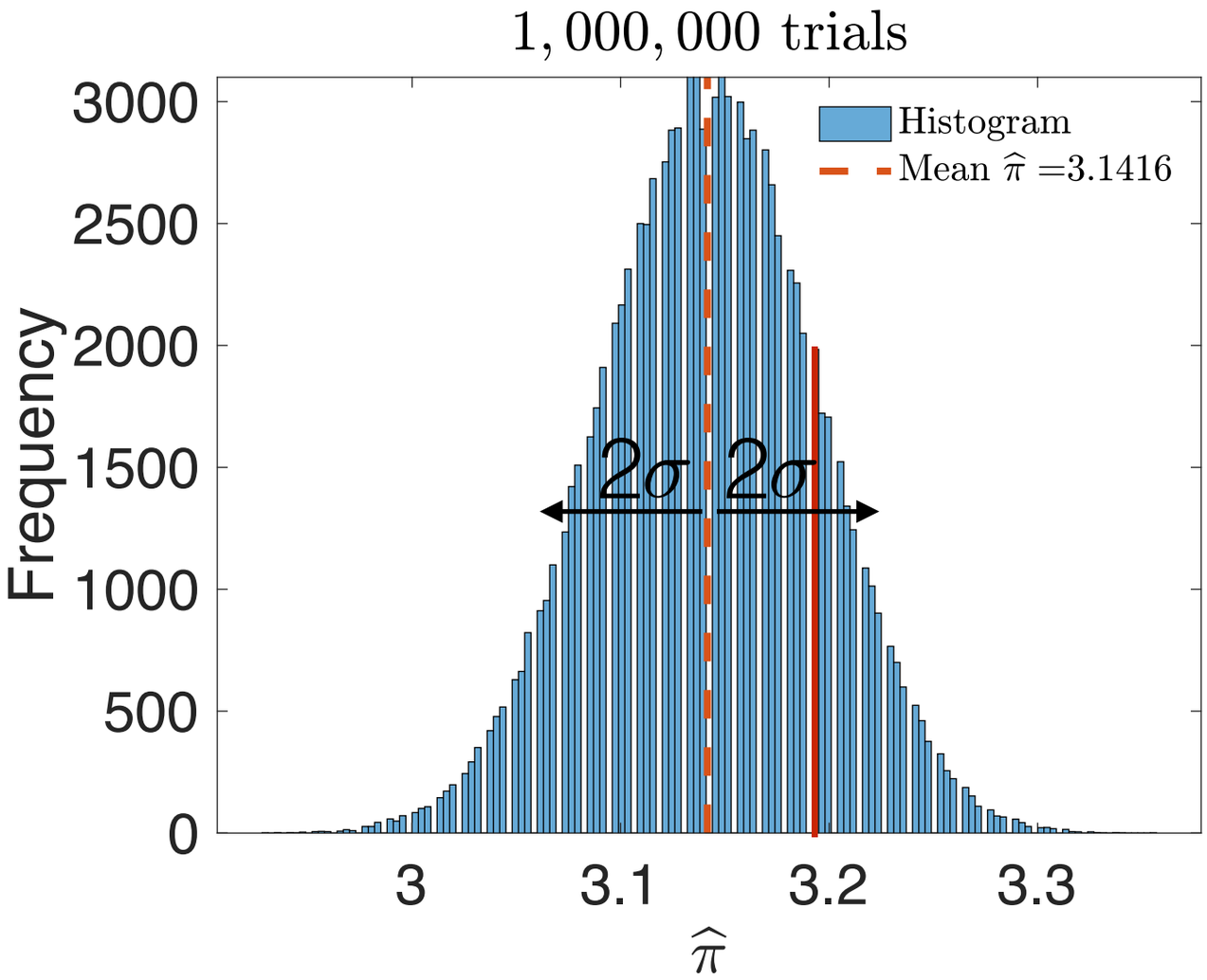
Correct value of π is $3.1415926535\ldots$


1,000,000 trials
Histogram
Mean $\widehat{\pi}$ =3.1416
Frequency
3000
2500
2000
1500
1000
500
0
3
3.1
3.2
3.3
$\widehat{\pi}$

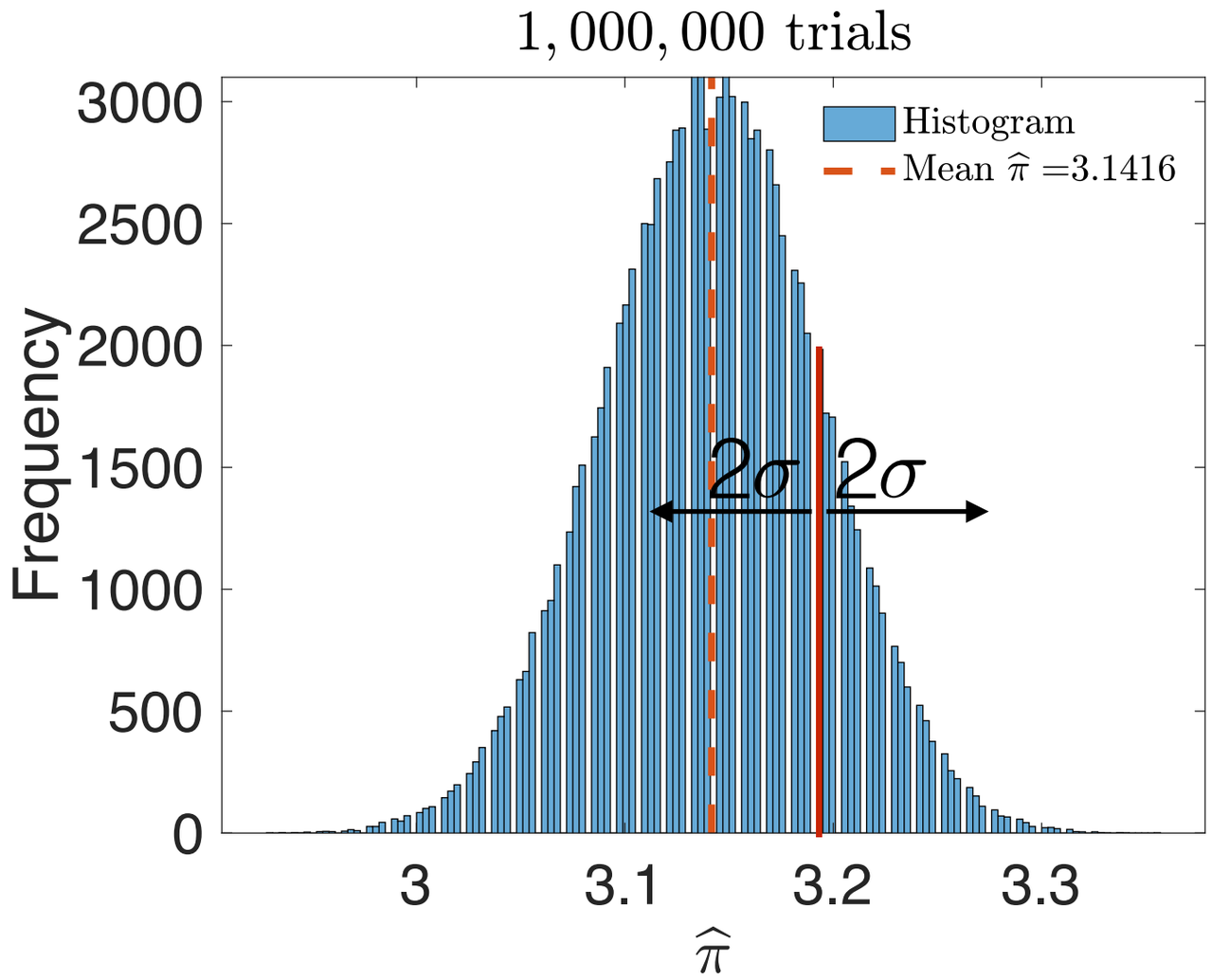

1,000,000 trials
Histogram
Mean $\widehat{\pi}$ =3.1416
Frequency
3000
2500
2000
1500
1000
500
0
3
3.1
3.2
3.3
$\widehat{\pi}$



In 95% of cases, the red value is contained within 2σ of the true value



In 95% of cases, the red value is contained within 2σ of the true value

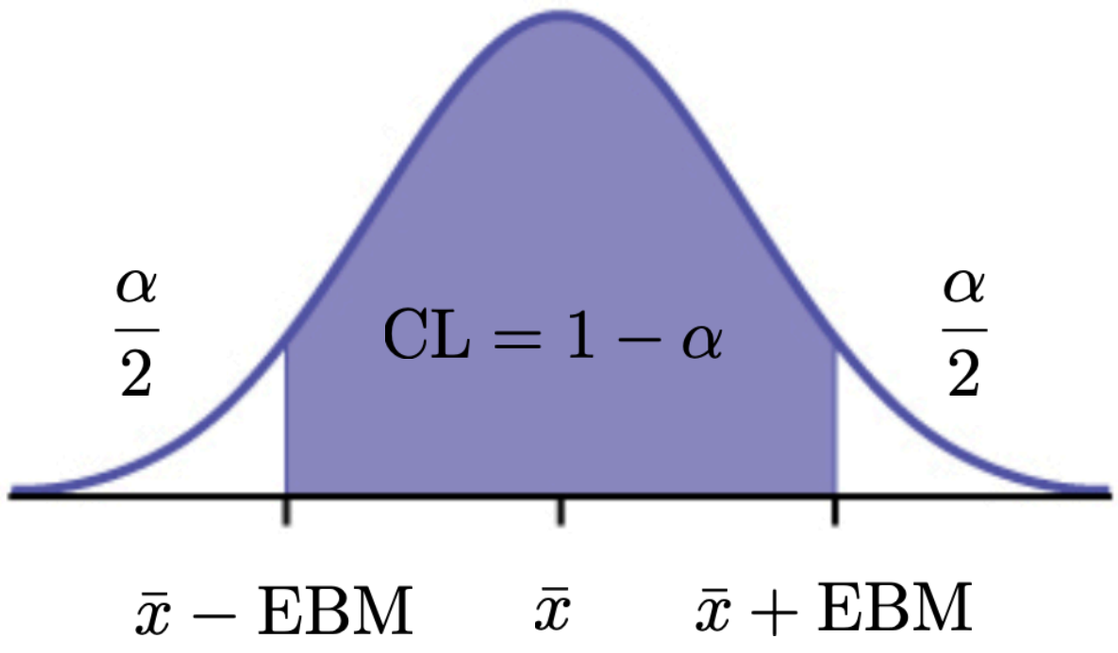


In 95% of cases, the true value is contained within 2σ of the red value

Confidence interval: a range of plausible values for an unknown parameter, which is defined as $(\bar{x} - \mathrm{EBM}, \bar{x} + \mathrm{EBM})$

$\overline{x}$: point estimate

EBM: error bound for the population mean. In other words, the margin of error, which depends on the value of **confidence level** (CL)



Example 1: In a poll published by the National Sleep Foundation, 1,508 randomly selected Americans were surveyed about their sleep. About 60% of them reported that they had some sleep problems. The margin of error is 2.5% at the 95% confidence level.

How do we interpret the results?

95% Confidence interval

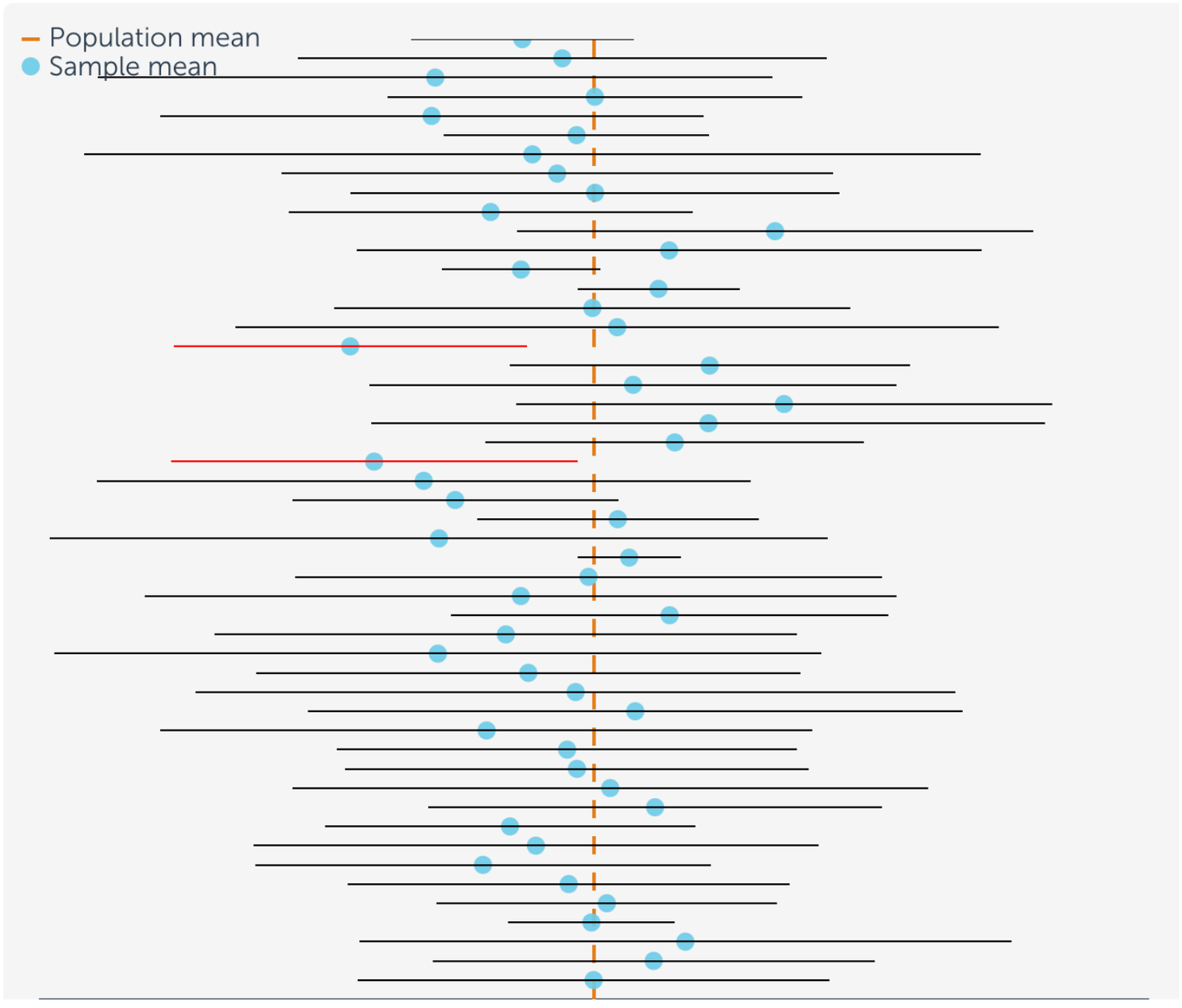
= (60% - 2.5%, 60% + 2.5%)

= (57.5%, 62.5%)

We are 95% confident that between 57.5% and 62.5% of *all* Americans experience some sleep problems.

If we repeat the same survey many times, the fraction of the calculated confidence intervals (different in each time) that encompass the true proportion is close to 95%

95% confidence intervals

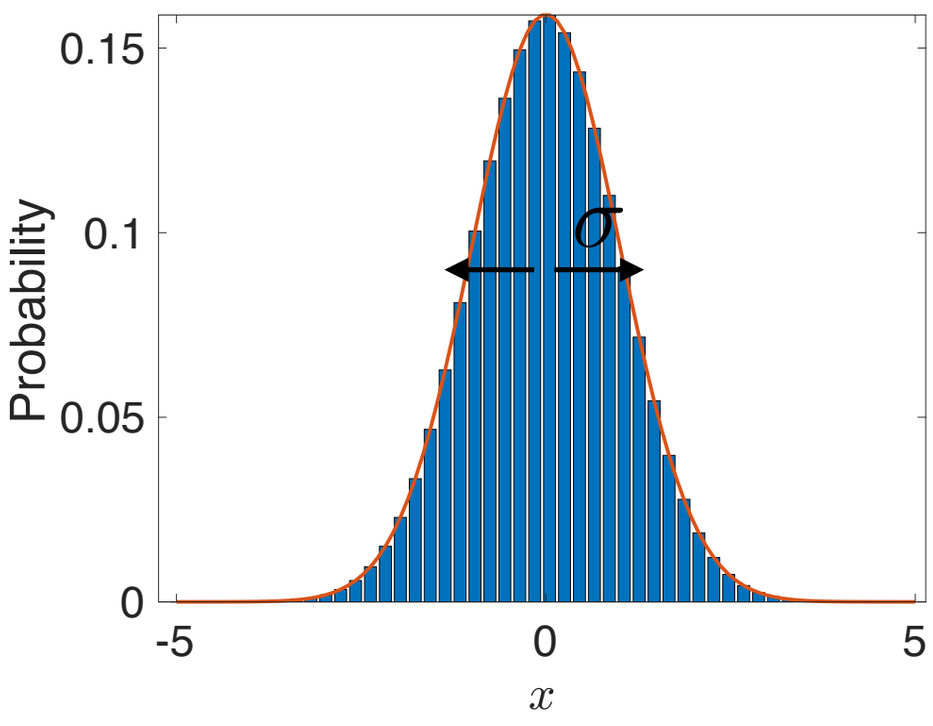


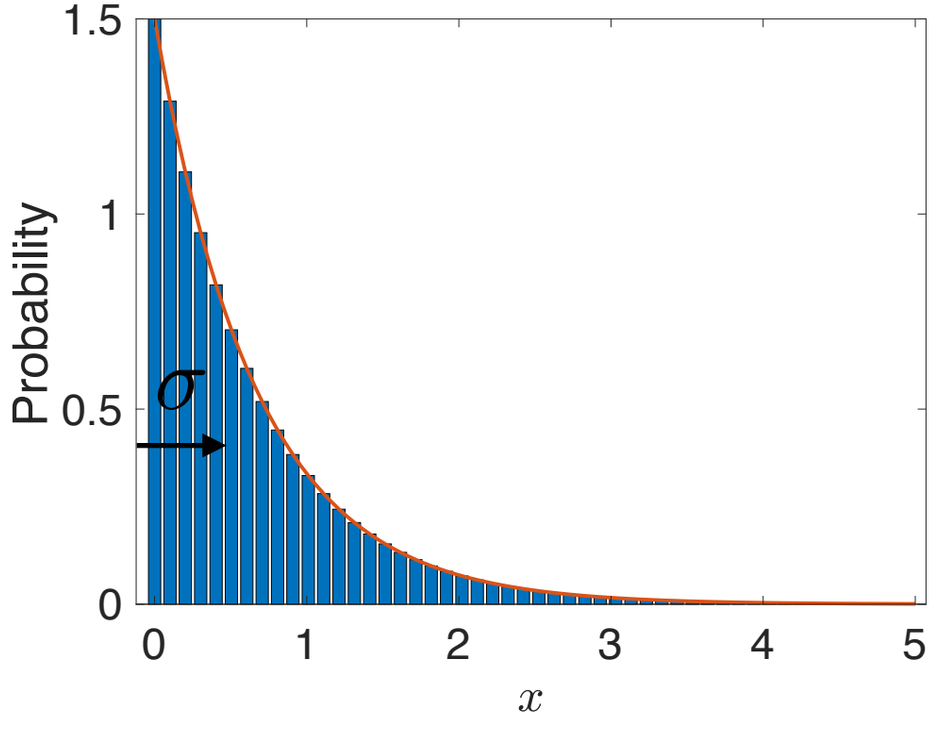
That is, if the true proportion lies outside the 95% confidence interval, then a sampling event has occurred which has a probability of 5% of happening by chance.

How do we compute confidence intervals?

Central limit theorem with more details

Let $\{x_1, x_2, \ldots, x_n\}$ be independent random samples drawn from a distribution of mean μ and standard deviation σ.





Recap of central limit theorem

Let $\{x_1, x_2, \ldots, x_n\}$ be independent random samples drawn from a distribution of mean μ and standard deviation σ. For large enough n (> 30), the distribution of the sample mean

$$\bar{x}_n = (x_1 + x_2 + \cdots + x_n)/n$$

is close to a normal distribution with mean μ and standard deviation $\sqrt{\sigma^2/n}$.

For smaller sample sizes, another distribution called the t-distribution must be used instead of the normal distribution

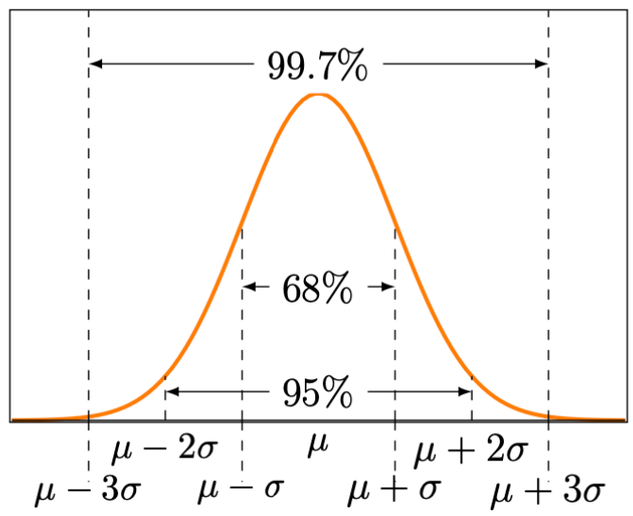
Example 2: In an orchard, there are hundreds of apples on the trees, and we want to measure the mean weight of the apples. You randomly choose just 46 apples and get the mean and the standard deviation of the weights as $\bar{x} = 86\ \mathrm{g},\ \sigma = 6.2\ \mathrm{g}$ What is the 95% confidence interval?

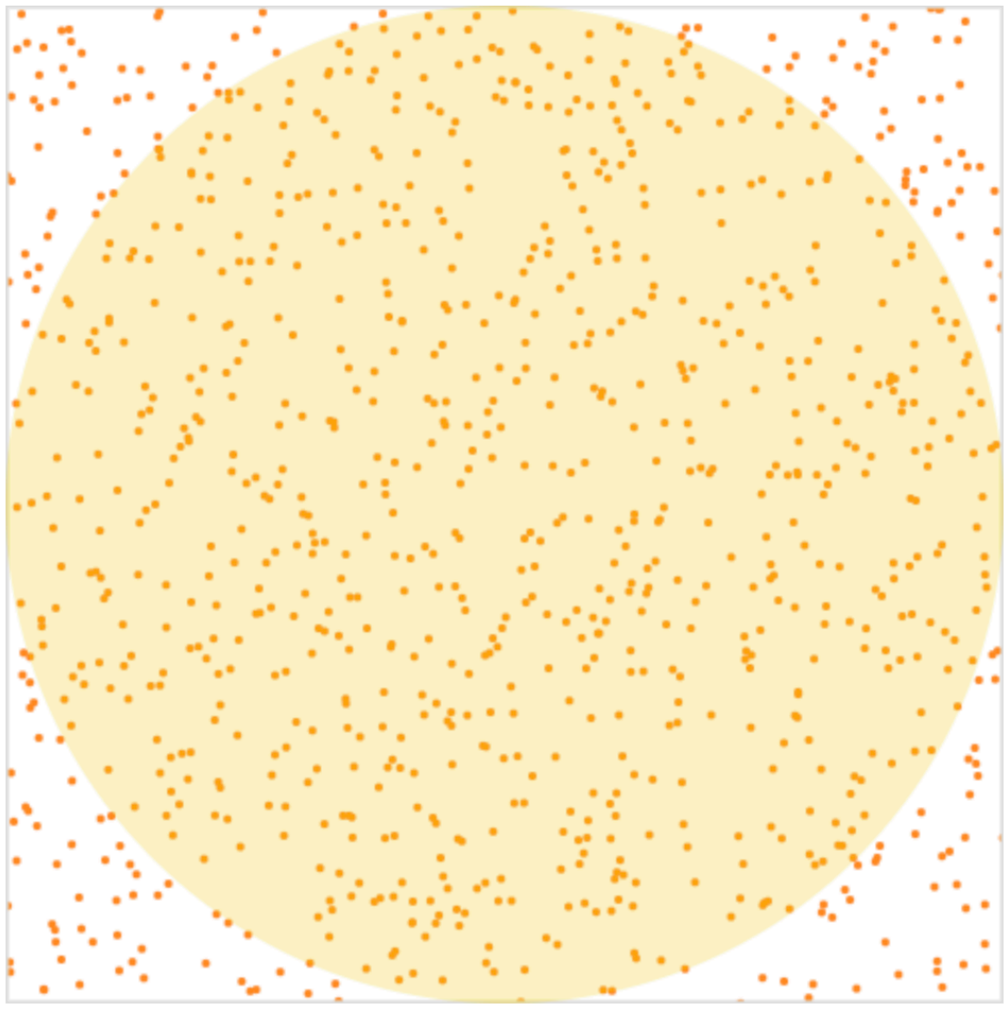
The standard deviation of the distribution of means $s = 6.2/\sqrt{46} = 0.9\ \mathrm{g}$

For 95% confidence level, the margin of error = $2s = 1.8\ \mathrm{g}$

The 95% confidence interval = (84.2, 87.8)

3-sigma rule





m: # of samples in the circle

n: # of samples dropped

m = 798, n = 1000

$$\hat{\pi} = \frac{4m}{n} = 3.192$$

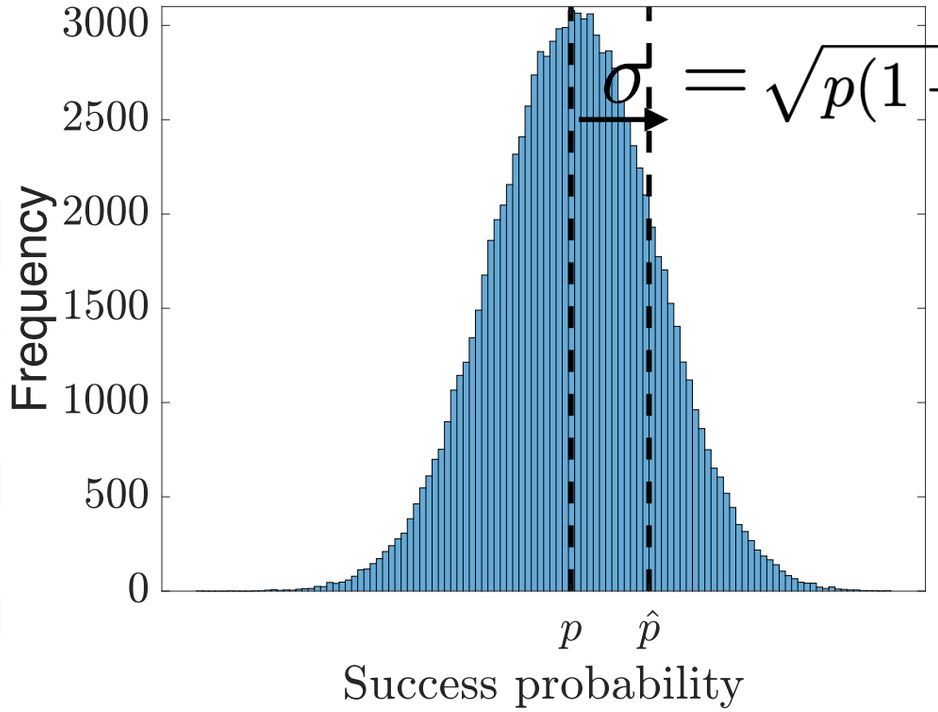
Then, how do we calculate the confidence interval for our estimate of π ?

Binomial proportion confidence interval is an interval estimate of a success probability p when only the number of experiments n and the number of successes n_s are known

Let $\hat{p} = n_s/n$ be the estimate for p

If the sample size is not too small, the distribution of $\hat{p}$ is close to normal, with mean value p and standard deviation $\sqrt{p(1-p)/n}$

We can approximate this by $\sqrt{\hat{p}(1-\hat{p})/n}$ with the hope that p is not close to 0 or 1 and $\hat{p}$ is not too far from p



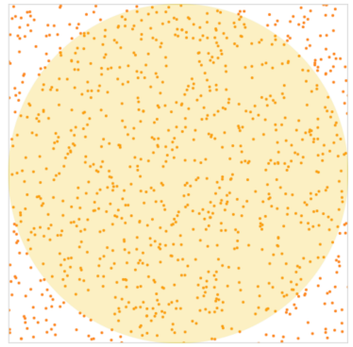
Binomial | **interval** is an interval est obability p when only the nu and the number of success

Let $\hat{p} = n_s/n$ be the estimate for p

If the sample size is not too small, the distribution of $\hat{p}$ is close to normal, with mean value p and standard deviation $\sqrt{p(1-p)/n}$

We can approximate this by $\sqrt{\hat{p}(1-\hat{p})/n}$ with the hope that p is not close to 0 or 1 and $\hat{p}$ is not too far from p

Example 3: Estimation for the value of π

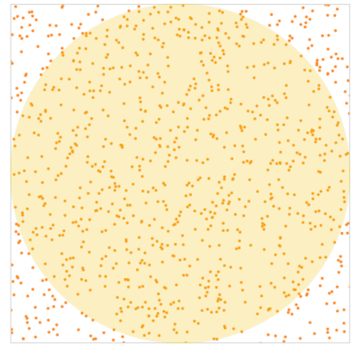


1. Compute the point estimate

$$\hat{p} = \frac{n_s}{n} = 0.798$$

$n_s = 798$, $n = 1000$

Example 3: Estimation for the value of π



$n_s = 798$, $n = 1000$

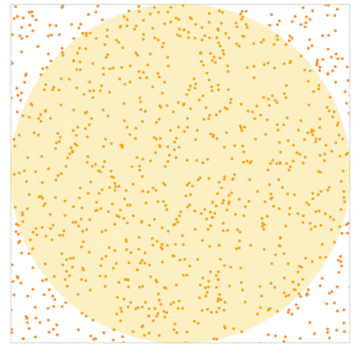
1. Compute the point estimate

$$\hat{p} = \frac{n_s}{n} = 0.798$$

2. Compute the standard deviation

$$s = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{\frac{0.798(1-0.798)}{1000}} = 0.013$$

Example 3: Estimation for the value of π



$n_s = 798$, $n = 1000$

1. Compute the point estimate

$$\hat{p} = \frac{n_s}{n} = 0.798$$

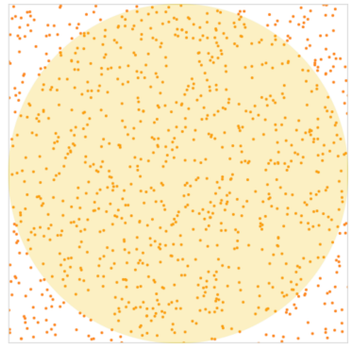
2. Compute the standard deviation

$$s = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{\frac{0.798(1-0.798)}{1000}} = 0.013$$

3.The 95% confidence interval for the success probability is

$$(\hat{p} - 2s, \hat{p} + 2s) = (0.772, 0.824)$$

Example 3: Estimation for the value of π



$n_s = 798, n = 1000$

1. Compute the point estimate

$$\hat{p} = \frac{n_s}{n} = 0.798$$

2. Compute the standard deviation

$$s = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{\frac{0.798(1-0.798)}{1000}} = 0.013$$

3.The 95% confidence interval for the success probability is

$$(\hat{p} - 2s, \hat{p} + 2s) = (0.772, 0.824)$$

Since $\pi = 4p$, the 95% confidence interval for π is (3.088, 3.296) (the conf. interval for p mult. by 4)

Statistical Bias

Common sampling bias

When a sample does not have the same characteristics as the population, we say this is a biased sample.

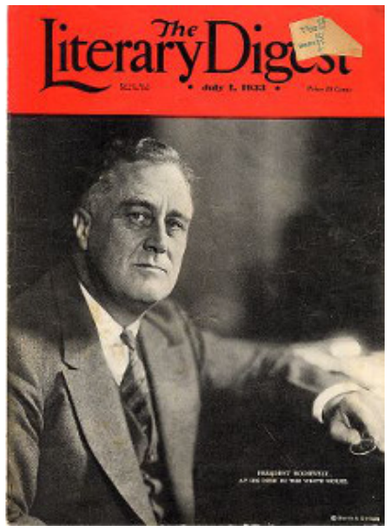
Common sampling bias

When a sample does not have the same characteristics as the population, we say this is a biased sample.

Example: poll by the magazine "The Literary Digest" for the 1936 presidential election.

Prediction: Alf Landon beats Franklin D Roosevelt by 57% to 43%.

Result: Roosevelt beats Landon by 62% to 38%



What went wrong in the 1936 poll?

Prediction: Alf Landon beats Franklin D Roosevelt by 57% to 43%.

Result: Roosevelt beats Landon by 62% to 38%

Questionnaires were sent out using lists of phone numbers, drivers' registrations, and country club memberships. 24% of those polled responded.

What went wrong?

What went wrong in the 1936 poll?

Questionnaires were sent out using lists of phone numbers, drivers' registrations, and country club memberships. 24% of those polled responded.

- The people polled were wealthy and the election was at the height of the Great Depression
- A large fraction of polled did not respond. Typically, those with strong feelings respond
- …

What went wrong in the 1936 poll?

Questionnaires were sent out using lists of phone numbers, drivers' registrations, and country club memberships. 24% of those polled responded.

- The people polled were wealthy and the election was at the height of the Great Depression
- A large fraction of polled did not respond. Typically, those with strong feelings respond
- …

Note that the poll was enormous (10 million questionnaires!), but bias still made the result useless!

Common sampling bias

When a sample does not have the same characteristics as the population, we say this is a biased sample.

We will talk about four kinds of bias:

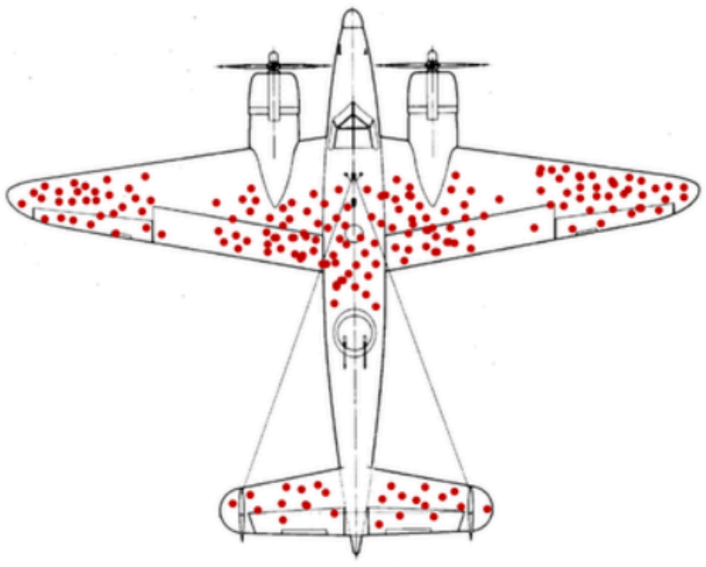
1. Survival bias
2. Self-selection bias
3. Confirmation bias
4. Undercoverage bias

Common sampling bias

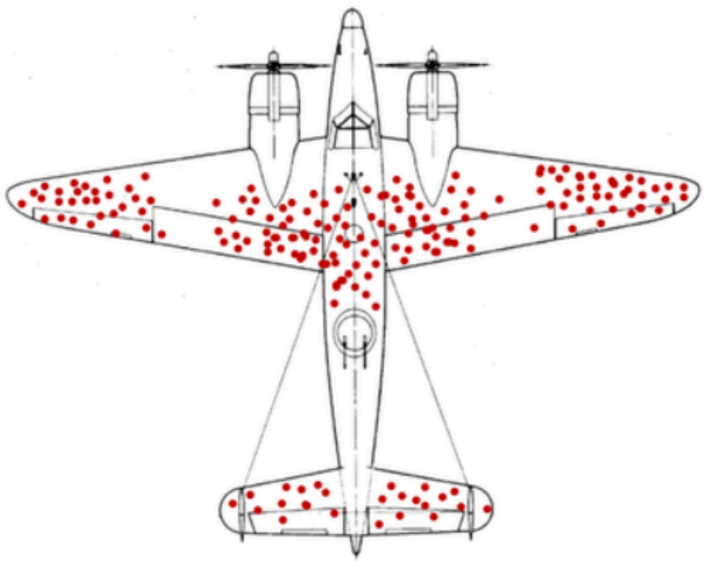
When a sample does not have the same characteristics as the population, we say this is a biased sample.

Survival bias: only the portion of the population that has survived some process can be sampled

Example 1: In scientific journals, there is strong publication bias towards positive results. Successful research outcomes are published far more often than null findings.



Example 2: During World War II, the statistician Abraham Wald examined the damage done to aircrafts that had returned from missions. The US military previously concluded that the most-hit areas of the plane needed additional armor.



Example 2: During World War II, the statistician Abraham Wald examined the damage done to aircrafts that had returned from missions. The US military previously concluded that the most-hit areas of the plane needed additional armor. Wald instead recommended adding armor to the areas that showed the least damage.

Self-selection bias: People with specific characteristics are more likely to agree to take part in a study than others. This often leads to a polarization of responses with extreme perspectives being given a disproportionate weight in the summary.

Example: people who have strong opinions or substantial knowledge may be more willing to spend time answering a survey than those who do not.

Confirmation bias: People display this bias when they select information that supports their views.

Example 1: In social media, personalized search displays to individuals only information they are likely to agree with, while excluding opposing views.

Example 2: The decision made by a doctor may be strongly influenced by the disorders described in a recently-read paper, without considering multiple possibilities based on evidence.

Undercoverage bias: Some members of a population are inadequately represented in the sample.

Example 1: Administering general national surveys online may miss groups with limited internet access, such as the elderly and lower-income households.

Example 2: Researchers want to know what citizens in a particular city think of a new traffic law so they give out a questionnaire to people that walk by at a local mall.