MS&E 226: Fundamentals of Data Science

Lecture 9: Frequentist properties of estimators

Ramesh Johari
ramesh.johari@stanford.edu

Frequentist inference

Thinking like a frequentist

Suppose that for some population distribution with parameters $\boldsymbol{\theta}$, you have a process that takes observations $\mathbf{Y}$ and constructs an estimator $\hat{\boldsymbol{\theta}}$.

How can we quantify our uncertainty in $\hat{\boldsymbol{\theta}}$?

By this we mean: *how sure are we of our guess?*

We take a *frequentist* approach: How well would our procedure would do if it was repeated many times?

Thinking like a frequentist

The rules of thinking like a frequentist:

- The parameters $\boldsymbol{\theta}$ are fixed (non-random).
- Given the fixed parameters, there are many possible realizations ("parallel universes") of the data given the parameters.
- We get one of these realizations, and use only the universe we live in to reason about the parameters.

The sampling distribution

The distribution of $\hat{\boldsymbol{\theta}}$ if we carry out this "parallel universe" simulation is called the *sampling distribution*.

The sampling distribution is the heart of frequentist inference! Nearly all frequentist quantities we derive come from the sampling distribution.

The idea is that the sampling distribution quantifies our uncertainty, since it captures how much we expect the estimator to vary if we were to repeat the procedure over and over ("parallel universes").

Example: Flipping a biased coin

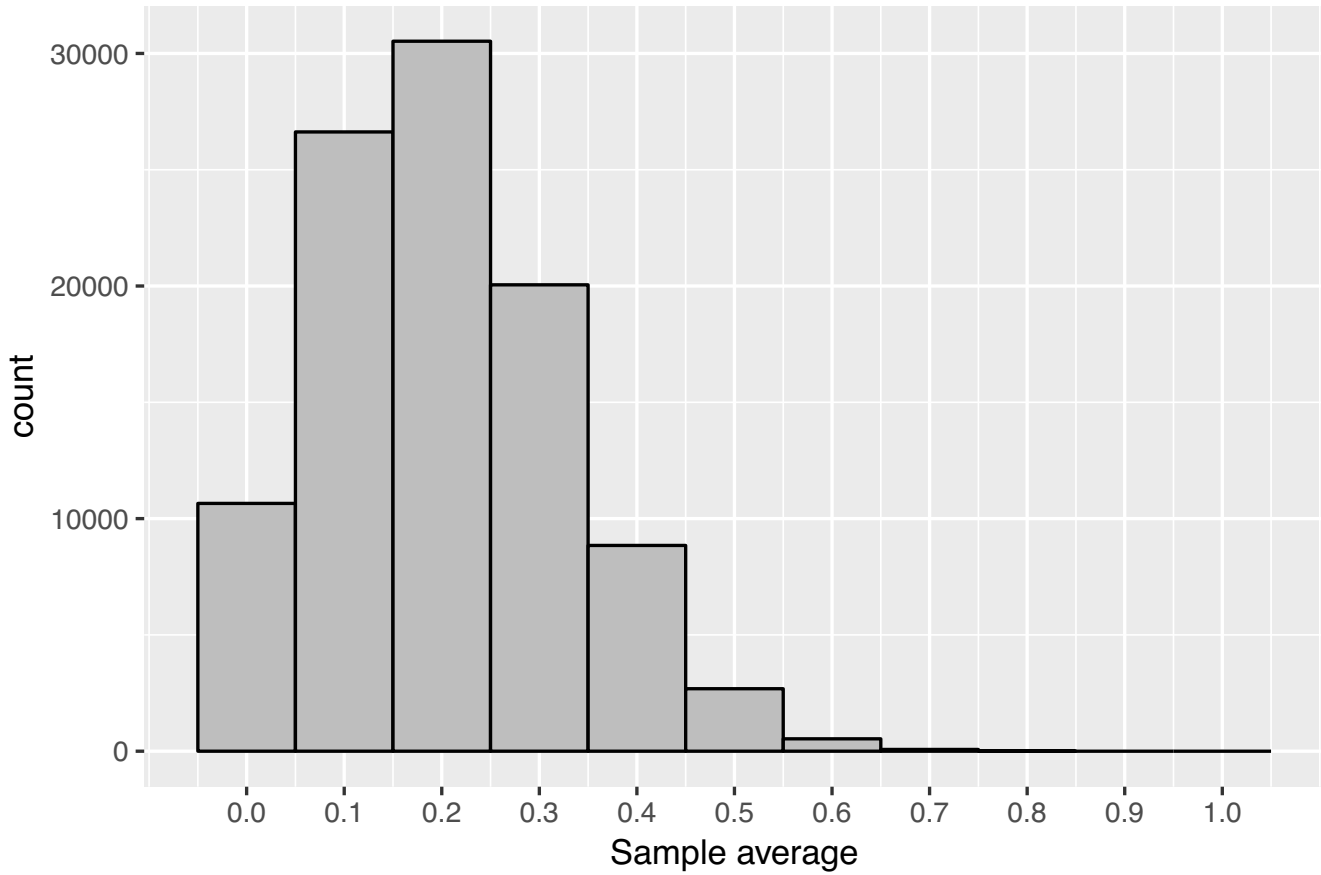
Suppose we are given the outcome of $n = 10$ flips of a biased coin, and take the *sample average* of these ten flips as our estimate of the true bias q (i.e., compute the maximum likelihood estimator of the bias).

What is the sampling distribution of this estimator?

Example: Flipping a biased coin

Suppose true bias is $q = 0.20$.

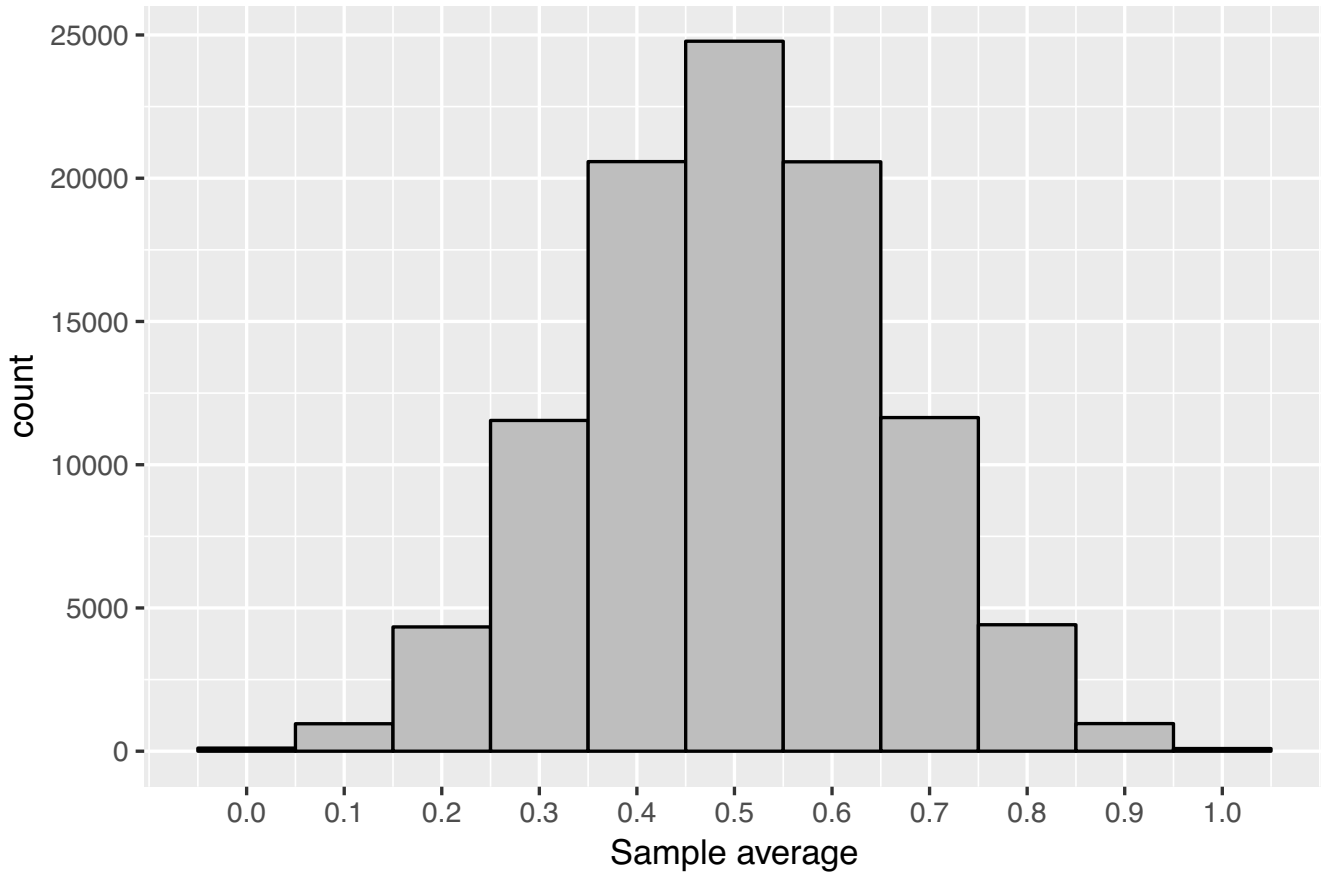
100,000 parallel universes, with $n = 10$ flips in each:



Example: Flipping a biased coin

Suppose true bias is $q = 0.5$.

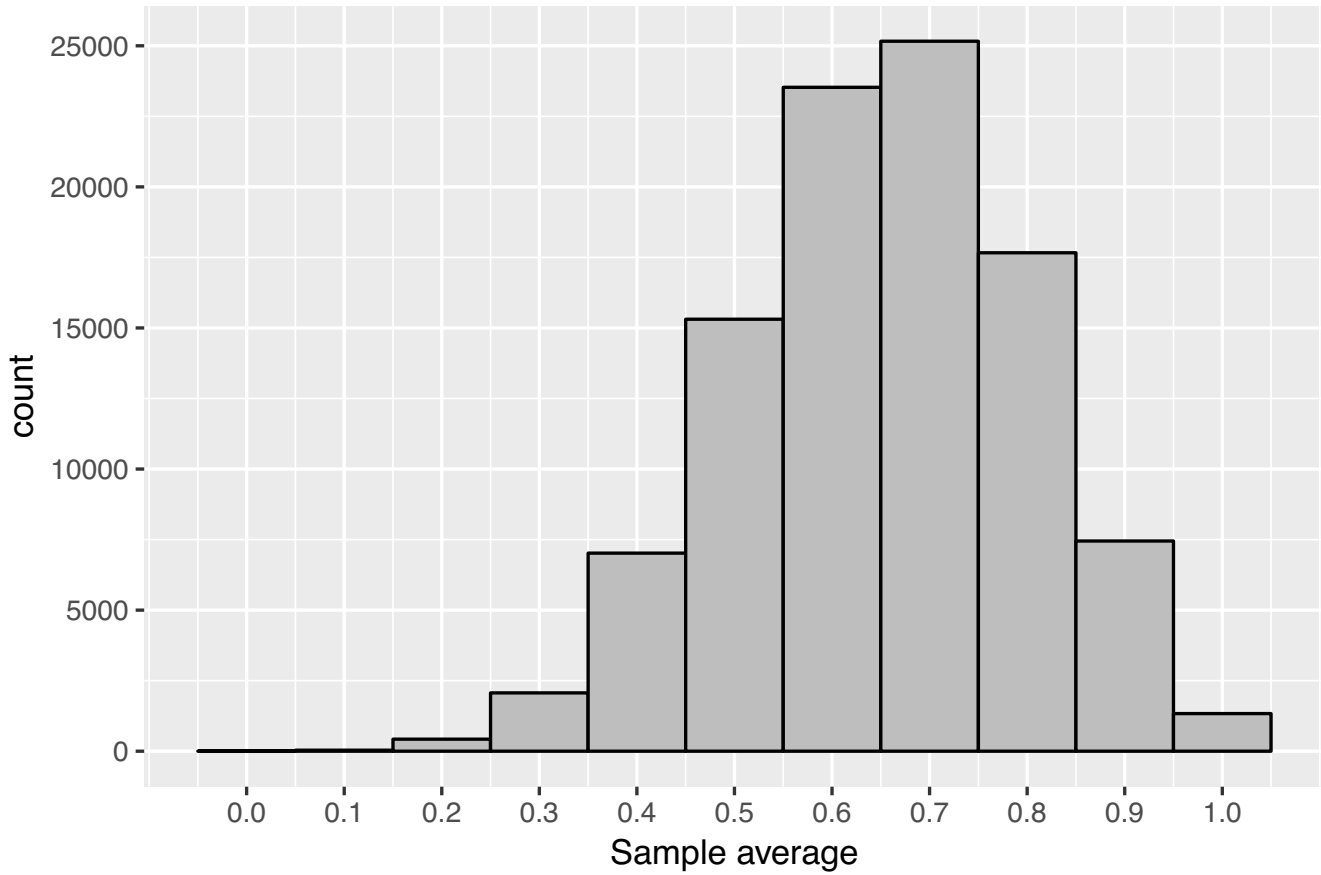
100,000 parallel universes, with $n = 10$ flips in each:



Example: Flipping a biased coin

Suppose true bias is $q = 0.65$.

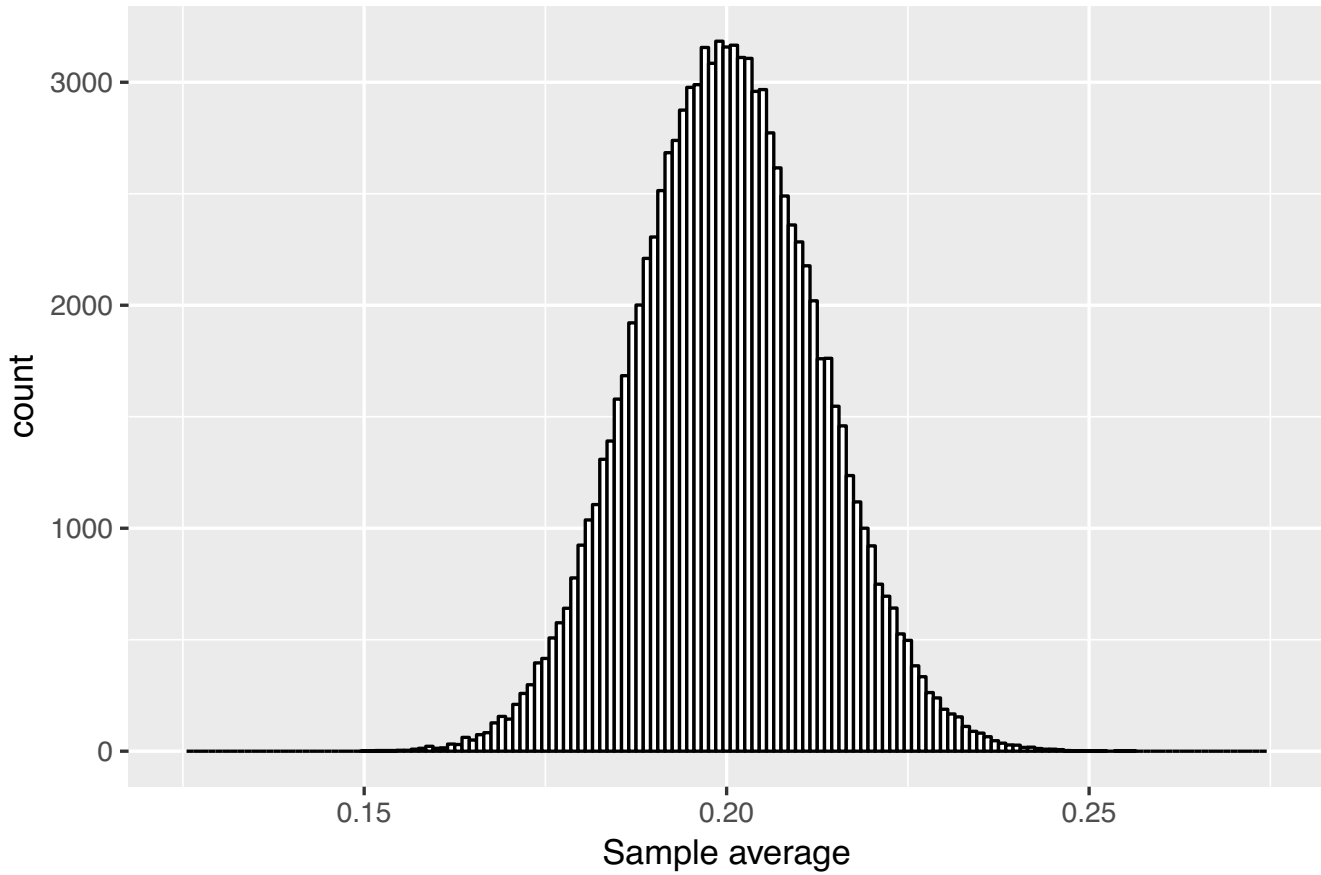
100,000 parallel universes, with $n = 10$ flips in each:



Example: Flipping a biased coin

Suppose true bias is $q = 0.20$.

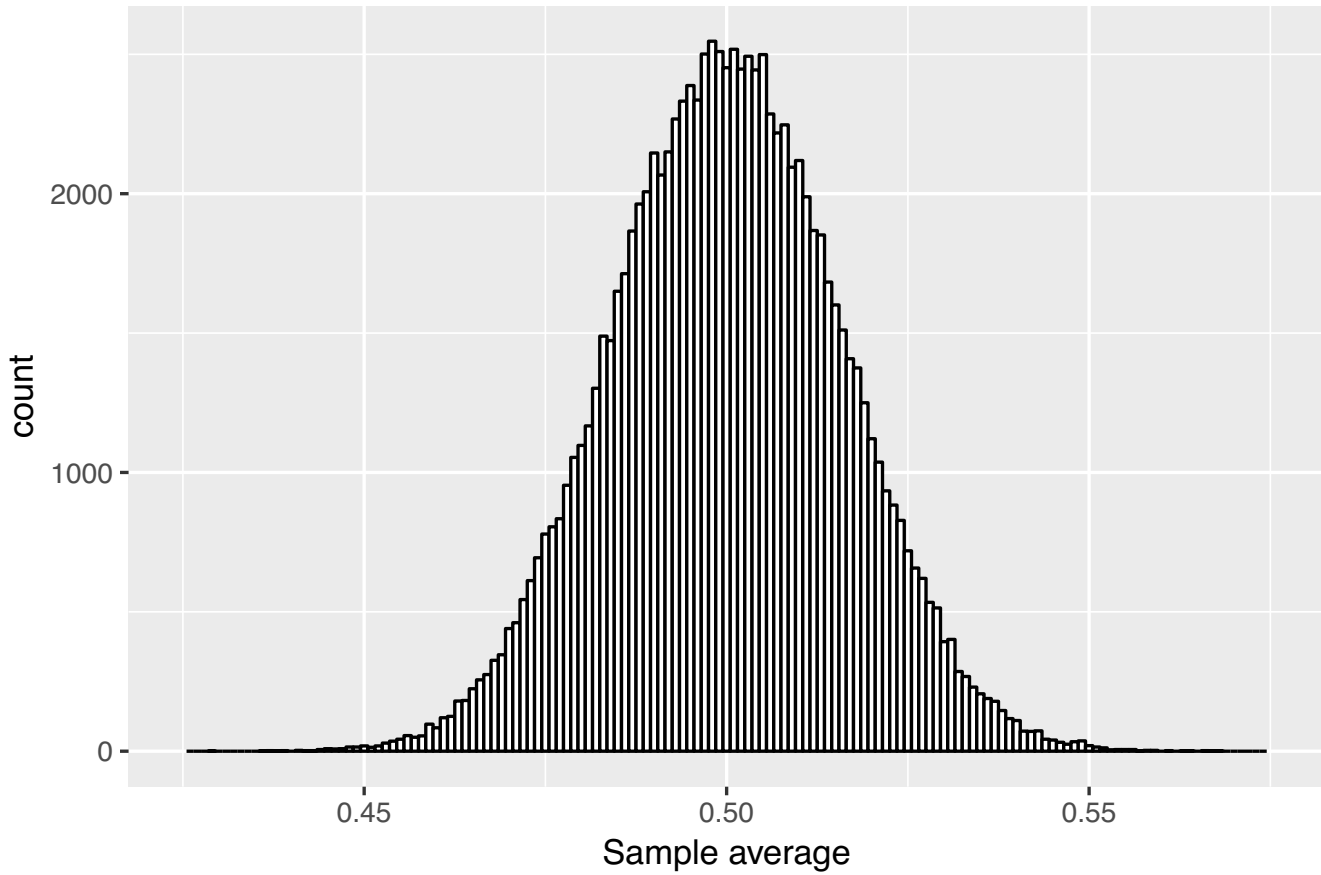
100,000 parallel universes, with $n = 1000$ flips in each:



Example: Flipping a biased coin

Suppose true bias is $q = 0.5$.

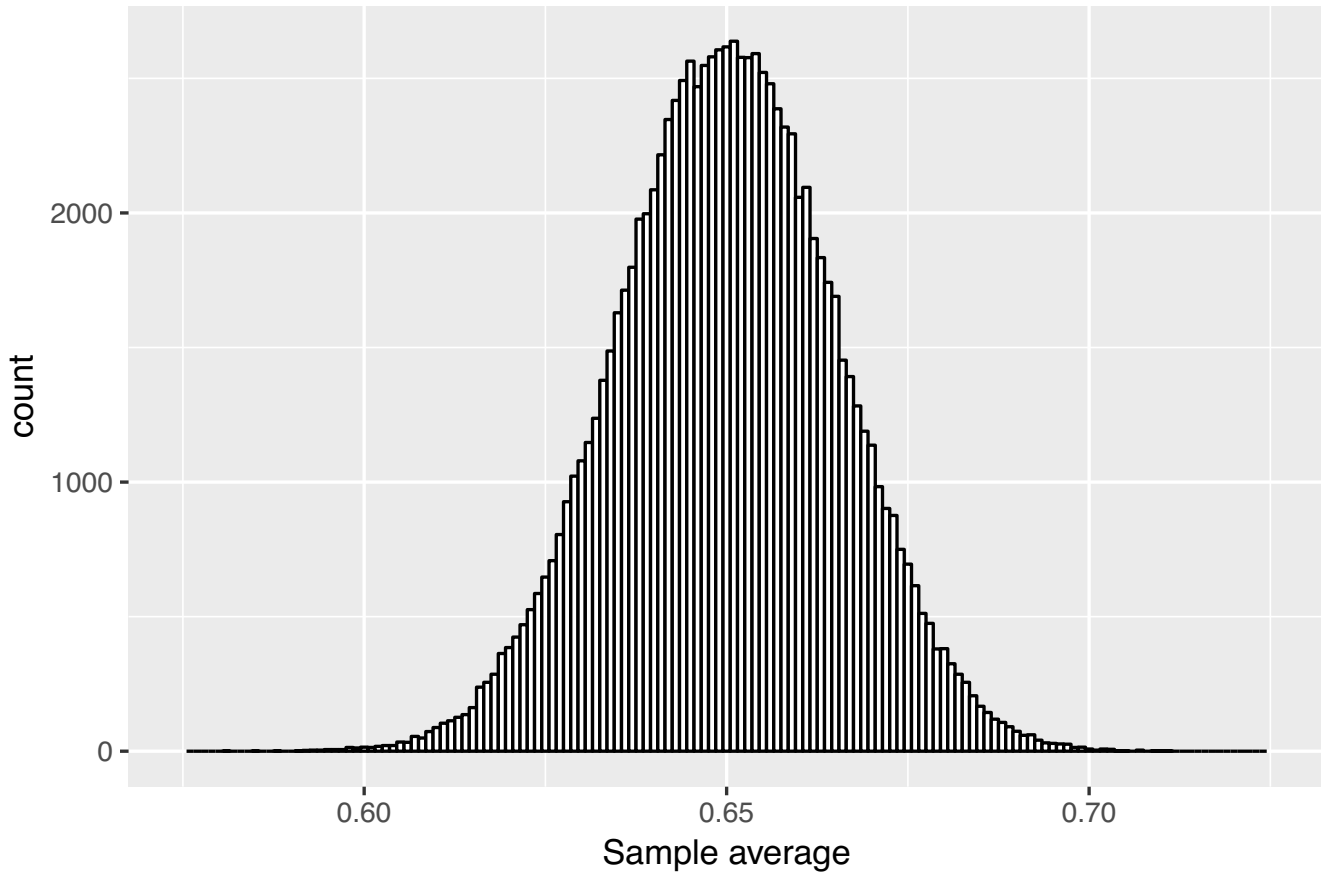
100,000 parallel universes, with $n = 1000$ flips in each:



Example: Flipping a biased coin

Suppose true bias is $q = 0.65$.

100,000 parallel universes, with $n = 1000$ flips in each:



Example: Flipping a biased coin

The sampling distribution is what we get if *the number of parallel universes* $\to \infty$.

In this case: the sampling distribution is $\frac{1}{n}\text{Binomial}(n, q)$.

MLE = # of successes / # of flips

Note that:

$\frac{nq}{n} = q$

- In this case, the *mean* of the sampling distribution *is* the true parameter. This need not always be true.
- The standard deviation of the sampling distribution is called the *standard error* (SE) of the estimator.[1]
- The sampling distribution looks *asymptotically normal* as $n \to \infty$.

[1] We also use the term "standard error" to refer to an estimate of the true SE, computed from a specific sample. In this case we write $\widehat{\text{SE}}$ to clarify the dependence on the observed data.

Using the sampling distribution

How might you use the sampling distribution?

Suppose you receive 10 flips of a coin, of which 2 are heads. You compute the MLE, $\hat{q}_{\text{MLE}} = 0.2$.

To quantify your uncertainty in this guess, you can use the sampling distribution in two ways:

- First, you might simulate the sampling distribution with $n = 10$, *assuming the true bias* q *was actually equal to your estimate of* 0.2. You can use, e.g., the standard deviation of this simulated distribution to measure your uncertainty; this is an example of an *estimated standard error* ($\widehat{SE}$).

Using the sampling distribution

How might you use the sampling distribution?

Suppose you receive 10 flips of a coin, of which 2 are heads. You compute the MLE, $\hat{q}_{\text{MLE}} = 0.2$.

To quantify your uncertainty in this guess, you can use the sampling distribution in two ways:

- First, you might simulate the sampling distribution with $n = 10$, *assuming the true bias* q *was actually equal to your estimate of* 0.2. You can use, e.g., the standard deviation of this simulated distribution to measure your uncertainty; this is an example of an *estimated standard error* ($\widehat{SE}$).
- Second, you might simulate the sampling distribution with $n = 10$, *assuming a fixed value for the true bias, e.g.,* $q = 0.5$. With this simulation you can answer the following question: *if* the true bias were q, what is the chance that you would have observed the MLE that you did?

Using the sampling distribution

How might you use the sampling distribution?

Suppose you receive 10 flips of a coin, of which 2 are heads. You compute the MLE, $\hat{q}_{\text{MLE}} = 0.2$.

To quantify your uncertainty in this guess, you can use the sampling distribution in two ways:

- First, you might simulate the sampling distribution with $n = 10$, *assuming the true bias* q *was actually equal to your estimate of* 0.2. You can use, e.g., the standard deviation of this simulated distribution to measure your uncertainty; this is an example of an *estimated standard error* ($\widehat{SE}$).
- Second, you might simulate the sampling distribution with $n = 10$, *assuming a fixed value for the true bias, e.g.,* $q = 0.5$. With this simulation you can answer the following question: *if* the true bias were q, what is the chance that you would have observed the MLE that you did?

We discuss both approaches to quantifying uncertainty in the subsequent slides.

Analyzing the MLE

Bias and consistency of the MLE

The MLE may not be unbiased with finite n. However, under reasonable technical conditions, the MLE is converges to the true parameter when the amount of data grows. This is called *consistency*.

Theorem

If appropriate technical conditions hold, the MLE is consistent*:*

If the true parameter is $\boldsymbol{\theta}$*, then as* $n \to \infty$*, there holds* $\hat{\boldsymbol{\theta}}_{\mathsf{MLE}} \to \boldsymbol{\theta}$ *in probability.*

(See Theorem 9.13 in [AoS]for a precise technical statement.)

Asymptotic normality of the MLE

Theorem

If appropriate technical conditions hold, then for large n, the sampling distribution of the MLE $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is approximately a normal distribution (with mean that is the true parameter $\boldsymbol{\theta}$).

The variance, and thus the standard error, can be explicitly characterized in terms of the *Fisher information* of the population model. See Theorem 9.18 in [AoS]for a precise technical statement.

Example 1: Biased coin flipping

Let $\hat{q}_{\text{MLE}}$ be MLE estimate of q, the bias of the coin; recall it is just the sample average:

$$\hat{q}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}.$$

- $\hat{q}_{\text{MLE}}$ is unbiased, regardless of n.
- The standard error of $\hat{q}_{\text{MLE}}$ can be computed directly:

$$\text{SE} = \sqrt{\text{Var}(\hat{q}_{\text{MLE}})} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} q(1-q)} = \sqrt{\frac{q(1-q)}{n}}.$$

- We can estimate the standard error of $\hat{q}$ by $\widehat{\text{SE}} = \sqrt{\hat{q}_{\text{MLE}}(1-\hat{q}_{\text{MLE}})/n}$.
- For large n, $\hat{q}_{\text{MLE}}$ is approximately normal, with mean q and variance $\widehat{\text{SE}}^2$.

Example 2: Linear normal model

$Y_i = \beta_0 + \sum_{j=1}^{n} \beta_j X_{ij} + \varepsilon_i \quad (i.i.d) \quad \varepsilon_i \sim N(0, \sigma^2)$

Recall that in this case, $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ is the OLS solution. (for β)

- It is unbiased (cf. the Gauss-Markov theorem), regardless of n.
- The covariance matrix of $\hat{\boldsymbol{\beta}}$ can be shown to be given by $\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$. (Using similar analysis to the bias-variance decomposition for linear regression.)
  In particular, the standard error SE_j of $\hat{\beta}_j$ is the square root of the j'th diagonal entry of this matrix.
- To estimate this covariance matrix (and hence SE_j), we use an estimator $\hat{\sigma}^2$ for σ^2.
- For large n, $\hat{\boldsymbol{\beta}}$ is approximately normal.

Example 2: Linear normal model

What estimator $\hat{\sigma}^2$ to use?

Recall that $\hat{\sigma}^2_{\text{MLE}}$ is the following sample variance:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} r_i^2.$$

But this is *not unbiased!* In fact it can be shown that:

$$\mathbb{E}_{\mathbf{Y}}[\hat{\sigma}^2_{\text{MLE}}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}] = \frac{n-p}{n}\sigma^2.$$

Example 2: Linear normal model

In other words, $\hat{\sigma}^2_{\text{MLE}}$ *underestimates* the true error variance.

This is because the MLE solution $\hat{\boldsymbol{\beta}}$ was *chosen* to minimize squared error on the training data. We need to account for this "favorable selection" of the variance estimate by "reinflating" it. So an unbiased estimate of σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2.$$

The quantity $n - p$ is called the *degrees of freedom* (DoF).

Example 2: Linear normal model

R output after running a linear regression:

```
lm(formula = Ozone ~ 1 +
   Solar.R + Wind + Temp,
   data = airquality)
             coef.est coef.se
(Intercept) -64.34    23.05
Solar.R       0.06     0.02
Wind         -3.33     0.65
Temp          1.65     0.25
---
n = 111, k = 4
residual sd = 21.18, R-Squared = 0.61
```

4×4 matrix $\rightarrow \hat{\sigma}^2 (X^T X)^{-1}$

$(p = 3)$

Example 2: Linear normal model

In the case of simple linear regression (one covariate with intercept), we can explicitly calculate the standard error of $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\widehat{\mathsf{SE}}_0 = \frac{\hat{\sigma}}{\hat{\sigma}_X \sqrt{n}} \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}};$$
$$\widehat{\mathsf{SE}}_1 = \frac{\hat{\sigma}}{\hat{\sigma}_X \sqrt{n}},$$

where $\hat{\sigma}$ is the estimate of standard deviation of the error.

Note that both standard errors decrease proportional to $1/\sqrt{n}$. This is always true for standard errors in linear regression.

Standard error vs. estimated standard error

Note that the standard error *depends* on the unknown parameter(s)!

- *Biased coin flipping*: SE depends on q.
- *Linear normal model*: SE depends on σ^2.

This makes sense: the sampling distribution is determined by the unknown parameter(s), and so therefore the standard deviation of this distribution must also depend on the unknown parameter(s).

In each case we *estimate* the standard error, by plugging in an estimate for the unknown parameter(s).

Additional properties of the MLE

Two additional useful properties of the MLE:

- It is *equivariant*: If you want the MLE of a function of $\boldsymbol{\theta}$, say $g(\boldsymbol{\theta})$, you can get it by just computing $g(\hat{\boldsymbol{\theta}}_{\text{MLE}})$. (See Theorem 9.14 in [AoS].)
- It is *asymptotically optimal*: As the amount of data increases, the MLE has asymptotically lower variance than any other asymptotically normal consistent estimator of $\boldsymbol{\theta}$ you can construct (in a sense that can be made precise). (See Theorem 9.23 in [AoS].)

Additional properties of the MLE

Asymptotic optimality is also referred to as *asymptotic efficiency* of the MLE.

The idea is that the MLE is the most "informative" estimate of the true parameter(s), given the data.

However, it's important to also understand when MLE estimates might *not* be efficient.

A clue to this is provided by the fact that the guarantees for MLE performance are all *asymptotic* as n grows large (consistency, normality, efficiency).

In general, when n is not "large" (and in particular, when the number of covariates is large relative to n), the MLE may not be efficient.

Confidence intervals

Quantifying uncertainty

The combination of the standard error, consistency, and asymptotic normality allow us to quantify uncertainty directly through *confidence intervals*:

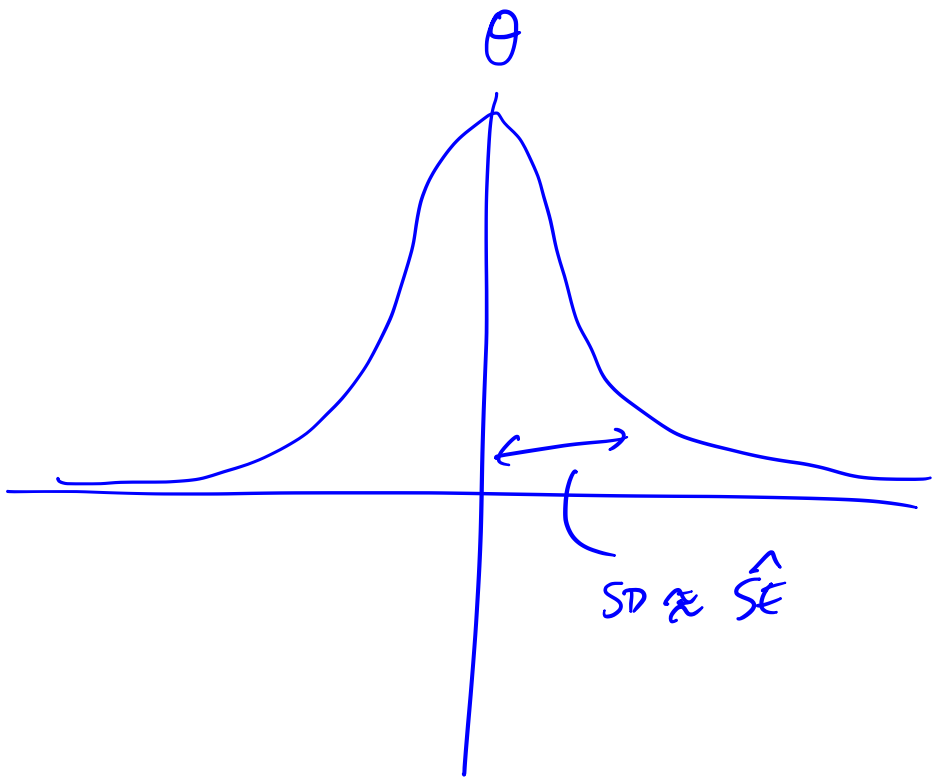
In particular, for large n:

Quantifying uncertainty

The combination of the standard error, consistency, and asymptotic normality allow us to quantify uncertainty directly through *confidence intervals*:

In particular, for large n:

- The sampling distribution of the MLE $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is approximately normal with mean $\boldsymbol{\theta}$, and standard deviation $\widehat{\text{SE}}$.

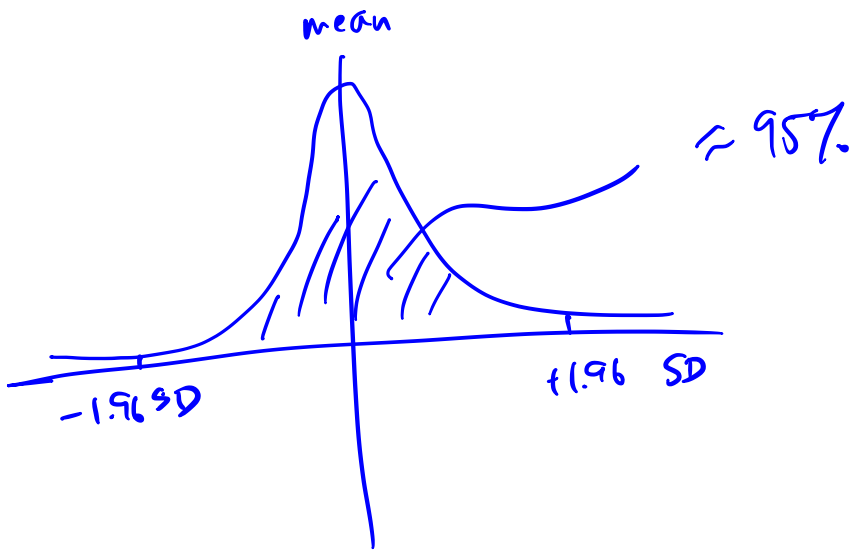


Quantifying uncertainty

The combination of the standard error, consistency, and asymptotic normality allow us to quantify uncertainty directly through *confidence intervals*:

In particular, for large n:

- The sampling distribution of the MLE $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is approximately normal with mean $\boldsymbol{\theta}$, and standard deviation $\widehat{\text{SE}}$.
- A normal distribution has $\approx 95\%$ of its mass within 1.96 standard deviations of the mean.



Quantifying uncertainty

The combination of the standard error, consistency, and asymptotic normality allow us to quantify uncertainty directly through *confidence intervals*:

In particular, for large n:

- The sampling distribution of the MLE $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is approximately normal with mean $\boldsymbol{\theta}$, and standard deviation $\widehat{\text{SE}}$.
- A normal distribution has $\approx 95\%$ of its mass within 1.96 standard deviations of the mean.
- Therefore, in 95% of our "universes", $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ will be within 1.96 $\widehat{\text{SE}}$ of the true value of $\boldsymbol{\theta}$.

Quantifying uncertainty

The combination of the standard error, consistency, and asymptotic normality allow us to quantify uncertainty directly through *confidence intervals*:

In particular, for large n:

- The sampling distribution of the MLE $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is approximately normal with mean $\boldsymbol{\theta}$, and standard deviation $\widehat{\text{SE}}$.
- A normal distribution has $\approx 95\%$ of its mass within 1.96 standard deviations of the mean.
- Therefore, in 95% of our "universes", $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ will be within $1.96\ \widehat{\text{SE}}$ of the true value of $\boldsymbol{\theta}$.
- In other words: in 95% of our universes:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} - 1.96\ \widehat{\text{SE}} \leq \boldsymbol{\theta} \leq \hat{\boldsymbol{\theta}}_{\text{MLE}} + 1.96\ \widehat{\text{SE}}.$$

$$|\hat{\theta}_{MLE} - \theta| \leq 1.96 \hat{SE}$$

Confidence intervals

We refer to $[\hat{\boldsymbol{\theta}}_{\mathsf{MLE}} - 1.96\ \widehat{\mathsf{SE}}, \hat{\boldsymbol{\theta}}_{\mathsf{MLE}} + 1.96\ \widehat{\mathsf{SE}}]$ as a *95% confidence interval* for $\boldsymbol{\theta}$.

More generally, let z_α be the unique value such that $P(Z \leq z_\alpha) = 1 - \alpha$ for $\mathcal{N}(0, 1)$ random variable. Then:

$$[\hat{\boldsymbol{\theta}}_{\mathsf{MLE}} - z_{\alpha/2}\widehat{\mathsf{SE}}, \hat{\boldsymbol{\theta}}_{\mathsf{MLE}} + z_{\alpha/2}\widehat{\mathsf{SE}}]$$

is a $1 - \alpha$ confidence interval for $\boldsymbol{\theta}$.

In R, you can get z_α using the `qnorm` function.

Confidence intervals

Comments:

- Note that *the interval is random, and $\boldsymbol{\theta}$ is fixed*!
- When $\alpha = 0.05$, then $z_{\alpha/2} \approx 1.96$.
- Confidence intervals can always be enlarged; so the goal is to construct the smallest interval possible that has the desired property.
- Other approaches to building $1 - \alpha$ confidence intervals are possible, that may yield asymmetric intervals.

Example: Linear regression

In the regression `Ozone ~ 1 + Solar.R + Wind + Temp`, the coefficient on `Temp` is 1.65, with $\widehat{\mathsf{SE}} = 0.25$.

Therefore a 95% confidence interval for this coefficient is: $[1.16, 2.14]$.

Concluding thoughts on frequentist inference

A thought experiment

You run business intelligence for an e-commerce startup.

Every day t your marketing department gives you the number of clicks from each visitor i to your site that day ($V_i^{(t)}$), and your sales department hands you the amount spent by each of those visitors ($R_i^{(t)}$).

[2] Ignore seasonality: let's suppose the true value of this multiplier is the same every day.

A thought experiment

You run business intelligence for an e-commerce startup.

Every day t your marketing department gives you the number of clicks from each visitor i to your site that day ($V_i^{(t)}$), and your sales department hands you the amount spent by each of those visitors ($R_i^{(t)}$).

Every day, your CEO asks you for an estimate of how much each additional click by a site visitor is "worth".[2]

So you:

- Run a OLS linear regression of $\mathbf{R}^{(t)}$ on $\mathbf{V}^{(t)}$.
- Compute intercept $\hat{\beta}_0^{(t)}$ and slope $\hat{\beta}_1^{(t)}$.
- Report $\hat{\beta}_1^{(t)}$.

But your boss asks: *"How sure are you of your guess?"*

[2] Ignore seasonality: let's suppose the true value of this multiplier is the same every day.

A thought experiment

Having taken MS&E 226, you also construct a 95% confidence interval for your guess $\hat{\beta}_1^{(t)}$ each day:

$$C^{(t)} = [\hat{\beta}_1^{(t)} - 1.96\ \widehat{SE}_1^{(t)}, \hat{\beta}_1^{(t)} + 1.96\ \widehat{SE}_1^{(t)}].$$

You tell your boss:

> *"I don't know what the real β_1 is, but I am 95% sure it lives in the confidence interval I give you each day."*

After one year, your boss goes to an industry conference and discovers the true value of β_1, and now he looks back at the guesses you gave him every day.

A thought experiment

How does he evaluate you? A picture:

The benefit of frequentist inference

This example lets us see why frequentist evaluation can be helpful.

More generally, the meaning of reporting 95% confidence intervals is that you "trap" the true parameter in 95% of the claims that you make, even across *different* estimation problems. (See Section 6.3.2 of [AoS].)

This is the defining characteristic of estimation procedures with good frequentist properties: *they hold up to scrutiny when repeatedly used.*