Statistical Estimation and Sampling Distributions

Mahbub Latif, PhD

November 2025

Plan

- Estimation
- Sampling distributions
  - Sampling distribution of sample mean
  - Sampling distribution of sample proportion
- Confidence intervals
  - Confidence intervals for population mean
  - Confidence intervals for population proportion

Statistical inference

- Descriptive statistics deal with describing observations of a sample, e.g. sample men, sample variance, etc. are tools of descriptive statistics
- Inferential statistics deal with making a statement (conclusion) about a population characteristic using the infromation obtained from a sample
- There are two methods of statistical inference
  - Estimation and test of hypothesis
- Two methods of estimation
  - point estimation and interval estimation (confidence interval)

Point Estimation

Parameters

- A parameter is a property of a probability distribution
  - E.g. mean, variance or a particular quantile of a probability distribution may be a parameter
- Parameters are usually unknown and the main goal of statistical inference is to estimate parameters using sample data

Example (Machine breakdown)

- Let p_0 be the probability of machine breakdown due to "operator misuse"
- p_0 is a parameter because it is an unknown quantity of the corresponding probability distribution

Example (Milk contents)

- Let μ and σ^2 be mean and variance of probability distribution of milk contents of a container
- μ and σ^2 are parameters of the distribution of milk contents.

Statistics

- Statistic is a property of a sample, e.g. sample mean, sample variance, etc. are example of statistics
- Statistic is a function of sample observations and it can be used to estimate unknown parameters
- Statistics are random variables whose observed values can be calculated from a set of observed data observations.

Statistics

- Let $x_1, \ldots, x_n$ be a random sample from a probability distribution $f(x)$
- Any function the sample observations, say $T(x_1, \ldots, x_n)$, is a statistic
  - Sample mean $\bar{x}$ is a statistic

$$T_1(x_1, \ldots, x_n) = \bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

  - Sample variance s^2 is also a statistic

$$T_2(x_1, \ldots, x_n) = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Estimation

- *Estimation* is a procedure by which the information contained within a sample is used to investigate properties of the population from which the sample is drawn

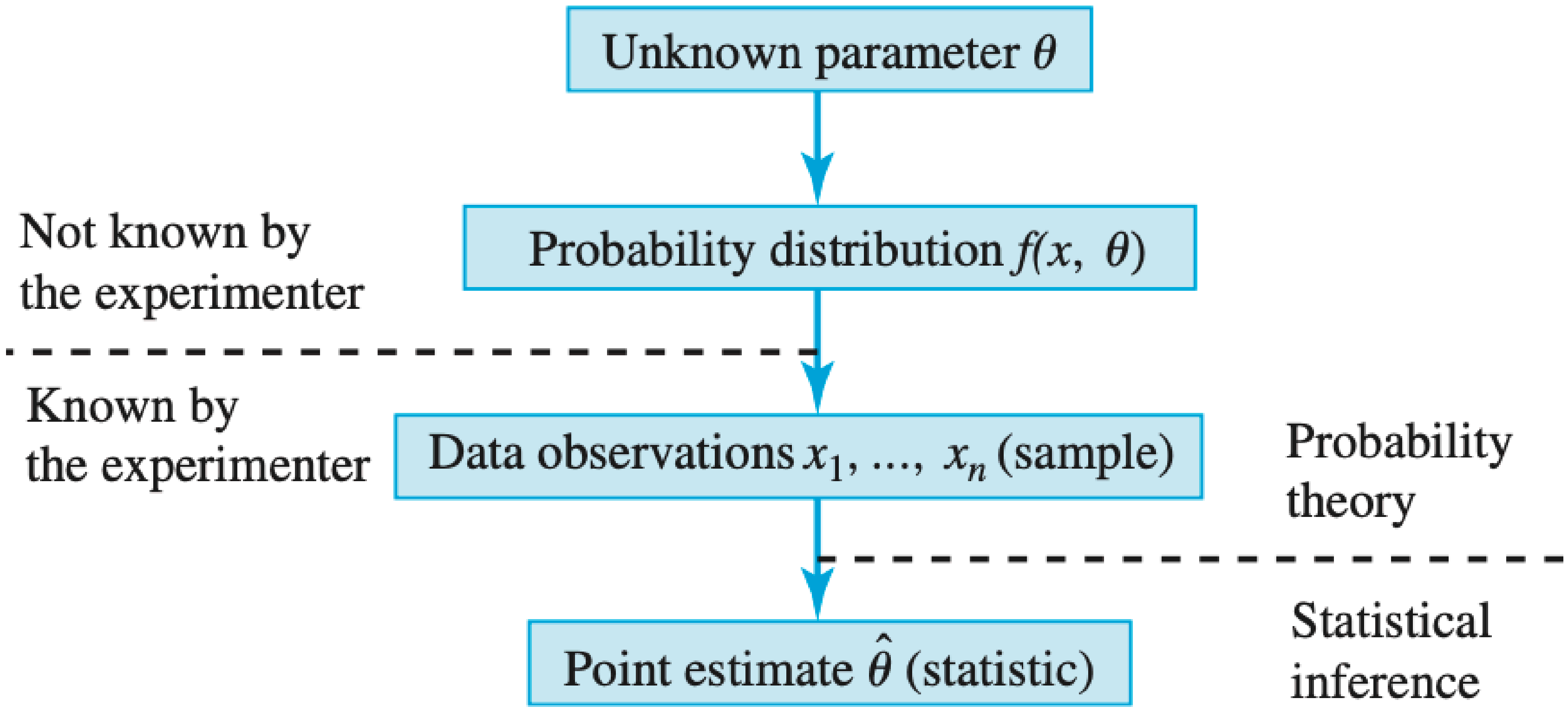
- A point estimate of an unknown parameter θ is a statistic $\hat{\theta}$ that represents a "*best guess*" of the value of θ

Estimation

- A point estimate $\hat{\theta}$ may not be exactly equal to the corresponding parameter θ, but a good point estimator is a good indicator of the actual value of the parameter

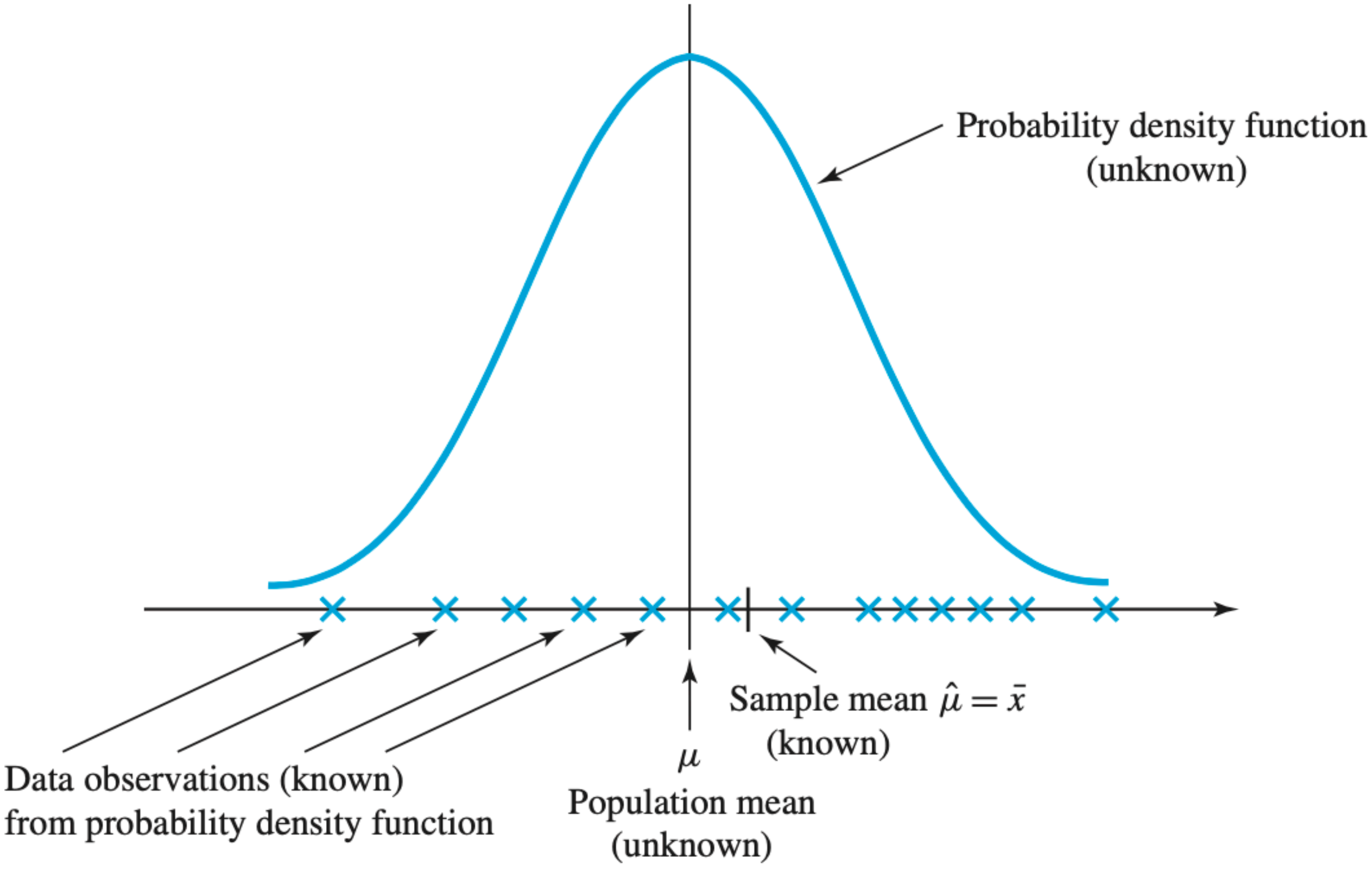
- A point estimate can only be as good as the data set from which they are calculated, e.g., whether sample is randomly selected, representative of the population, etc.

The relationship between a point estimate $\hat{\theta}$ and a parameter θ

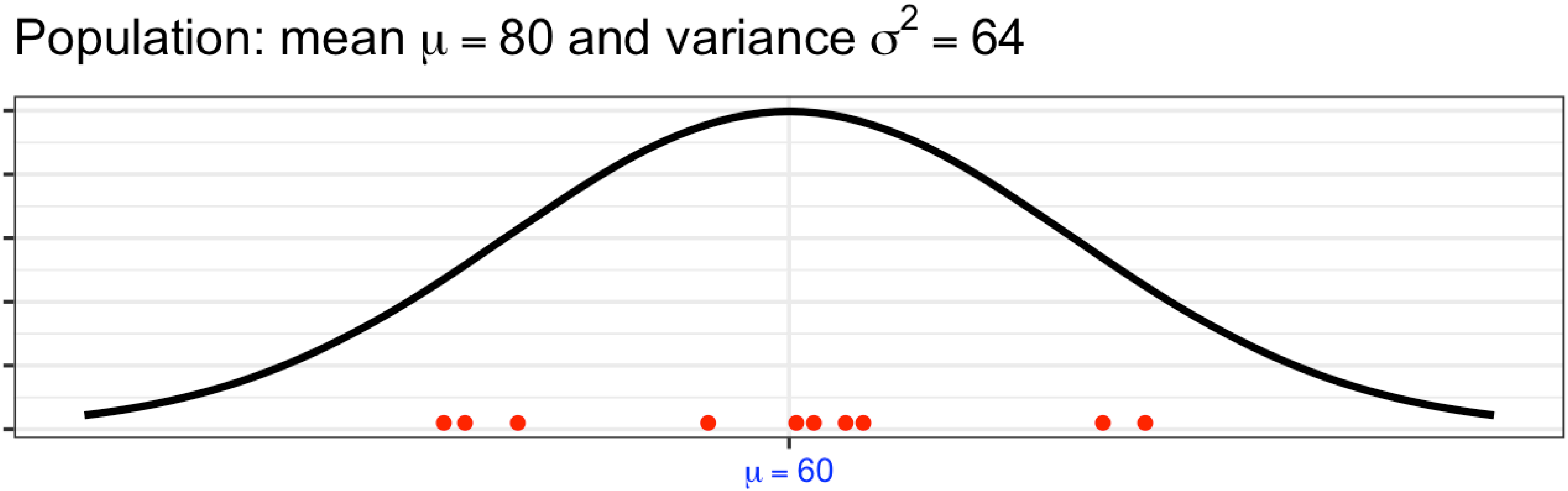


The statistic $\hat{\theta}$ is the "best guess" of the parameter θ.

Estimation of the population mean

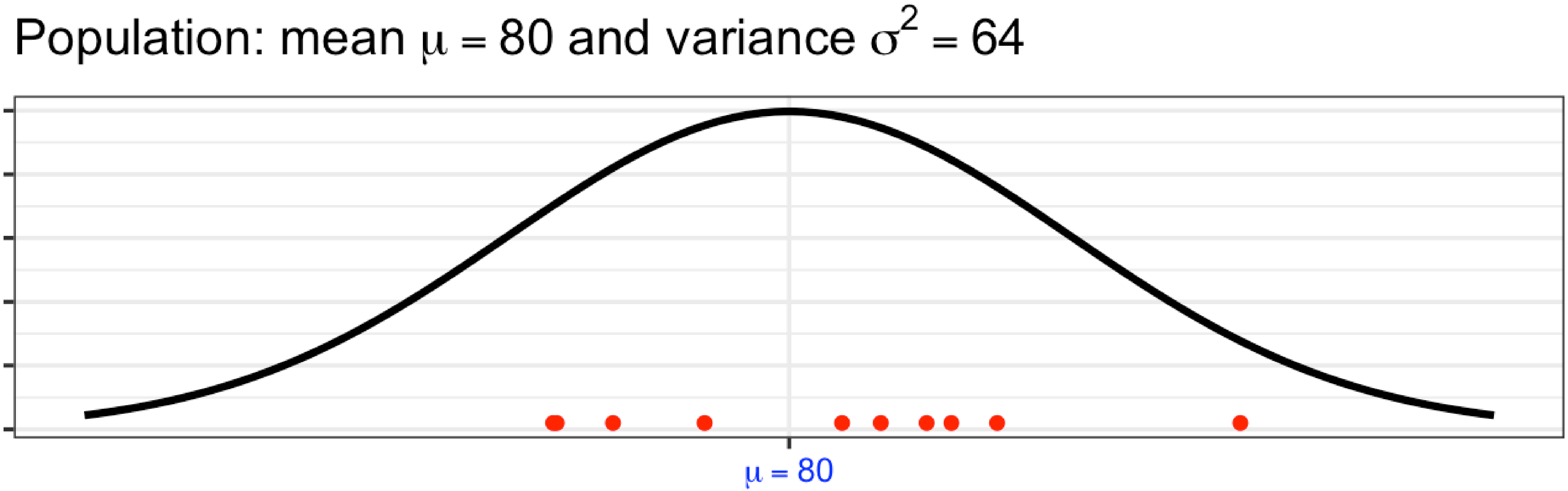


Estimation (Sample I)



- Sample observations: **72.3, 77.7, 82.1, 70.8, 81.6, 80.2, 80.7, 88.9, 70.2, 90.1**
- Point estimates: $\bar{x} = \hat{\mu} = 79.46$ and $s^2 = \hat{\sigma}^2 = 47.9$

Estimation (Sample II)



- Sample observations: **75, 81.5, 73.3, 92.8, 82.6, 73.4, 83.9, 85.9, 84.6, 77.6**
- Point estimates: $\bar{x} = \hat{\mu} = 81.06$ and $s^2 = \hat{\sigma}^2 = 39.11$

Estimation of population proportion

- Consider estimating a parameter p_0, the probability that a machine breakdown is due to operator misuse
- Suppose a sample of n machine breakdowns is recorded, of which x_0 are due to operator misuse
- The point estimate of unknown parameter p_0

$$\hat{p}_0 = \frac{x_0}{n}$$

Estimation of population proportion

- Estimate of population proportion due to operator misuse

Sample I

type	Frequency
Electrical	9
Mechanical	24
Misuse	13

Sample II

type	Frequency
Electrical	10
Mechanical	25
Misuse	11

$$\hat{p}_0 = \frac{13}{46} = 0.28$$

$$\hat{p}_0 = \frac{11}{46} = 0.24$$

Summary

- Let $x_1, \ldots, x_n$ be a random sample from a population with mean μ and variance σ^2
  - The point estimate of μ

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{m} x_i$$

  - The point estimate of σ^2

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Summary

- Suppose $x_1, \ldots, x_n$ be a random sample from $B(1, p)$
  - $\sum X \sim B(n, p)$
  - The point estimate of p

$$\hat{p} = \frac{\sum x_i}{n}$$

Sampling distribution

Sampling distribution

- Probability distribution of a *statistic* (e.g., $\hat{\mu}$, $\hat{\theta}$, $\hat{p}_0$, etc.) is known as a sampling distribution of the statistic
  - E.g. probability distribution of the sample mean $\bar{X}$ is its sampling distribution
- The main goal of statistical inference is to estimate unknown population characteristics (parameters) using sample observations

Sampling distribution

- A number of samples of a specific size (say n) can be drawn from the population and from each of these samples, the statistic of interest (e.g. $\bar{X}$, $\hat{p}$, etc.) can be calculated

- The distribution of these statistics constitute the sampling distribution

Sampling distribution of a sample mean

- Let $x_1, \ldots, x_n$ be a random sample from a population with mean μ and variance σ^2
  - The sample mean: $\hat{\mu} = \bar{x} = (1/n) \sum x$
- For a large n, the sampling distribution of sample mean $\bar{x}$ follows a **normal distribution**

$$\hat{\mu} = \bar{X} \sim N(\mu, \sigma^2/n)$$

- Note this result is valid for any population, e.g. skewed or bell-shaped

Sampling distribution of a sample mean

- The standard deviation (sd) of a sampling distribution is known as the *standard error* (se)
  - The standard error of sample mean

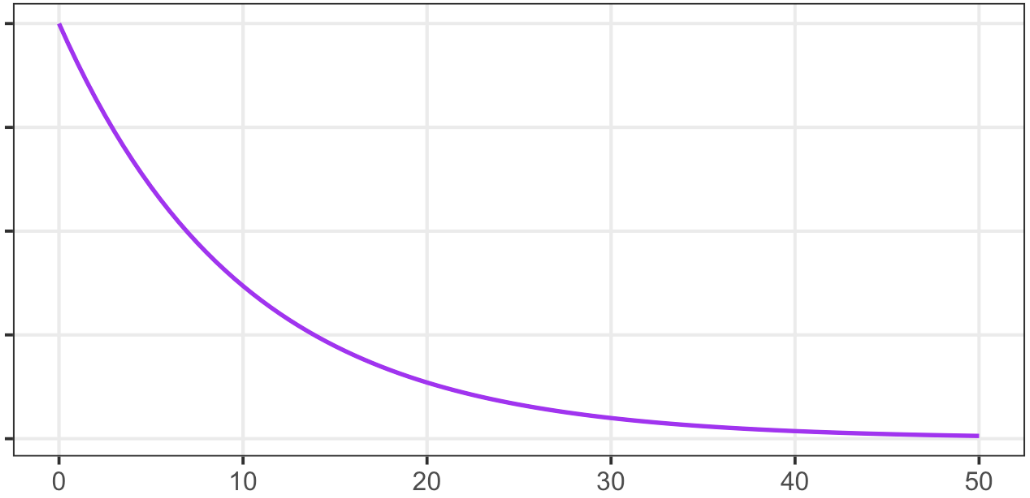
$$se(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

  - The corresponding Z statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

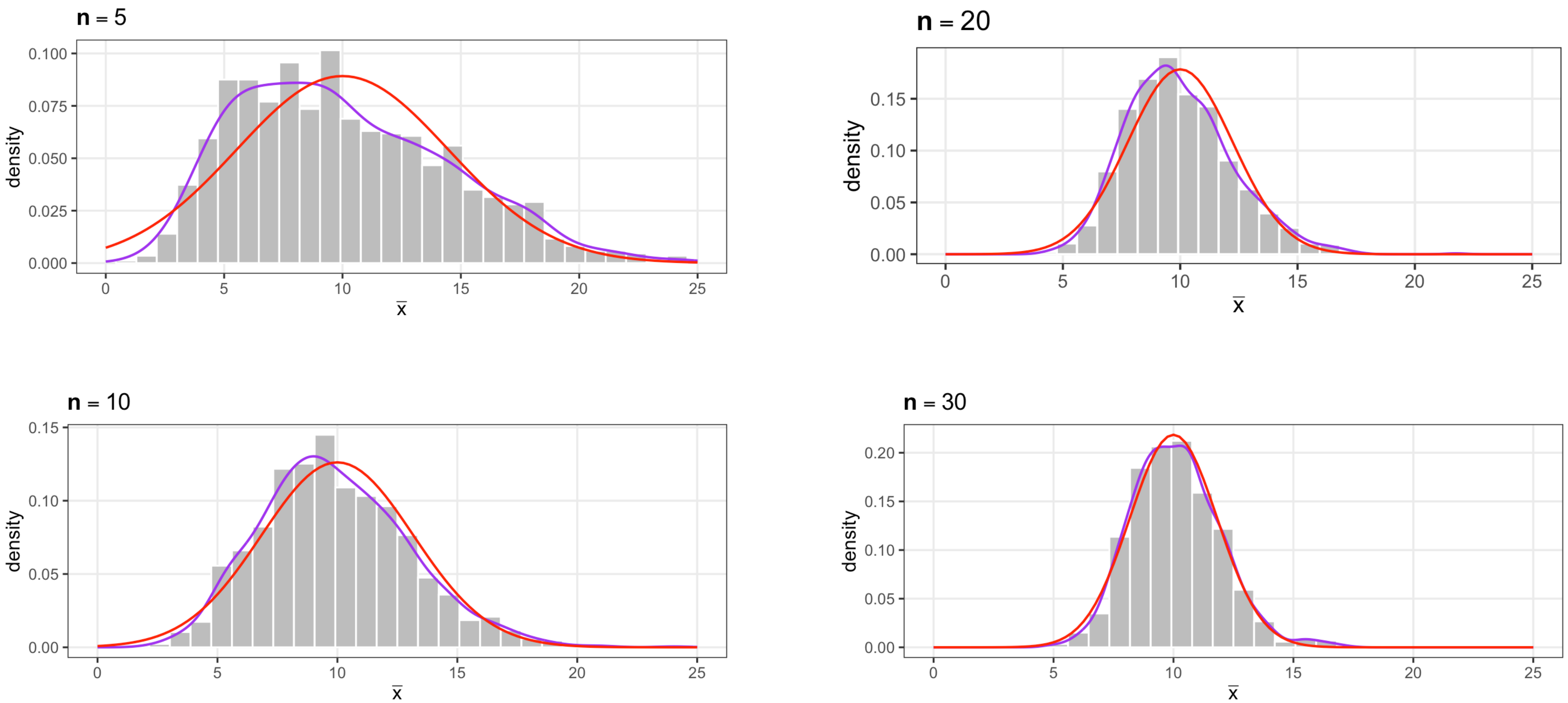
Sampling distribution of a sample mean

Population: Exponential distribution with parameter $\lambda = 0.1$

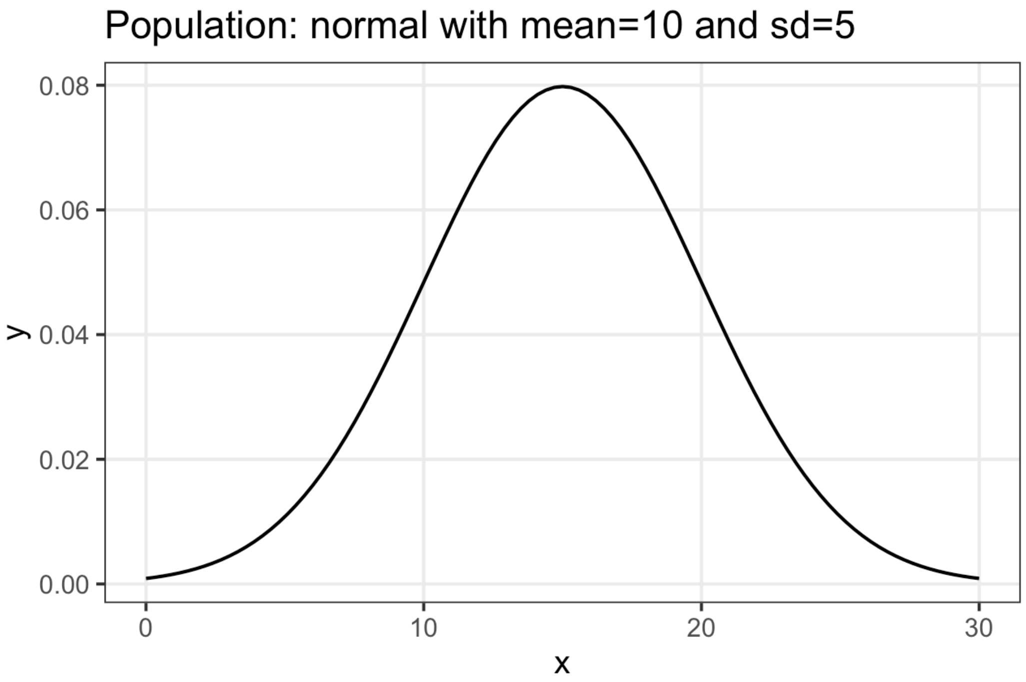


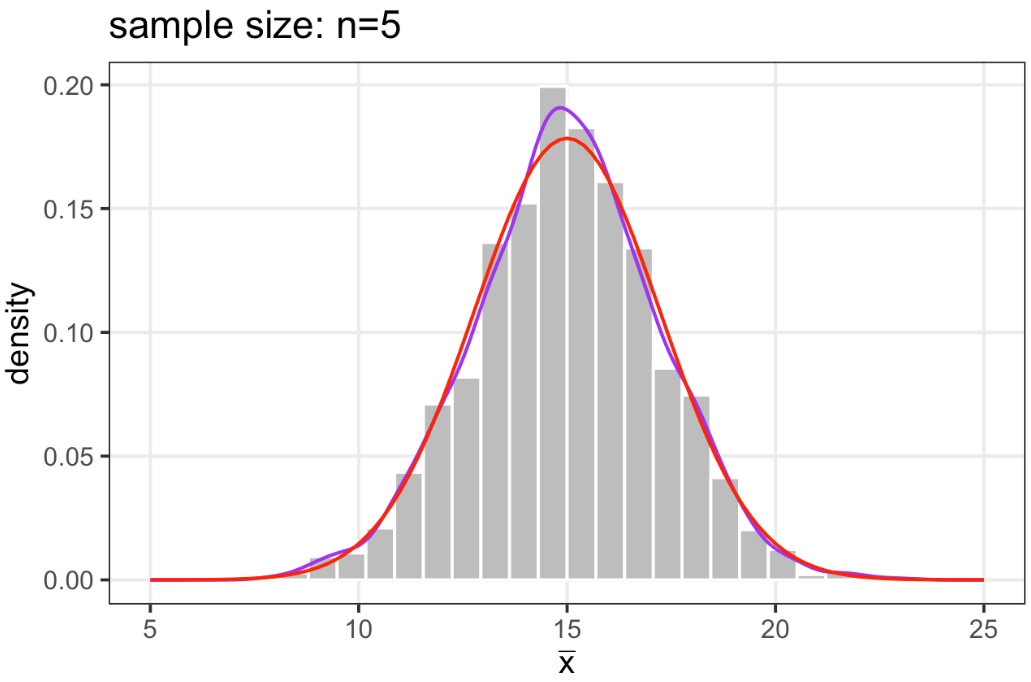
- **Sample I**: 9.2, 7.3, 7.5, 2.4, 10.8
  - Sample mean: $\bar{x} = 7.44$
- **Sample II**
  - 10.3, 12.9, 12.5, 5.5, 3
  - $\bar{x} = 8.84$
- **Sample III**
  - 12.9, 9.9, 5.1, 20.1, 4.2
  - $\bar{x} = 10.44$

Population: Exponential distribution



Population: Normal distribution





Exercise 7.3.24

- Suppose that components have weights that are normally distributed with $\mu = 341$ and $\sigma = 2$.
- An experimenter measures the weights of a random sample of 20 components in order to estimate μ.
- What is the probability that the experimenter's estimate of μ will be less than 341.5?

Exercise 7.3.24

- The estimate of μ is the sample mean

$$\hat{\mu} = \bar{x} = \sum x/20$$

- The sampling distribution of sample mean

$$\hat{\mu} = \bar{x} \sim N\left(341, 4/20\right)$$

- The probability that $\hat{\mu}$ less than 341.5 is

$$P\left(\hat{\mu} < 341.5\right) = \Phi\left(\frac{341.5 - 341}{2/\sqrt{20}}\right) = \Phi\left(1.12\right) = 0.8686$$

Exercise 7.3.2

- Consider a sample of $X_1, \ldots, X_n$ of normally distributed random variables with mean μ and variance $\sigma^2 = 1$.
- If $n = 10$, what is the probability that

$$P\Big(|\mu - \bar{X}| \leq 0.3\Big)$$

- What is this probability when $n = 30$?

t-distribution

- We have shown, irrespective of the population distribution, for a large n

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- If population variance is unknown and needs to be estimated from the sample of n observations

> The corresponding Z statistic does not follow the standard normal distribution and it follows a distribution known as t-distribution

- Similar to the standard normal distribution, t-distribution has a mean 0 and is symmetric about its mean

t-distribution

- The following t statistic follows a t distribution with $(n-1)$ degrees of freedom

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

  - $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ $\rightarrow$ the sample standard deviation

- Quantiles of t-distribution with different df can be obtained from a t-table

Comparison between t and standard normal distributions

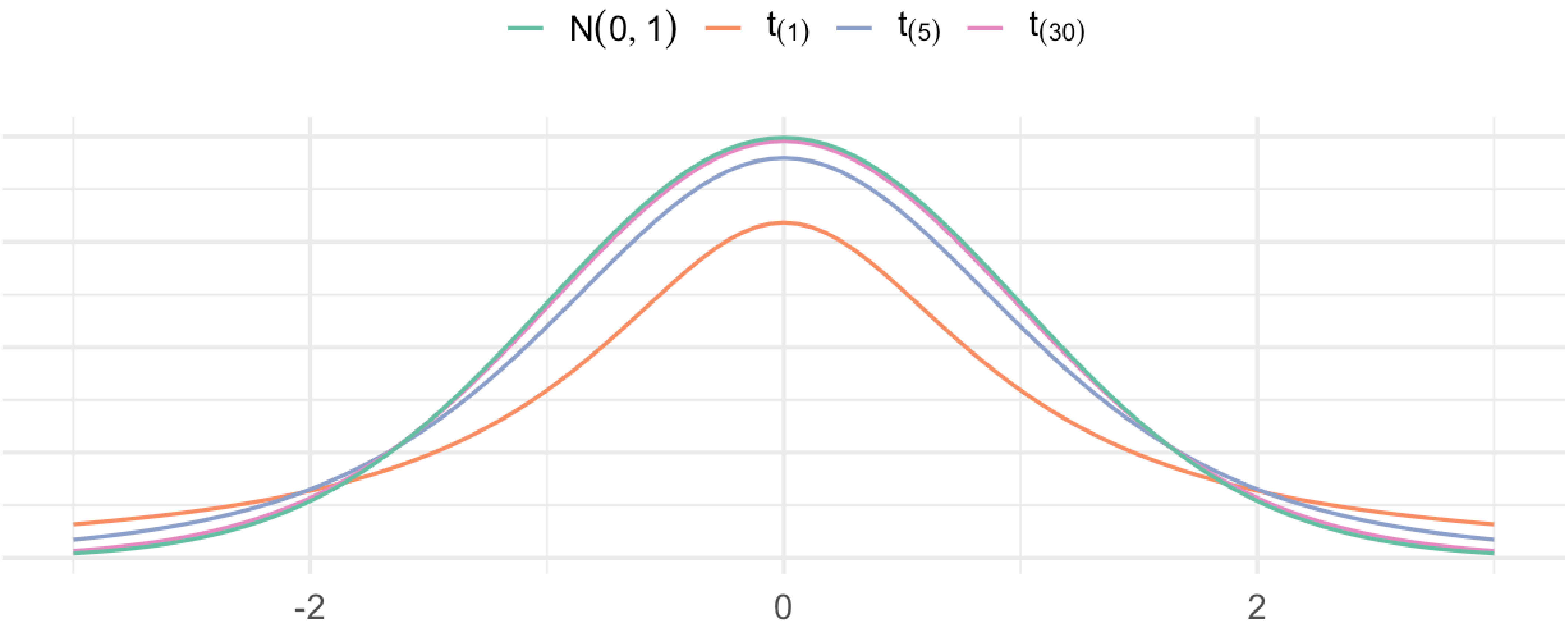
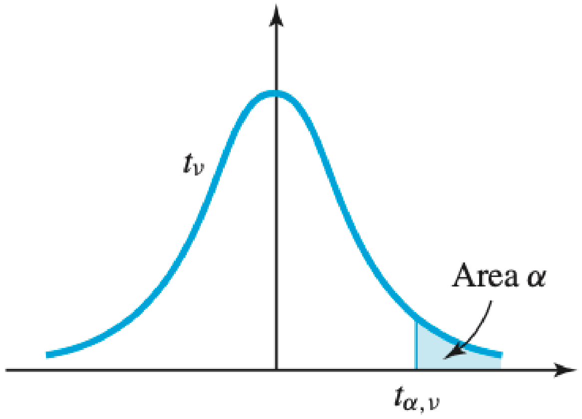


Table III: Critical Points of the t-Distribution



Degrees of freedom ν	α 0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781

$$P(t_{(1)} > 6.314) = 0.05$$

$$P(t_{(9)} > 1.833) = 0.05$$

$$P(Z > 1.645) = 0.05$$

Exercise 7.3.9

- The breaking strengths of 35 pieces of cotton thread are measured.
- The sample mean is $\bar{x} = 974.3$ and the sample variance is $s^2 = 452.1$.
- Construct a point estimate of the average breaking strength of this type of cotton thread.
- What is the standard error of your point estimate?

Exercise 7.3.7

- Consider a sample $X_1, \ldots, X_n$ of normally distributed random variables with mean μ. Suppose that $n = 21$.

- What is the value of c for which

$$P\left(\left|\frac{\bar{X} - \mu}{S}\right| \leq c\right) = 0.95?$$

- What is the value of c for which

$$P\left(\left|\frac{\bar{X} - \mu}{\sigma}\right| \leq c\right) = 0.95?$$

Sampling distribution of a sample proportion

Sampling distribution of a sample proportion

- Let $X \sim B(n, p)$, the sample proportion $\hat{p} = X/n$ follows a normal distribution

$$\hat{p} \sim N\big(p,\ p(1-p)/n\big)$$

  - $se(\hat{p}) = \sqrt{p(1-p)/n}$
  - The corresponding Z statistic

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$$

Sampling distribution of a sample proportion

- A coin that is suspected of being biased is tossed many times in order to investigate the possible bias. Consider the following two scenarios:
  - Scenario I

    > The coin is tossed 100 times and 40 heads are obtained

  - Scenario II

    > The coin is tossed 1000 times and 400 heads are obtained

- What is the difference, if any, between the interpretations of these two sets of experimental results?

Sampling distribution of a sample proportion

- Point estimate $\hat{p} = .40$ is the same for both scenarios, but standard errors are different

- Scenario I

$$se(\hat{p}) = \sqrt{.4(1-.4)/100} = 0.0024$$

- Scenario II

$$se(\hat{p}) = \sqrt{.4(1-.4)/1000} = 0.0002$$

Exercise 7.3.21

- Unknown to an experimenter, when a coin is tossed there is a probability of $p = 0.63$ of obtaining a head.
- The experimenter tosses the coin 300 times in order to estimate the probability p.
- What is the probability that the experimenter's point estimate of p will be within the interval $(0.62, 0.64)$?

Exercise 7.3.8

- In a consumer survey, 234 people out of a representative sample of 450 people say that they prefer product A to product B.
- Let p be the proportion of all consumers who prefer product A to product B.
    - Construct a point estimate of p.
    - What is the standard error of your point estimate?

Exercise 7.3.20

- The capacitances of certain electronic components have a normal distribution with parameters $\mu = 174$ and $\sigma = 2.8$.
- If an engineer randomly selects a sample of $n = 30$ components and measures their capacitances
- What is the probability that the engineer's point estimate of the mean μ will be within the interval $(173, 175)$?

Exercise 7.3.23

- A scientist reports that the proportion of defective items from a process is 12.6%.
- If the scientist’s estimate is based on the examination of a random sample of 360 items from the process
- What is the standard error of the scientist's estimate?

Exercise 7.3.34

- Consider the data set with five observations

$$\{7, 9, 14, 15, 22\}$$

    - Obtain the standard error of the sample mean.

Exercise 7.7.18

- The probability that a medical treatment is effective is 0.68, unknown to a researcher.
- In an experiment to investigate the effectiveness of the treatment, the researcher applies the treatment in 140 cases and measures whether the treatment is effective or not.
- What is the probability that the researcher's estimate of the probability that the medical treatment is effective is within 0.05 of the correct answer?

Confidence Intervals

Confidence Interval Construction

- A confidence interval for a parameter θ is an interval that contains a set of plausible values of the parameter θ
- Confidence interval is associated with a confidence level, denoted by $(1-\alpha)$, which measures the probability that the confidence interval actually contains the parameter value
- Often confidence intervals are obtained for 90% or 95% confidence levels
  - $1-\alpha = 0.95 \;\Rightarrow\; \alpha = 0.05$
  - $1-\alpha = 0.90 \;\Rightarrow\; \alpha = 0.10$

Confidence Interval for a Poplation Mean

- Confidence interval for population mean μ can be obtained based on a t-statistic if
  - population variance σ^2 is unknown
  - sample size n is large
  - for a small n, data must be drawn from a normal population

Confidence Interval for a Poplation Mean

- The t-statistic

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

- Confidence interval of μ can also be obtained using Z-statistic when σ is known

Confidence Interval for a Poplation Mean

- For $0 < \alpha < 0.5$, $(1 - \alpha)$ quantile of $Z \sim N(0, 1)$ is denoted by z_α

$$P(Z \leq z_\alpha) = 1 - \alpha \;\Rightarrow\; \Phi(z_\alpha) = 1 - \alpha \;\Rightarrow\; z_\alpha = \Phi^{-1}(1 - \alpha)$$

- It can be shown

$$\begin{aligned} P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) &= \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) \\ &= (1 - \alpha/2) - \alpha/2 = 1 - \alpha \end{aligned}$$

- Similarly, it can be shown for t-distribution

$$P(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 1 - \alpha$$

Confidence Interval for a Poplation Mean

- The $(1-\alpha)$ level confidence interval for μ

$$P\Big(-t_{\alpha/2} < t < t_{\alpha/2}\Big) = 1-\alpha$$

$$P\Big(-t_{\alpha/2} < \frac{\bar{X}-\mu}{s/\sqrt{n}} < t_{\alpha/2}\Big) = 1-\alpha \quad \Rightarrow \quad P\big(L < \mu < U\big) = 1-\alpha$$

  - $L = \bar{X} - (s/\sqrt{n})t_{n-1,\alpha/2}$
  - $U = \bar{X} + (s/\sqrt{n})t_{n-1,\alpha/2}$

Confidence Interval for a Poplation Mean μ

- Let $x_1, x_2, \ldots, x_n$ be a random sample from a population with mean μ and variance σ^2

- The $(1-\alpha)100\%$ confidence interval for μ

$$\bar{x} \pm (s/\sqrt{n})\, t_{\alpha/2} = \left(\bar{x} - (s/\sqrt{n})\, t_{\alpha/2},\ \bar{x} + (s/\sqrt{n})\, t_{\alpha/2}\right)$$

- $\hat{\mu} = \bar{x} = \sum x/n$
- $s = \sqrt{\sum (x_i - \bar{x})^2/(n-1)}$
- $s/\sqrt{n} = se(\hat{\mu})$

- $t_{\alpha/2} \rightarrow$ critical value
- e.g., $n = 20$ and $\alpha = 0.05$
    - $t_{19,.025} = 2.093$

General expression of a confidence Interval

- Confidence interval of a parameter θ can be expressed in terms of its point estimate, standard error, and critical point (which depends on the confidence level)

$$\text{point estimate} \pm \text{margin of error} \Rightarrow \hat{\theta} \pm (\text{critical value})\, se(\hat{\theta})$$

$$\hat{\mu} \pm t_{\alpha/2}\,(s/\sqrt{n}) \;\; and \;\; \hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

  - t- and z-value are considered as critical value in most cases

- Length of a confidence interval

$$L = 2t_{\alpha/2}\,(s/\sqrt{n}) = 2 \times (\text{critical value})(se)$$

Example 17 (Milk container contents)

A sample of 50 milk container provides estimates $\bar{x} = 2.0727$ *and* $s = .0711$*. Obtain a 95% confidence interval for population mean milk contents* μ

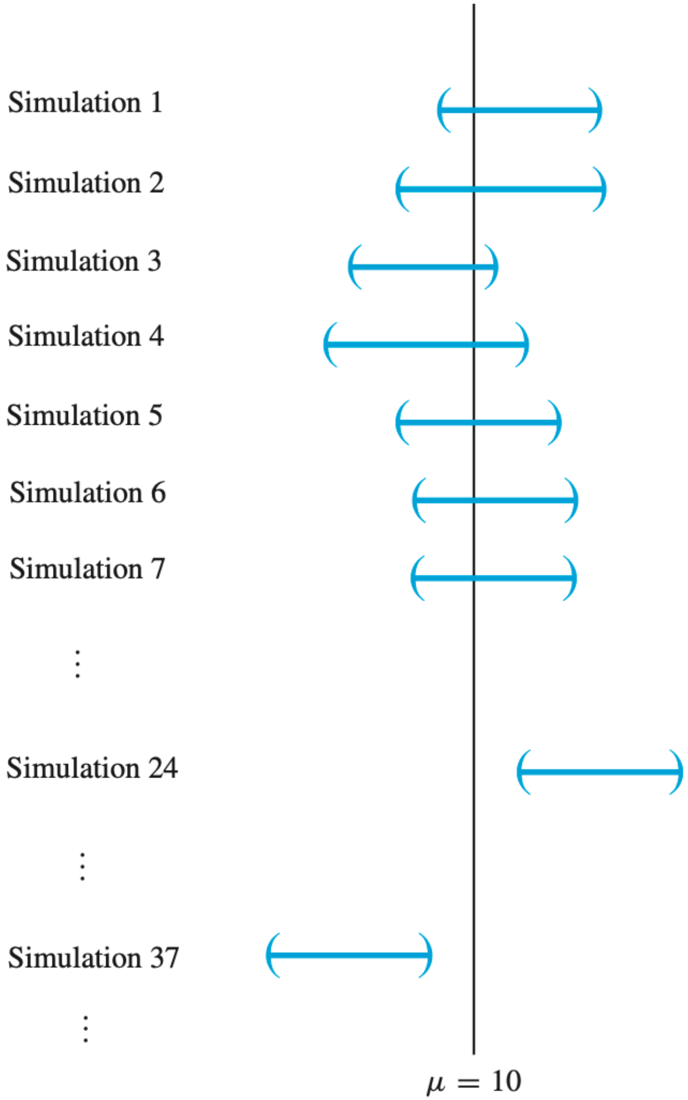
- The 95% confidence interval for μ

$$
\begin{aligned}
\bar{x} \pm (s/\sqrt{n})\, t_{\alpha/2} &= 2.0727 \pm (.0711/\sqrt{50})\, t_{.025}, \quad where \;\; t_{.025} \simeq 2.021 \\
&= 2.0727 \pm (0.1006)\,(2.021) \\
&= (2.0524, 2.093)
\end{aligned}
$$

Interpretation: The experimenter is 95% confident that the interval $(2.0524, 2.093)$ contains the true population mean milk contents μ

FIGURE 8.7

Confidence intervals from simulation experiment



Example 17 (Milk container contents)

- *A sample of 50 milk container provides estimates*

$$\bar{x} = 2.0727 \ \ and \ \ s = .0711$$

- *Obtain a 99% confidence interval for population mean milk contents μ.*
- *Compare it with the 95% confidence interval.*

Effect of the Confidence Level on the Confidence Interval Length

- The length of a confidence interval depends upon the confidence level $(1-\alpha)$ through the critical point

- As the confidence level $(1-\alpha)$ increases $\implies$ the length of the confidence interval also increases because critical value $(t_{\alpha/2})$ increases with the confidence level $(1-\alpha)$

- For a fixed confidence level $(1-\alpha)$

    - As sample size n increases $\implies$ the width of confidence interval decreases

Problem 8.1.1

- A sample of 31 data observations has a sample mean $\bar{x} = 53.42$ and a sample standard deviation $s = 3.05$.
- Construct a 95% confidence interval for the population mean.

Problem 8.1.2

- A random sample of 41 glass sheets is obtained and their thicknesses are measured.
- The sample mean is $\bar{x} = 3.04$ mm and the sample standard deviation is $s = 0.124$ mm.
- Construct a 99% confidence interval for the mean glass thickness.
- Do you think it is plausible that the mean glass thickness is 2.90 mm?

Problem 8.1.4

- A random sample of 16 one-kilogram sugar packets is obtained and the actual weights of the packets are measured.
- The sample mean is $\bar{x} = 1.053$ kg and the sample standard deviation is $s = 0.058$ kg.
- Construct a 99% two-sided t-interval for the average sugar packet weight.
- Do you think it is plausible that the average weight is 1.025 kg?

Confidence interval for a population proportion

- Let X be the number of successes in n binary trials, i.e., $X \sim B(n, p)$, where p is the probability of success and $\hat{p} = X/n$ is the point estimate

- The sampling distribution of $\hat{p}$

$$\hat{p} \sim N\big(p, p(1-p)/n\big) \quad \Rightarrow \quad Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

- The $(1-\alpha)100\%$ confidence interval for p

$$\hat{p} \pm z_{\alpha/2}\, se(\hat{p}) = \hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}$$

Exercise

- Suppose that $x = 21$ is an observation from a $B(27, p)$ random variable.
- Compute a 95% confidence interval for p.

Exercise

- In trials of a medical screening test for a particular illness, 23 cases out of 324 positive results turned out to be false-positive results.
- Construct a 99% confidence bound on p.