Basic Statistics

Sean McGrath

June 17, 2025

Yale SCHOOL OF PUBLIC HEALTH
Big Data Summer Immersion

Outline

- Basic Setup
- Point Estimation
- Hypothesis Testing
- Confidence Intervals

What is statistical inference?

Nobody has responded yet.

Hang tight! Responses are coming in.

Big Picture: Statistical Inference

- We are interest in some aspect, called a **parameter,** of our population of interest

- We collect a **sample** of data from our population

- We make **inference** on our parameter based on our sample
  - We construct an **estimate** of our parameter
  - We characterize the **uncertainty** of our estimate
  - We make a **decision/conclusion**

Brief Review: Random Variables and Distributions

- **Motivation:**
  – We can characterize probabilities of events (e.g., last lecture)
  – It is often convenient to re-express events into a type of summary variable

- **Example:**
  – Suppose one flips a coin 50 times and we are interested in the number of heads
  – It is quite complicated to express our events of interest in terms of the individual tosses
  – Instead, we define a **random variable** X to denote the number of heads
  – The **distribution** of X describes the probability that X equals any particular value
    - i.e., $P(X = 0), P(X = 1), \dots, P(X = 50)$
    - Typically denoted by $f(x|\theta)$ for some parameter θ

Setting

- Let θ denote our parameter of interest
  – E.g., average height of individuals in the USA

- We collect a random sample, denoted by $X_1, \ldots, X_n$
  – E.g., X_i is the height of the ith individual sampled

- Often, our sample is **independent** and **identically distributed** (i.i.d.)
  – **Independent:** The X_i's are all independent
  – **Identically Distribution:** The X_i all have the same probability distribution

- **Question:** What settings may involve data that are not i.i.d.?

Estimators

- **Estimator:** Any function of our sample $X_1, \ldots, X_n$
  - We will denote it by $\hat{\theta}$

- **Examples:**
  - The sample mean, $\frac{1}{n}\sum_{i=1}^{n} X_i$
  - The minimum value, $\min_i X_i$
  - The maximum value, $\max_i X_i$

- **Goal:** Construct a "good" estimator of our parameter of interest θ

How would you define a "good" estimator?

Nobody has responded yet.

Hang tight! Responses are coming in.

Bias and Variance

- **Bias:** The bias of an estimator, $\hat{\theta}$, is given by

$$\mathrm{Bias}(\hat{\theta}) := \mathrm{E}(\hat{\theta}) - \theta$$

  where E is the "expectation" (average value) of $\hat{\theta}$
  - **Intuition:** Characterizes how close the average value of our estimator is to the truth
  - Called **unbiased** if the bias equals 0 for all θ

- **Variance:** The variance of an estimator, $\hat{\theta}$, is given by

$$\mathrm{Variance}(\hat{\theta}) := \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)^2\right]$$

  - **Intuition:** Describes the variability of the estimator

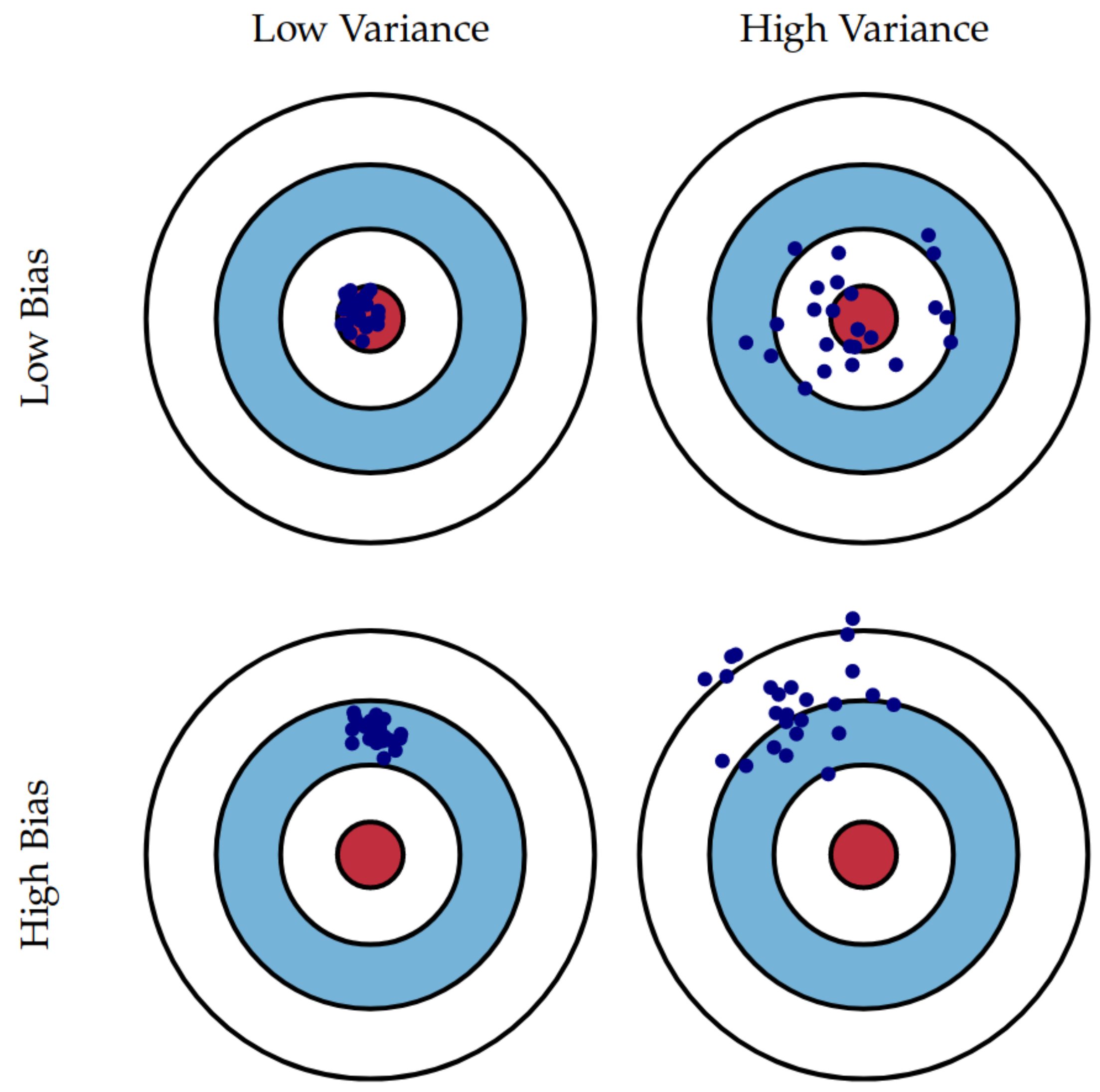


Fig. 1 Graphical illustration of bias and variance.

https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html

Mean Squared Error

- **Mean Squared Error (MSE):** The MSE of an estimator, $\hat{\theta}$, is given by

$$\mathrm{MSE}(\hat{\theta}) := \mathrm{E}\left[(\hat{\theta} - \theta)^2\right]$$

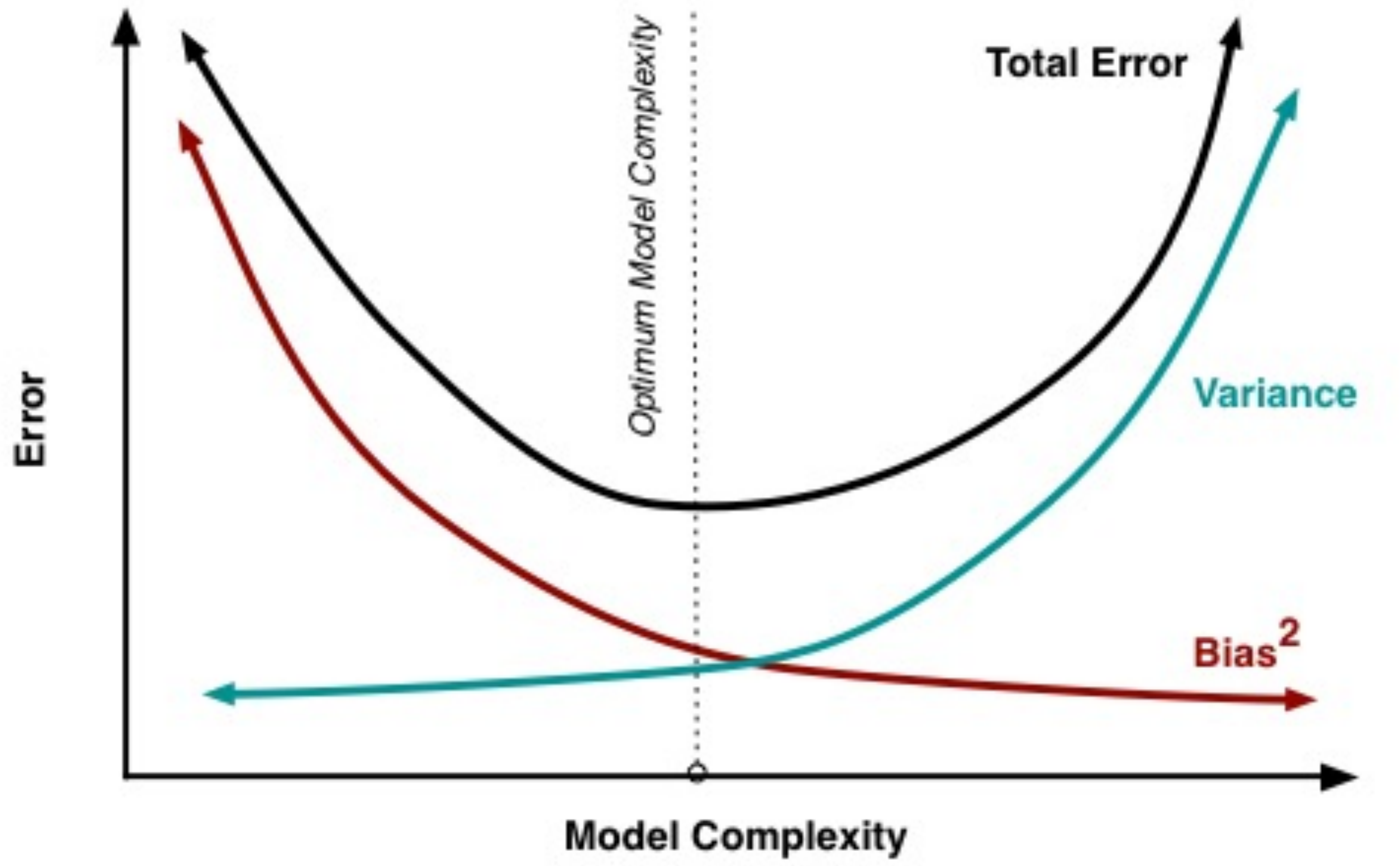
- **Intuition:** Characterizes how far away our estimator is from the truth on average

- **Important Feature:** The MSE incorporates both the bias and variance of $\hat{\theta}$**,**

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Bias}(\hat{\theta})^2 + \mathrm{Variance}(\hat{\theta})$$

- Optional Exercise: Prove the above equality

A Common Theme: Bias-Variance Tradeoff



https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

Example: Comparing the sample mean and median

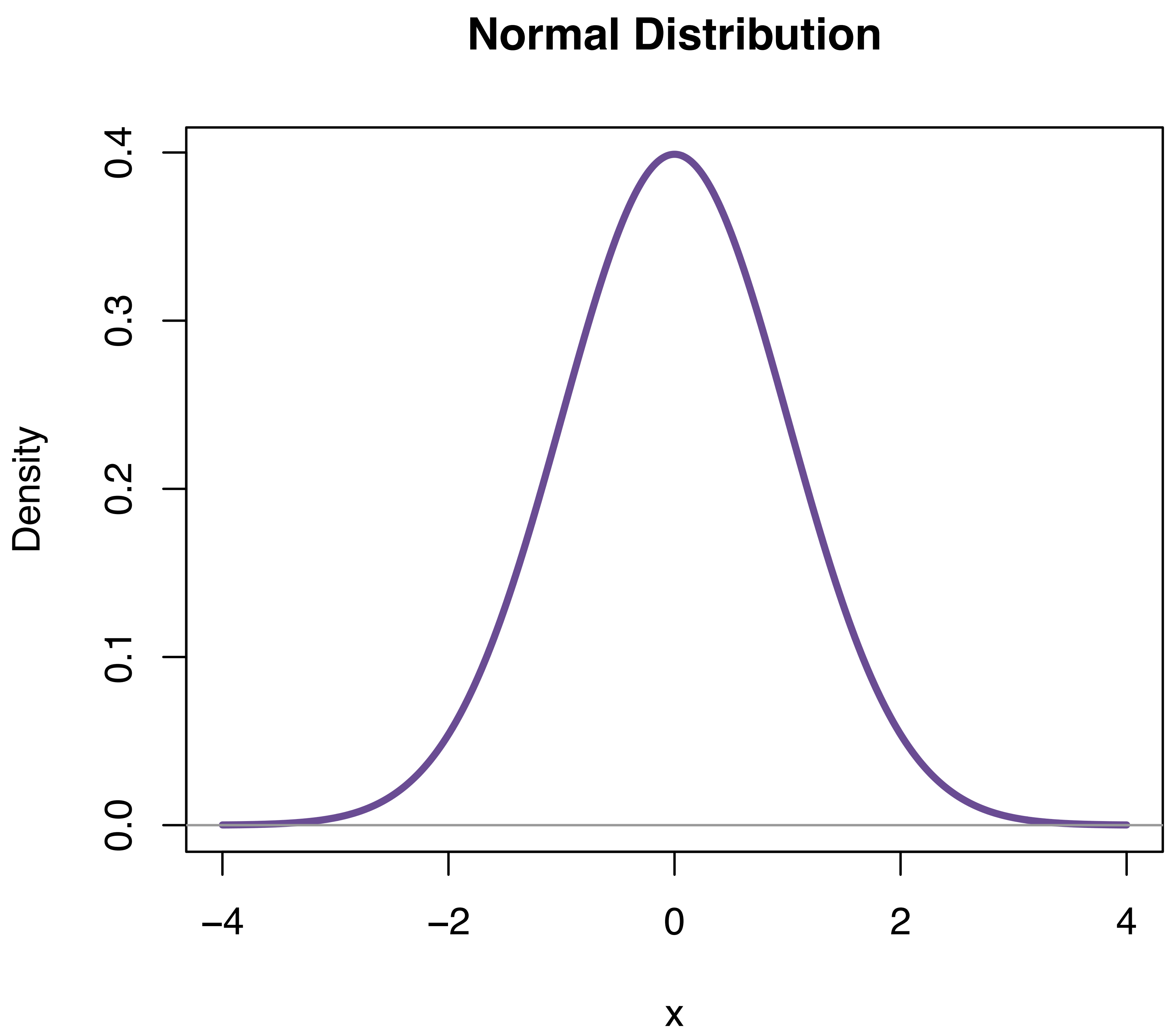
- Let's compare properties of the sample *mean* versus the sample *median*

- **Setting:**
  - Consider symmetric distributions, i.e., the *distribution* mean and median are equal
  - Consider the following three distributions
    - Normal
    - Logistic
    - Laplace
  - Sample size: $n = 21$

- **Criterion:**
  - Consider the mean squared error

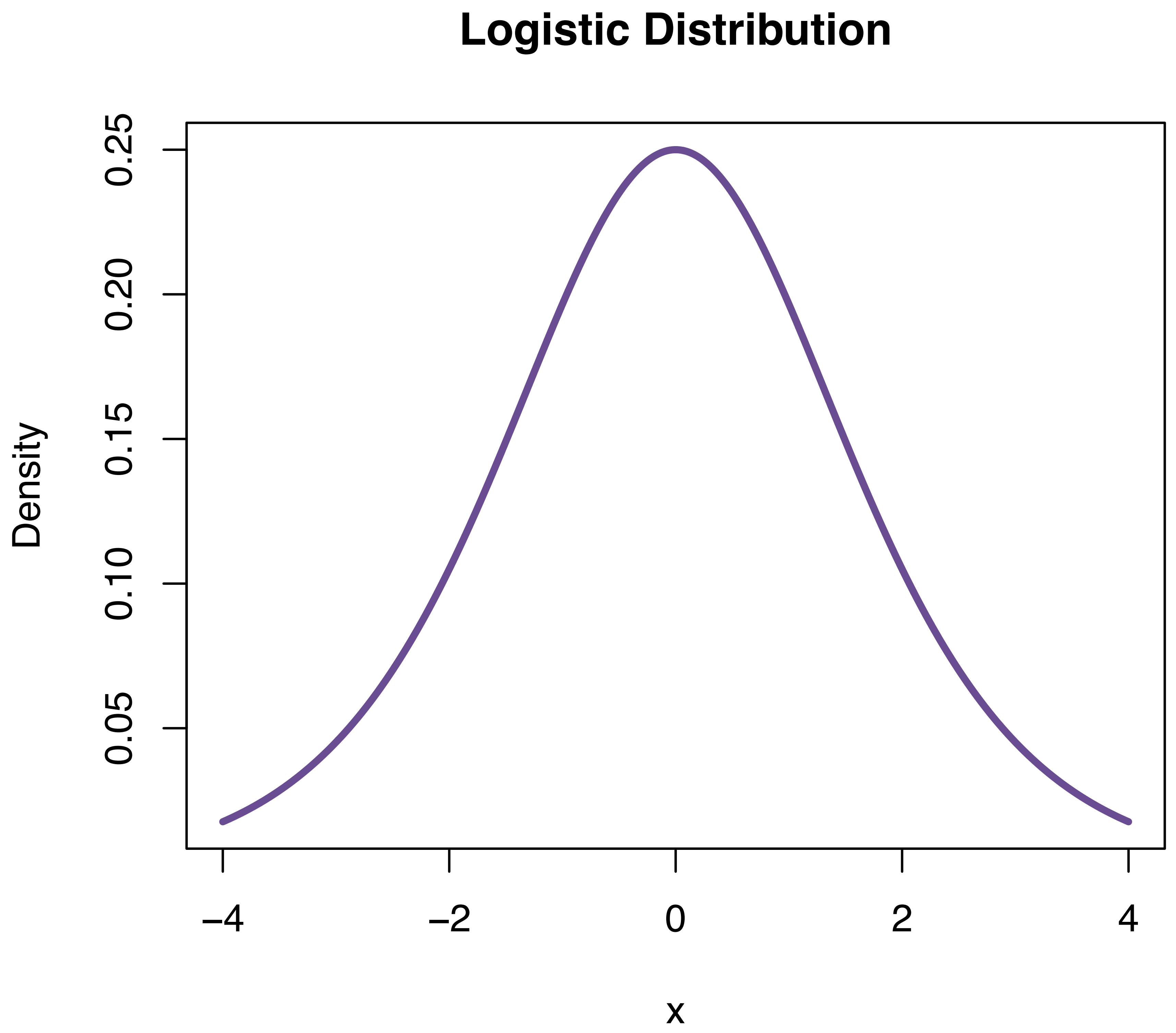
Normal Distribution

- MSE of sample mean: 0.0476
- MSE of sample median: 0.0733
- The sample *mean* is better!



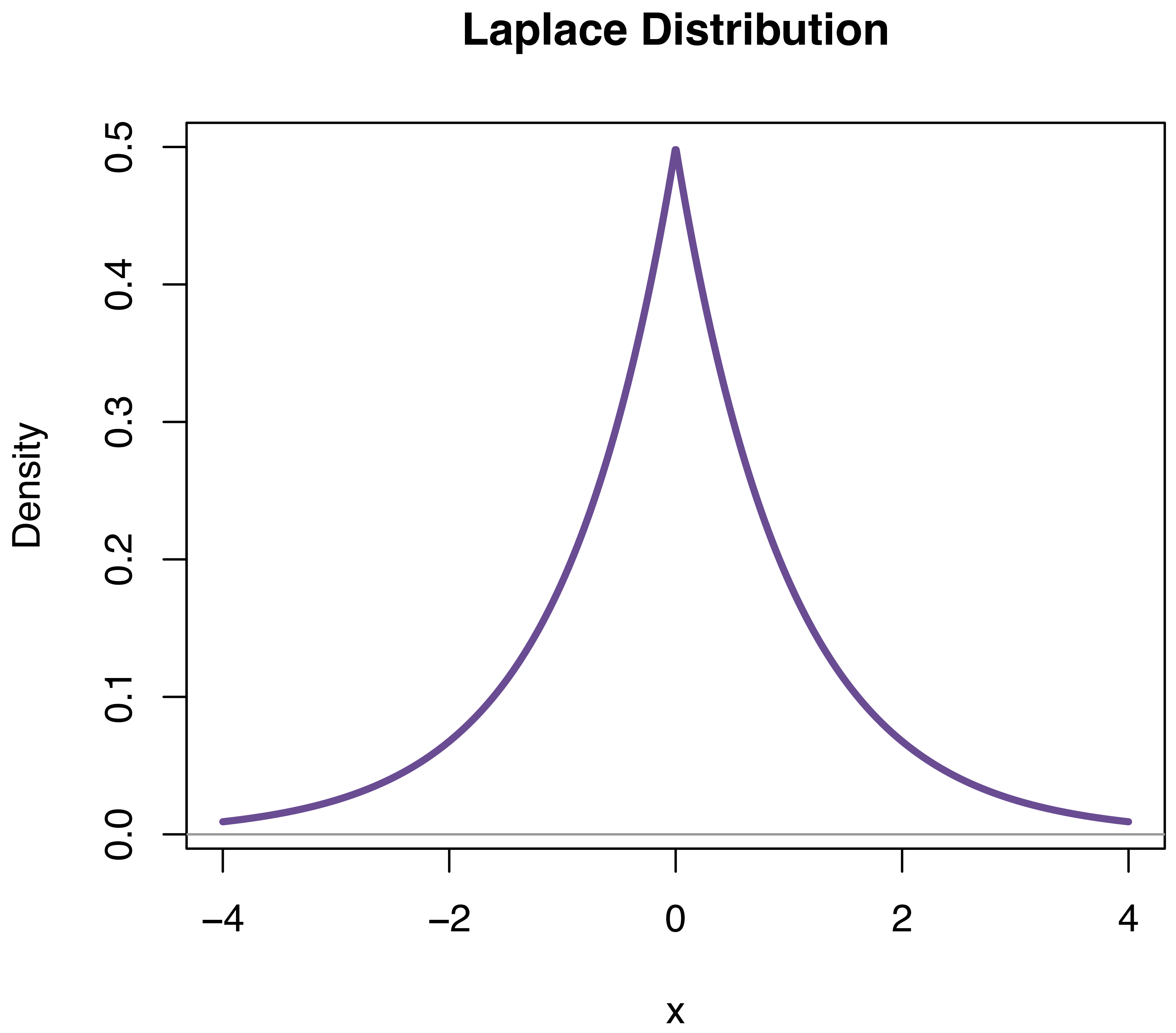
Logistic Distribution

- MSE of sample mean: 0.1567
- MSE of sample median: 0.1903
- Once again, the sample *mean* is better!



Laplace Distribution

- MSE of sample mean: 0.0952
- MSE of sample median: 0.0654
- Here, the sample *median* is better!



Exercise

- Let's revisit the sample mean versus median comparison for the normal distribution. Using your favorite programming language, perform a simulation-based evaluation of the MSE of the sample mean and sample median for a sample size of 51. How do the results compare to the setting with a sample size of 21?

- **Hint:**
  - For iteration $1, \ldots, n$ for some large n (e.g., 10,000), perform the following
    - Sample 51 observations from the normal distribution with mean 0 and variance 1
    - Compute the sample mean and standard deviation
  - Compute the $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta\right)^2$ where $\hat{\theta}_i$ is the sample mean (median) and θ is 0

- **If time permits:** Compare the MSE of the sample mean and sample median across a range of sample sizes from 10 to 100. Plot the MSE of the two estimators as a function of the sample size

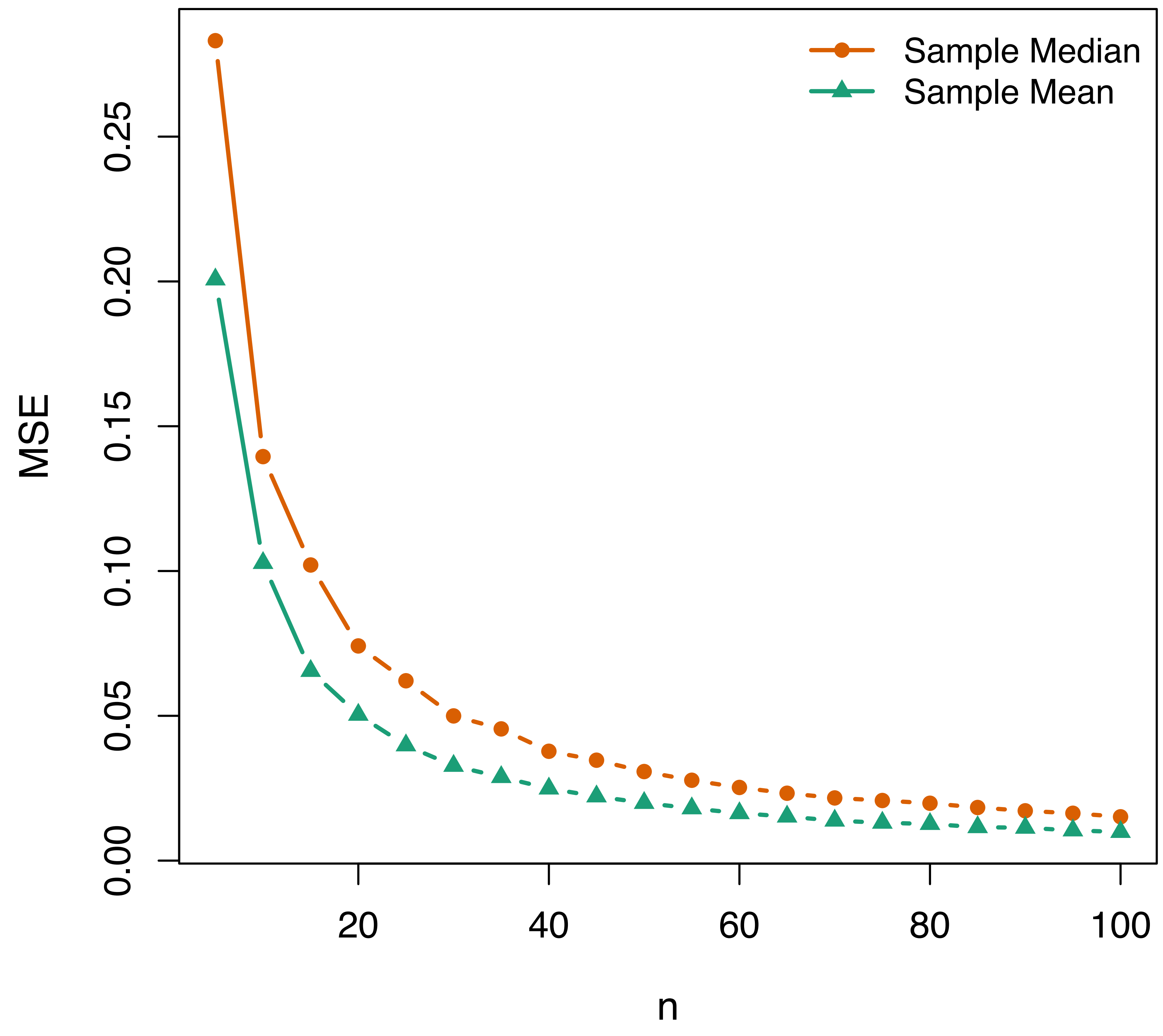
Solution (in R)

```r
# Set seed and constants
set.seed(1234)
n_rep <- 10000
mu <- 0; sd <- 1

# Iteratively sample data from normal distribution and compute mean and median
mean_ests <- median_ests <- rep(NA, times = n_rep)
for (i in 1:n_rep){
  dat <- rnorm(n = 51, mean = mu, sd = sd)
  mean_ests[i] <- mean(dat)
  median_ests[i] <- median(dat)
}

# Evaluate MSE
mean((mean_ests - mu)^2)  # 0.01969709
mean((median_ests - mu)^2)  # 0.03061826
```

Solution to Optional Exercise



Hypothesis Testing

- **Motivation:** We would like a framework to make decisions about θ

- **Hypothesis:** A statement about a parameter θ
  - **Null Hypothesis** $H_0: \theta \in \Theta_0$
  - **Alternative Hypothesis** $H_1: \theta \in \Theta_1$

- **Hypothesis Test:** A rule that specifies whether to reject H_0 or fail to reject H_0 based on a sample $X_1, \dots, X_n$

Example

- **Setting:** Suppose that the X_i follow a normal distribution with mean θ and variance 1

- **Two-Sided Hypothesis:**
  - $H_0: \theta = 0$
  - $H_1: \theta \neq 0$

- **One-Sided Hypothesis:**
  - $H_0: \theta \leq 0$
  - $H_1: \theta > 0$

Typical Application

- The null hypothesis is often set to reflect the status quo, especially in scientific applications
  - **Example:**
    - **Null hypothesis:** 5-year risk of death is the *same* between individuals receiving the experimental intervention versus those receiving the placebo
    - **Alternative hypothesis:** 5-year risk of death is *different* between individuals receiving the experimental intervention versus those receiving the placebo

- Therefore, rejecting the null hypothesis is thought to lead to a change in the status quo

- This also suggests being conservative in the sense that one would like to design hypothesis tests that rarely reject the null hypothesis when it is indeed true

Types of Errors

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = $1 - \beta$
Not rejected	Correct decision True negative Probability = $1 - \alpha$	Type II error False negative Probability = β

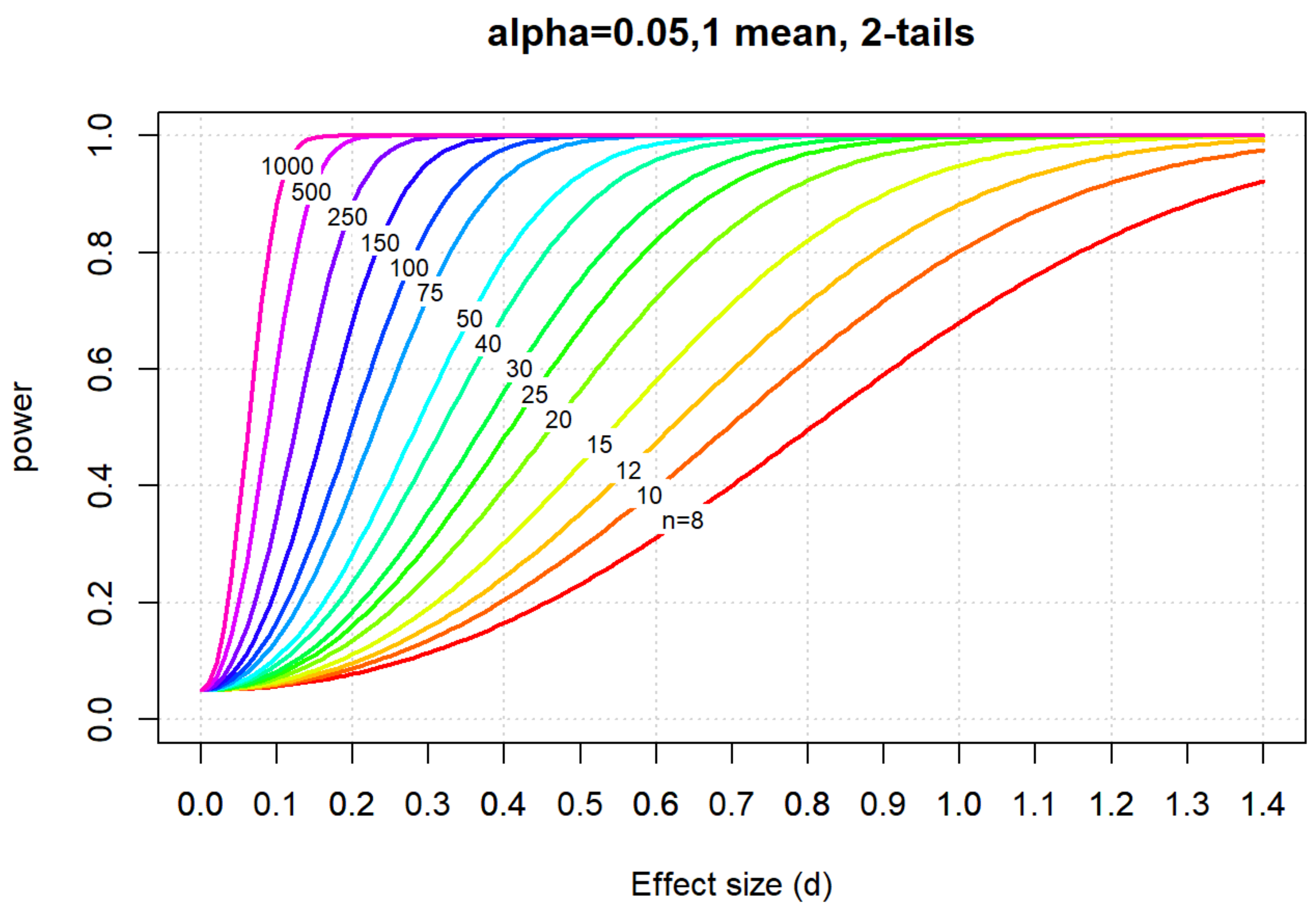
Scribbr

https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

Evaluating Hypothesis Tests

- Like our discussion of evaluating and comparing estimators, one can evaluate and compare hypothesis tests

- **Power:**
  - The **power function** for a test is $\beta(\theta) := P(\text{reject } H_0 | \theta)$
  - We want the power of a test to be *large* under H_1 and *small* under H_0
  - Rich literature on theory of optimal testing that is based on evaluating power of different hypothesis tests



https://courses.washington.edu/psy524a/_book/power.html

P-values

- **Informal Definition:**
  - The probability of observing a result as or more extreme that what was observed, if the null hypothesis were true

- **Intuition:** Small p-value indicates that the observed data were not very compatible with the state of nature implied by the null hypothesis

- **Usage:** Hypothesis tests can be constructed so that we reject the null hypothesis if the p-value is sufficiently small

Example: Wildfire Exposure and Test Scores

- **Setting:**
  - – We hypothesize that students living in regions with high exposure to wildfires have lower test scores than those in regions with low exposure
  - – Suppose we collect (or have access to) data on standardized test scores for high-school students in regions with high exposure and low exposure to wildfires

- **Hypotheses:**
  - **Null Hypothesis:** The average test score for students in regions with high exposure to wildfires (μ_1) is the *same* as the average test score for students in regions with low exposure (μ_0)
    - – i.e. $H_0: \mu_1 = \mu_0$
  - **Alternative Hypothesis:** The average test score for students in regions with high exposure to wildfires (μ_1) is *different* than the average test score for students in regions with low exposure (μ_0)
    - – i.e. $H_1: \mu_1 \neq \mu_0$

Example: Wildfire Exposure and Test Scores

- **Hypothetical Data:**
  - Region with high exposure:
    - Sample size: 59
    - Average test score: 72%
    - Standard deviation of test scores: 6
  - Region with low exposure:
    - Sample size: 137
    - Average test score: 75%
    - Standard deviation of test scores: 5.5

- **Two-sample t-test**
  - P-value: 0.0008
  - Reject the null hypothesis at the 0.05 significance level

Hypothesis Testing in Science

- The appropriateness of null hypothesis significance testing in scientific research has been of great debate

- **Discussion:** What are merits and limitations of hypothesis testing?

- **Some merits:**
  - Formal, structured decision-making framework
  - Supports control of certain error rates

- **Some limitations:**
  - No indication of the strength and direction of the association
  - Only provide information on the compatibility with the null hypothesis
  - Commonly misinterpreted in practice
  - Pressure to obtain "significant" results has led to p-hacking, selective reporting, publication bias, etc.

Confidence Intervals

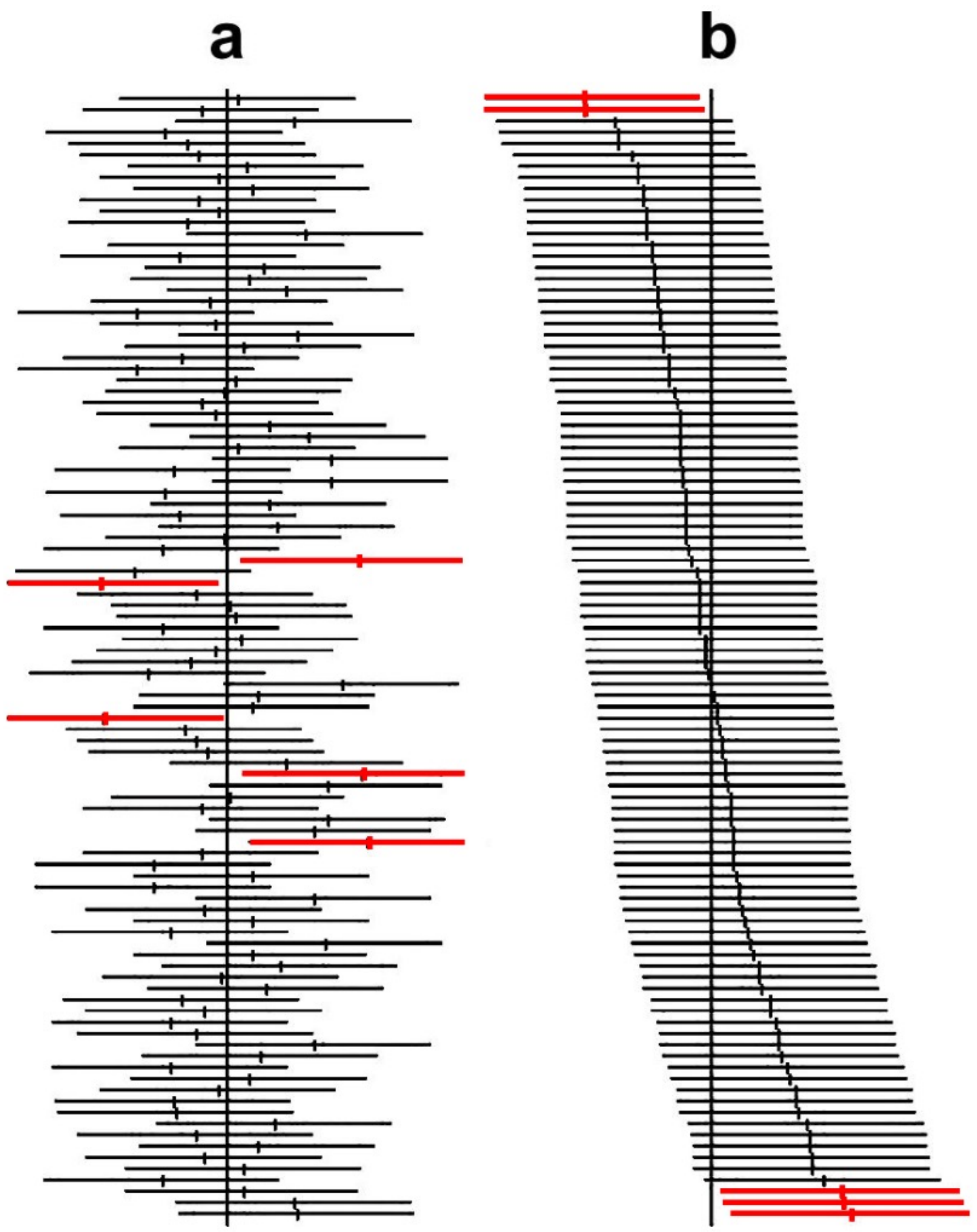
- Estimation of θ:
  - So far, we have talked about *point* estimators of θ
  - We now begin discussing *interval* estimators of θ

- Confidence Intervals (CI): A $100\times(1-\alpha)\%$ CI is an interval $[L(X), U(X)]$ such that for all θ

$$P(\theta \in [L(X), U(X)] | \theta) = 100\times(1-\alpha)\%$$

- **Interpretation:**
  - If one were to repeatedly perform the experiment and construct CIs in this manner, $100\times(1-\alpha)\%$ of the intervals would contain θ

Confidence Intervals



Christensen E. J Hepatol. 2007;46(5):947-54

- **Setting:** 100 estimates and CIs in repeated samples (via simulations)
- **Panel A:** Unsorted CIs
- **Panel B:** Sorted CIs (by lower bound)

Confidence Intervals: The Wald Method

- **Wald Method:**
  - One of the most common ways to construct CIs
  - CI takes the form

$$\hat{\theta} \pm z_{1-\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta})$$

    - $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of a standard normal distribution
    - $\widehat{\mathrm{SE}}(\hat{\theta})$ denotes the estimated standard error of $\hat{\theta}$ (i.e., the standard deviation of the distribution of $\hat{\theta}$)
  - Justified by asymptotic normality of $\hat{\theta}$

- **Example:**
  - Let's revisit our example on wildfire exposure and test scores
  - Recall that the group of students with low exposure had an average test score that was 3% higher than the high exposure group
  - The 95% CI around the percent difference is $[-4.74, -1.26]$

Confidence Intervals

- As with methods for point estimation and hypothesis tests, we can evaluate the performance of different methods to construct confidence intervals

- **Properties:**
  - Often based on examining the finite sample behavior of
    - **Coverage probabilities**, i.e., the proportion of CIs that contain θ
    - **Confidence interval length**, i.e., $U(X) - L(X)$

Exercise

- Consider a sample of $n = 20$ i.i.d. observations that follow a Bernoulli distribution with success probability p. Let $\hat{p}$ denote the sample proportion. A 95% CI for p based on the Wald method is given by

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Recall that the coverage probability of a confidence interval is the proportion of CIs that contain the true p. Using your favorite programming language, find the coverage probability when $p = 0.50$ and when $p = 0.01$

- **If time permits:** Plot the coverage probability as a function of p from 0 to 1

Solution (in R): p = 0.5

```
# Set seed and constants
set.seed(1234)
n_rep <- 10000
p <- 0.5
n <- 20

# Iteratively sample data, compute CI, and check if p falls within the CI
cov <- rep(NA, times = n_rep)
for (i in 1:n_rep){
  dat <- rbinom(n = n, size = 1, prob = p)
  point_est <- mean(dat)
  se_est <- sqrt((point_est * (1 - point_est)) / n)
  lb <- point_est - 1.96 * se_est
  ub <- point_est + 1.96 * se_est
  cov[i] <- ifelse(p >= lb & p <= ub, 1, 0)
}

# Evaluate coverage
mean(cov)  # 0.9585
```

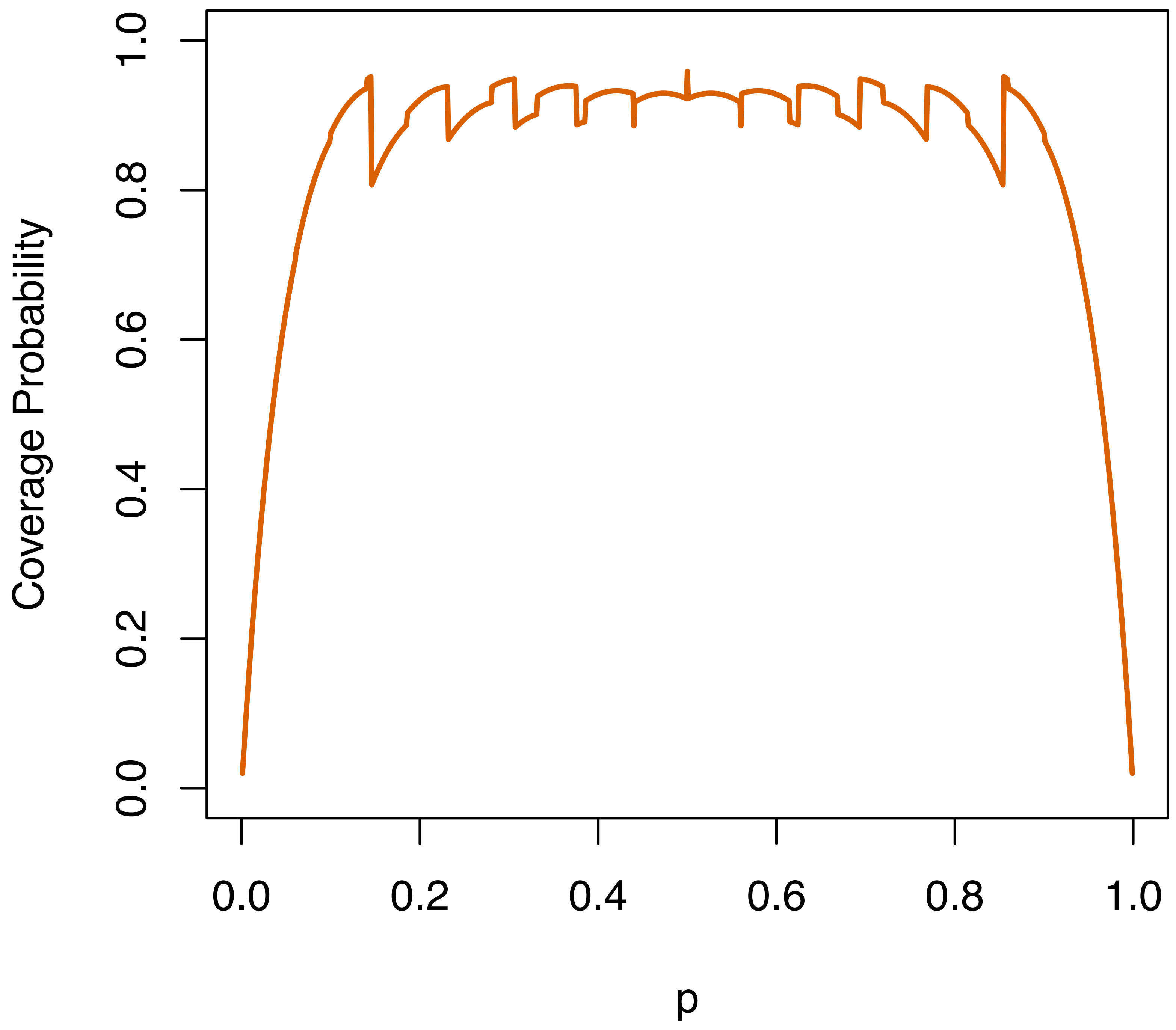
Solution (in R): p = 0.01

```r
# Set seed and constants
set.seed(1234)
n_rep <- 10000
p <- 0.01
n <- 20

# Iteratively sample data, compute CI, and check if p falls within the CI
cov <- rep(NA, times = n_rep)
for (i in 1:n_rep){
  dat <- rbinom(n = n, size = 1, prob = p)
  point_est <- mean(dat)
  se_est <- sqrt((point_est * (1 - point_est)) / n)
  lb <- point_est - 1.96 * se_est
  ub <- point_est + 1.96 * se_est
  cov[i] <- ifelse(p >= lb & p <= ub, 1, 0)
}

# Evaluate coverage
mean(cov)  # 0.186
```

Solution to Optional Exercise



Further Reading

- **Reference (again!):** *Statistical Inference*, 2nd Edition, by G. Casella and R. L. Berger, Duxbury: Thomson Learning Inc (2002).

- **Topics:** Further details on
  - Estimation strategies
  - Hypothesis testing
  - Confidence intervals