BSTA 551: Statistical Inference

Lesson 1: Introduction to Statistical Inference; Statistics

Jessica Minnier

2026-01-05

Lesson 1: Introduction to Statistical Inference

Welcome to Statistical Inference!

Course Focus: How do we learn about populations from samples?

Today's Goals

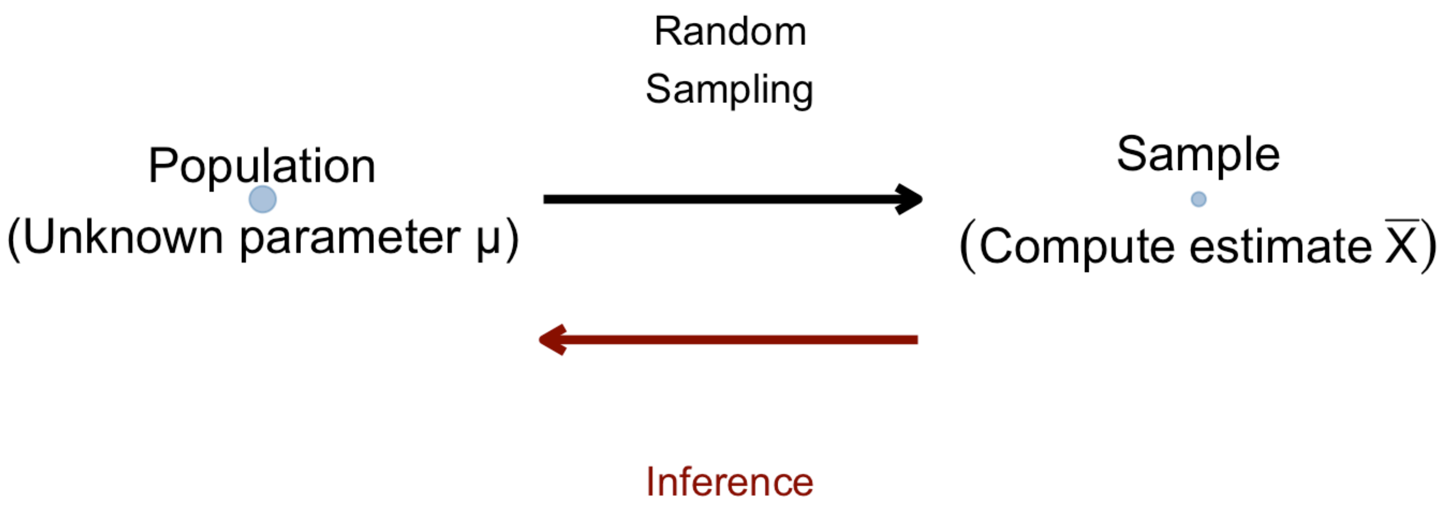
Motivating Example: Clinical Trial

A pharmaceutical company is testing a new blood pressure medication.

The Question: What is the true average reduction in systolic blood pressure?

Statistical Inference: The Big Picture

The Big Picture: Population vs Sample



Key insight: We use sample data to make inferences about population parameters.

Key Terminology (Devore 6.1)

Definitions

- **Population:** The entire collection of individuals or measurements of interest
- **Sample:** A subset of the population that we actually observe
- **Parameter:** A numerical characteristic of the population (e.g., μ, σ, p)
- **Statistic:** A numerical characteristic computed from sample data (e.g., $\bar{X}, S, \hat{p}$)

Parameters vs. Statistics

Concept	Population (Parameter)	Sample (Statistic)
Mean	μ	$\bar{X} = \frac{1}{n}\sum X_i$
Variance	σ^2	$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$
Proportion	p	$\hat{p} = X/n$
Maximum	θ (upper bound)	$\max(X_1, \ldots, X_n)$

The Sampling Process (Devore 6.2)

Random Sampling Assumptions:

1. Each observation X_i is a random variable
2. The X_i are **independent** of each other
3. Each X_i has the **same distribution** (identically distributed)

Review: Expected Value and Variance

Review: Expected Value

The **expected value** $E(X)$ is the long-run average of a random variable.

Key Properties We'll Use Today

1. $E(c) = c$ for any constant c
2. $E(cX) = c \cdot E(X)$
3. $E(X + Y) = E(X) + E(Y)$
4. $E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$

Worked Example: Expected Value of Sample Mean

Problem: If $X_1, X_2, \ldots, X_n$ are *iid* observations from a population with mean μ, what is $E(\bar{X})$?

Your Turn: Calculate Expected Value

Exercise: A hospital measures the recovery time (in days) for patients after surgery. Let X_1, X_2, X_3 be recovery times for 3 patients. The population mean recovery time is $\mu = 5$ days.

Questions:

1. What is $E(X_1)$?
2. What is $E(X_1 + X_2 + X_3)$?
3. What is $E(\bar{X})$ where $\bar{X} = \frac{X_1+X_2+X_3}{3}$?

Review: Variance

Variance measures the spread of a distribution: $\mathrm{Var}(X) = E[(X-\mu)^2]$

Key Properties We'll Use Today

1. $\mathrm{Var}(c) = 0$ for any constant
2. $\mathrm{Var}(cX) = c^2 \cdot \mathrm{Var}(X)$
3. If X and Y are **independent**: $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Worked Example: Variance of Sample Mean

Problem: If $X_1, \ldots, X_n$ are independent observations with variance σ^2, what is $\mathrm{Var}(\bar{X})$?

The Sampling Distribution

What is a Sampling Distribution?

Different samples give different values of a statistic. The **sampling distribution** describes this variability.

Sample	Sample Mean ($\bar{x}$)
1	9.90
2	10.31
3	10.03
4	10.85
5	9.17
6	9.31

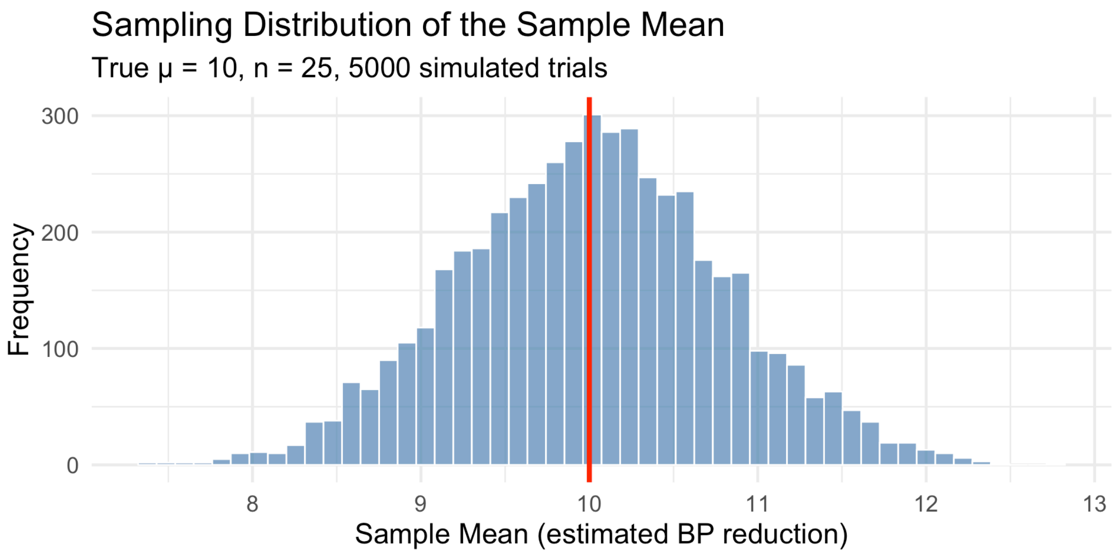
Simulation: Building a Sampling Distribution

```r
# Parameters
true_effect <- 10  # True mean BP reduction (mmHg)
true_sd <- 4       # Standard deviation
n_patients <- 25   # Patients per trial
n_trials <- 5000   # Number of simulated trials

# Simulate many clinical trials
sampling_distribution <- tibble(trial = 1:n_trials) |>
  mutate(
    sample_mean = map_dbl(trial, \(t) {
      patients <- rnorm(n_patients, true_effect, true_sd)
      mean(patients)
    })
  )

# Visualize
sampling_distribution |>
  ggplot(aes(x = sample_mean)) +
  geom_histogram(bins = 50, fill = "steelblue", alpha = 0.7, color = "white") +
  geom_vline(xintercept = true_effect, color = "red", linewidth = 1.5) +
  labs(title = "Sampling Distribution of the Sample Mean",
       subtitle = str_glue("True μ = {true_effect}, n = {n_patients}, {n_trials} simulated
       x = "Sample Mean (estimated BP reduction)", y = "Frequency")
```

Simulation: Building a Sampling Distribution



Simulation: Seeing Variance Decrease

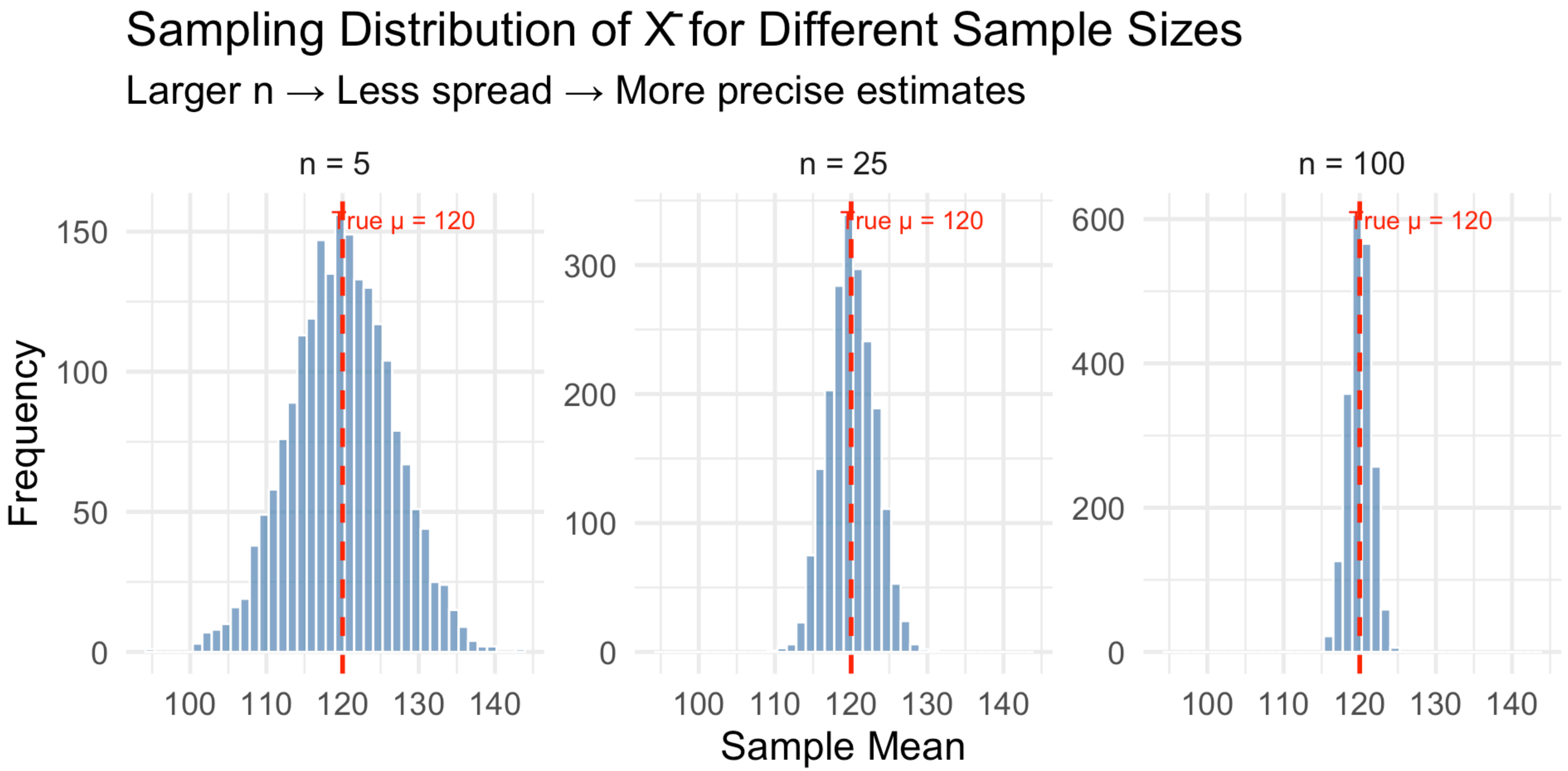
```r
# Population parameters
true_mean <- 120  # True mean systolic BP
true_sd <- 15     # Population standard deviation

# Simulate sample means for different sample sizes
simulation_data <- tibble(n = c(5, 25, 100)) |>
  cross_join(tibble(sim = 1:2000)) |>
  mutate(
    sample_mean = map2_dbl(n, sim, \(size, s) {
      mean(rnorm(size, true_mean, true_sd))
    })
  )

# Calculate observed standard deviation for each sample size
simulation_data |>
  group_by(n) |>
  summarize(
    observed_sd = sd(sample_mean),
    theoretical_sd = true_sd / sqrt(first(n))
  )
```

```
# A tibble: 3 × 3
      n observed_sd theoretical_sd
  <dbl>       <dbl>          <dbl>
```

Visualizing the Effect of Sample Size



Understanding Optimization

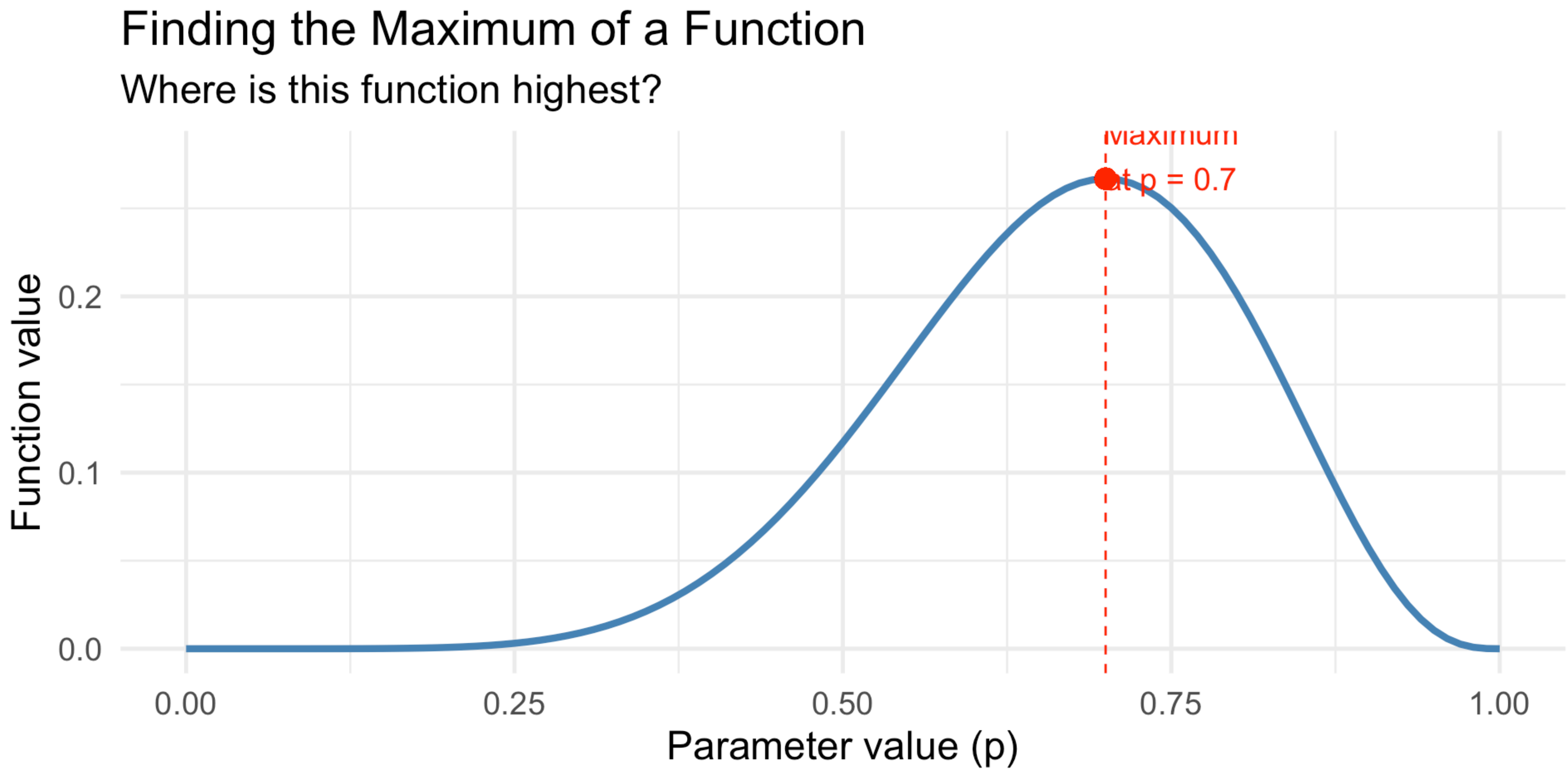
Why Optimization Matters in Statistics

Many statistical methods require finding the "best" value of a parameter.

Examples:

- **Maximum Likelihood:** Find the parameter value that makes the observed data most probable
- **Least Squares:** Find the parameter value that minimizes prediction errors
- **Minimum Variance:** Find the estimator with the smallest spread

Optimization: The Graphical Intuition



The maximum occurs where the function reaches its peak.

Concrete Example: Finding the Best Estimate

Scenario: In a clinical trial, 7 out of 10 patients respond to treatment. What's the best estimate of the true response rate p?

Grid Search: A Simple Numerical Approach

Idea: Try many values and see which gives the largest result.

```r
# Try different values of p
grid_search <- tibble(p = seq(0.01, 0.99, by = 0.01)) |>
  mutate(
    likelihood = dbinom(7, size = 10, prob = p)
  )

# Find the maximum
grid_search |>
  slice_max(likelihood, n = 1)
```

```
# A tibble: 1 × 2
      p likelihood
  <dbl>      <dbl>
1   0.7      0.267
```

Using R’s Optimizer

R has built-in functions to find maximums and minimums more precisely:

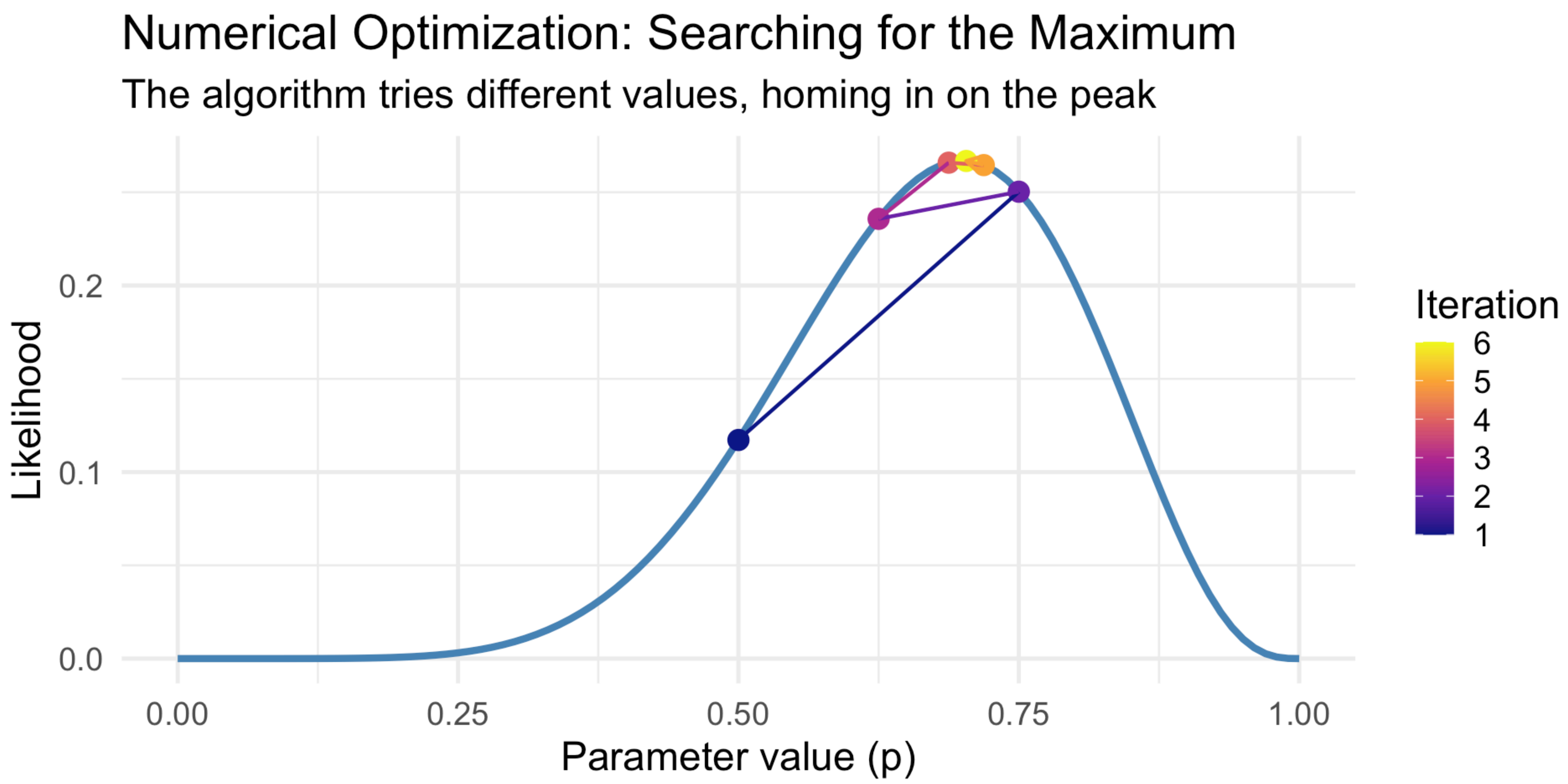
```r
# Define the likelihood function
likelihood_function <- function(p) {
  dbinom(7, size = 10, prob = p)
}

# Use optimize() to find the maximum
# Note: optimize finds MINIMUM by default, so we negate for maximum
result <- optimize(
  f = function(p) -likelihood_function(p),  # Negative to find max
  interval = c(0, 1)                         # Search between 0 and 1
)

# The maximum occurs at:
cat("Maximum likelihood estimate: p =", result$minimum)
```

```
Maximum likelihood estimate: p = 0.6999843
```

How Numerical Optimization Works



Key idea: The algorithm evaluates the function at different points and iteratively narrows in on the maximum.

Your Turn: Numerical Optimization

Exercise: A diagnostic test correctly identifies a disease in 18 out of 25 patients who have it. Find the maximum likelihood estimate for the test's sensitivity p.

```r
# Fill in the blanks:
likelihood_fn <- function(p) {
  dbinom(___, size = ___, prob = p)  # What goes here?
}

result <- optimize(
  f = function(p) -likelihood_fn(p),
  interval = c(0, 1)
)

result$minimum  # This is the MLE
```

When Optimization Gets Harder

Sometimes we need to optimize over multiple parameters or complex functions:

```r
# Example: Finding mean and SD that best fit data
patient_data <- c(120, 135, 128, 142, 131, 125, 138, 129, 133, 127)

# Negative log-likelihood for normal distribution
neg_log_lik <- function(params) {
  mu <- params[1]
  sigma <- params[2]
  if (sigma <= 0) return(Inf)  # sigma must be positive
  -sum(dnorm(patient_data, mean = mu, sd = sigma, log = TRUE))
}

# Use optim() for multiple parameters
result <- optim(par = c(130, 10), fn = neg_log_lik)
cat("MLE for mean:", round(result$par[1], 2), "\n")
```

```
MLE for mean: 130.8
```

```r
cat("MLE for SD:", round(result$par[2], 2), "\n")
```

```
MLE for SD: 6.13
```

```r
cat("Compare to sample mean:", round(mean(patient_data), 2))
```

```
Compare to sample mean: 130.8
```

Summary and Looking Ahead

Lesson 1 Summary

Key Concepts:

1. **Statistical Inference:** Using sample data to learn about population parameters
2. **Parameters vs. Statistics:**
   - Parameters (μ, σ, p): Fixed but unknown population values
   - Statistics ($\bar{X}, S, \hat{p}$): Calculated from sample data
3. **Sampling Distribution:** The distribution of a statistic across many samples
   - $E(\bar{X}) = \mu$ (centered at population mean)
   - $\mathrm{Var}(\bar{X}) = \sigma^2/n$ (precision improves with larger n)
4. **Numerical Optimization:** Finding maximum/minimum values
   - Grid search: try many values
   - `optimize()`: efficient numerical search

Lesson 1 Practice Problems

1. Calculate $E(\bar{X})$ and $\mathrm{Var}(\bar{X})$ for a sample of size $n = 16$ from a population with $\mu = 50$ and $\sigma = 12$.
2. Use `optimize()` to find the MLE for p when you observe 23 successes in 40 trials.
3. A quality control engineer samples 5 items from a production line. If the population mean weight is 100g with SD = 5g, what is the expected value and variance of the sample mean?
4. Simulate the sampling distribution of the sample median for $n = 30$ observations from a Normal(100, 15) distribution. Compare to the sampling distribution of the sample mean.

Next Lesson Preview

Lesson 2: Point Estimation; Bias, Variance, and MSE

- What is a point estimator?
- Bias: systematic error in estimation
- Standard error: precision of estimators
- Mean Squared Error: combining bias and variance
- The bias-variance tradeoff

References

- Devore, Berk, and Carlton. *Modern Mathematical Statistics with Applications* (Springer). Chapters 6.1, 6.2
- Chihara and Hesterberg. *Mathematical Statistics with Resampling and R* (Wiley). Chapter 6.

Questions?

Thank you!