

CA MAS-I Chapter 2

2.1.0 Overview

 5m

In this section, we shift gears from theoretical probability models to statistical ideas that incorporate observable data.

Distribution parameters are important because they are quantifiable attributes or properties of a population. But since full information on a population is usually impossible to obtain, the true parameter values are often unknown. In the previous section, we simply assumed the values of parameters. Instead of making arbitrary guesses, we can estimate the parameter values by using observed data from a sample of the population.

Parameter estimation can be accomplished with point estimation and/or interval estimation. We will focus on **point estimation** first; the objective is to obtain a single value as our "best guess" for a parameter. There are many ways to estimate parameters. In this subsection, we will cover three methods:

1. Method of moments
2. Percentile matching
3. Maximum likelihood estimation (MLE)

As parameters dictate the shape of a distribution, we may wish to directly estimate the shape of a distribution instead. Thus, we will cover kernel density estimation as well.

2.1.1 Method of Moments

The sample data should ideally be similar to the presumed distribution. The *method of moments* applies this idea to the moments of a distribution.

Let x_i be the i^{th} observed value from the sample of size n . The k^{th} sample moment is

$$\frac{\sum_{i=1}^n x_i^k}{n}$$

If the distribution of X has r parameters that require estimation, then their estimates are the values that satisfy the following set of equations:

$$\mathbb{E}[X^k] = \frac{\sum_{i=1}^n x_i^k}{n}, \quad k = 1, 2, \dots, r \quad (2.1.1.1)$$

In short, this method determines the parameter values that cause the theoretical moments to equal the sample moments.

Coach's Remarks

Equation 2.1.1.1 calls for the default approach of matching the **first r** moments. However, if a question instructs to match a different set of r moments, simply follow as instructed.

Example 2.1.1.1

The following data records the test scores of 33 students:

Test Score	Number of Students
60	3
70	9
80	12
90	7
100	2

Suppose that test scores follow an exponential distribution with parameter θ .

Estimate θ using the method of moments.

Solution

Since there is only one parameter to estimate, only one equation is needed:

$$\mathbb{E}[X] = \frac{\sum_{i=1}^n x_i}{n}$$

From the exam table, θ refers to the mean of the exponential distribution, i.e. $\mathbb{E}[X] = \theta$.

From the data,

$$\begin{aligned}\frac{\sum_{i=1}^{33} x_i}{33} &= \frac{60 + 60 + 60 + 70 + \dots + 90 + 100 + 100}{33} \\ &= \frac{3(60) + 9(70) + 12(80) + 7(90) + 2(100)}{33} \\ &= 78.7879\end{aligned}$$

Thus, the estimate of θ using the method of moments is

$$\hat{\theta} = \mathbf{78.788}$$



Coach's Remarks

The "hat" symbol will denote a generic use of point estimation. Hence, $\hat{\theta}$ here translates to "estimate of θ ".

Example 2.1.1.2

The following data records the test scores of 33 students:

Test Score	Number of Students
60	3
70	9
80	12
90	7
100	2

Suppose that test scores follow a gamma distribution with parameters α and θ .

Estimate α and θ using the method of moments.

Solution

With two parameters to estimate, two equations are needed:

$$\begin{aligned} E[X] &= \frac{\sum_{i=1}^n x_i}{n} \\ E[X^2] &= \frac{\sum_{i=1}^n x_i^2}{n} \end{aligned}$$

Determine from the exam table that

$$\begin{aligned} E[X] &= \alpha\theta \\ E[X^2] &= (\alpha + 1)\alpha\theta^2 \end{aligned}$$

From the data,

$$\frac{\sum_{i=1}^{33} x_i}{33} = 78.7879$$

$$\begin{aligned}\frac{\sum_{i=1}^{33} x_i^2}{33} &= \frac{3(60^2) + 9(70^2) + 12(80^2) + 7(90^2) + 2(100^2)}{33} \\ &= 6,315.1515\end{aligned}$$

Thus, the two equations are

$$\begin{aligned}\alpha\theta &= 78.7879 \\ (\alpha + 1)\alpha\theta^2 &= 6,315.1515\end{aligned}$$

Solve for α and θ . First, rewrite the second equation in terms of $\alpha\theta$.

$$\begin{aligned}(\alpha + 1)\alpha\theta^2 &= (\alpha\theta)^2 + (\alpha\theta)\theta \\ 6,315.1515 &= 78.7879^2 + 78.7879\theta \\ \theta &= \frac{6,315.1515 - 78.7879^2}{78.7879} \\ &= 1.366\end{aligned}$$

$$\begin{aligned}\alpha &= \frac{78.7879}{1.366} \\ &= 57.679\end{aligned}$$

The estimates are the parameter values that satisfy the equations. Thus,

$$\hat{\alpha} = \mathbf{57.679}, \quad \hat{\theta} = \mathbf{1.366}$$



Example 2.1.1.3

You observe the following loss amounts for an insurance product with a policy limit of 35:

8 13 20 22 27 30 34 35 35 35

Suppose that ground-up losses follow a Pareto distribution with parameters $\alpha = 2$ and θ .

Estimate θ using the method of moments.

Solution

If X follows the Pareto distribution, realize that the available data corresponds to $(X \wedge 35)$. This is because the policy limit of 35 causes the data to be capped at 35. As a result, the equation needed to estimate θ is:

$$\mathbb{E}[X \wedge 35] = \frac{\sum_{i=1}^n x_i}{n}$$

Determine from the exam table that

$$\begin{aligned}\mathbb{E}[X \wedge 35] &= \frac{\theta}{\alpha - 1} \left[1 - \left(\frac{\theta}{35 + \theta} \right)^{\alpha-1} \right] \\ &= \frac{\theta}{2 - 1} \left[1 - \left(\frac{\theta}{35 + \theta} \right)^{2-1} \right] \\ &= \theta \left[\frac{35 + \theta - \theta}{35 + \theta} \right] \\ &= \frac{35\theta}{35 + \theta}\end{aligned}$$

From the data,

$$\frac{\sum_{i=1}^{10} x_i}{10} = \frac{8 + 13 + \dots + 35}{10} = 25.9$$

Equating the two and solving for θ produces the estimate

$$\begin{aligned}\frac{35\theta}{35 + \theta} &= 25.9 \\ 35\theta &= 25.9(35) + 25.9\theta \\ \theta(35 - 25.9) &= 25.9(35) \\ \theta &= \frac{25.9(35)}{35 - 25.9}\end{aligned}$$

$$\hat{\theta} = \mathbf{99.615}$$



2.1.2 Percentile Matching

(L) 30m

Aside from moments, there are other characteristics that should ideally align between the sample data and distribution. The **percentile matching** method focuses on using percentiles to obtain parameter estimates.

As a reminder, we denote π_q as the $100q^{\text{th}}$ percentile of random variable X , i.e.

$$\Pr(X \leq \pi_q) = F(\pi_q) = q$$

In addition, denote $\hat{\pi}_q$ as the $100q^{\text{th}}$ percentile of the sample.

Basic Principle of Percentile Matching

If the distribution of X has r parameters that require estimation, then their estimates are the values that satisfy the following set of equations:

$$\pi_{q_k} = \hat{\pi}_{q_k}, \quad k = 1, 2, \dots, r \quad (2.1.2.1)$$

where the q_k 's are arbitrarily chosen probabilities. Problems on the exam will specify which percentiles should be matched.

Assume the median of a sample is 250. Suppose the data came from an exponential distribution.

Estimate θ by matching the medians.

Recall that the median is another name for the 50^{th} percentile. Therefore, we need to match

$$\pi_{0.5} = \hat{\pi}_{0.5} = 250$$

The exam table gives the formula for the CDF of an exponential distribution. This leads to

$$F(\pi_{0.5}) = 1 - e^{-\pi_{0.5} / \theta} = 0.5$$

Evaluate $\pi_{0.5} = 250$ and solve for θ to calculate its estimate.

$$\begin{aligned} 1 - e^{-250 / \theta} &= 0.5 \\ e^{-250 / \theta} &= 1 - 0.5 \\ -\frac{250}{\theta} &= \ln(1 - 0.5) \end{aligned}$$

$$\hat{\theta} = -\frac{250}{\ln(1 - 0.5)} = \mathbf{360.674}$$

In the example above, we were conveniently given that the sample median was 250. In reality, there are different approaches to computing sample percentiles, each potentially producing a different answer. This exam uses the *smoothed empirical percentile* approach.

Smoothed Empirical Percentile — Unique Values

Rather than memorizing a formula to calculate $\hat{\pi}_q$, we encourage you to follow these steps:

1. Let $x_{(i)}$ denote the i^{th} observed value in **ascending order**, e.g. $x_{(10)}$ is the 10^{th} smallest sample data point. Also, define $b = \lfloor q(n+1) \rfloor$, i.e. round $q(n+1)$ **down** to the nearest integer.
2. Calculate $q(n+1)$.
 - For example, if $q = 0.65$ and $n = 34$, then $q(n+1) = 22.75$. You may interpret this to loosely mean "the 65^{th} sample percentile is the 22.75^{th} observed value in ascending order".
3. If $q(n+1)$ is a non-integer, calculate $\hat{\pi}_q$ by linearly interpolating between $x_{(b)}$ and $x_{(b+1)}$.
 - Since 22.75 is between 22 and 23, we need to interpolate between $x_{(22)}$ and $x_{(23)}$, the 22^{nd} and 23^{rd} observed values in ascending order.

Take the numbers after the decimal of $q(n+1)$ as the weight that is multiplied to the larger value $x_{(b+1)}$. Then, the smaller value $x_{(b)}$ gets the complement weight. In other

words, $\hat{\pi}_{0.65} = 0.25x_{(22)} + 0.75x_{(23)}$, where **0.75** is taken from **22.75**, and $0.25 = 1 - \textcolor{red}{0.75}$.

4. If $q(n+1)$ is an integer, then $\hat{\pi}_q = x_{(b)}$.

- This is consistent with the interpolation technique above; integers have 0's after the decimal, so the larger value $x_{(b+1)}$ will receive a weight of 0.

There are a couple of caveats worth mentioning:

- $q(n+1)$ must be between 1 and n , inclusive. For this exam, we may take this for granted; otherwise, $\hat{\pi}_q$ is undefined.
- The interpolation technique in Step 3 assumes that all observed values from the sample are unique, i.e. there are no repeated values. We will address how to handle repeated values shortly.

Let's see the steps in action with a few examples.

Write the expression that calculates the 34th percentile of a sample of size 7.

$$q(n+1) = 0.34(7+1) = 2.72$$

2.72 tells us to interpolate between the 2nd and 3rd observed values in ascending order, with respective weights $1 - 0.72 = 0.28$ and 0.72 . Thus,

$$\hat{\pi}_{0.34} = \mathbf{0.28}x_{(2)} + \mathbf{0.72}x_{(3)}$$

Write the expression that calculates the 84th percentile of a sample of size 17.

$$q(n+1) = 0.84(17+1) = 15.12$$

15.12 tells us to interpolate between the 15th and 16th observed values in ascending order, with respective weights $1 - 0.12 = 0.88$ and 0.12 . Thus,

$$\hat{\pi}_{0.84} = \mathbf{0.88}x_{(15)} + \mathbf{0.12}x_{(16)}$$

Write the expression that calculates the 40th percentile of a sample of size 19.

$$q(n+1) = 0.4(19+1) = 8$$

Since 8 is an integer, no interpolation is necessary.

$$\hat{\pi}_{0.4} = x_{(8)}$$

Now, let's apply this concept to a proper example.

Example 2.1.2.1

A soccer fan records the time it takes his favorite team to score one goal in 16 random matches.

15	35	60	85	33	69	88	44
36	78	90	32	2	68	23	19

Assume that the scoring times follow a Weibull distribution.

Estimate θ and τ by matching the 40th and 60th percentiles.

Solution

First, organize the data in ascending order.

2	15	19	23	32	33	35	36
44	60	68	69	78	85	88	90

Next, calculate the 40th and 60th sample percentiles.

For the 40th sample percentile,

$$q(n+1) = 0.4(16+1) = 6.8$$

$$\begin{aligned}\hat{\pi}_{0.4} &= 0.2x_{(6)} + 0.8x_{(7)} \\ &= 0.2(33) + 0.8(35) \\ &= 34.6\end{aligned}$$

For the 60th sample percentile,

$$q(n+1) = 0.6(16+1) = 10.2$$

$$\begin{aligned}\hat{\pi}_{0.6} &= 0.8x_{(10)} + 0.2x_{(11)} \\ &= 0.8(60) + 0.2(68) \\ &= 61.6\end{aligned}$$

Therefore, we need to match

$$\begin{aligned}\pi_{0.4} &= \hat{\pi}_{0.4} = 34.6 \\ \pi_{0.6} &= \hat{\pi}_{0.6} = 61.6\end{aligned}$$

With the Weibull CDF from the exam table, the equations are

$$\begin{aligned}1 - e^{-(34.6/\theta)^7} &= 0.4 \\ 1 - e^{-(61.6/\theta)^7} &= 0.6\end{aligned}$$

Next, simplify each equation to arrive at

$$\begin{aligned}-\left(\frac{34.6}{\theta}\right)^\tau &= \ln 0.6 \\-\left(\frac{61.6}{\theta}\right)^\tau &= \ln 0.4\end{aligned}$$

We can first solve for τ by dividing the two equations. As a result,

$$\begin{aligned}-\left(\frac{34.6}{\theta}\right)^\tau / -\left(\frac{61.6}{\theta}\right)^\tau &= \ln 0.6 / \ln 0.4 \\ \left(\frac{34.6}{61.6}\right)^\tau &= 0.5575 \\ \tau \ln\left(\frac{34.6}{61.6}\right) &= \ln 0.5575 \\ \tau &= \ln 0.5575 / \ln\left(\frac{34.6}{61.6}\right) \\ &= 1.013\end{aligned}$$

$$\begin{aligned}-\left(\frac{34.6}{\theta}\right)^{1.013} &= \ln 0.6 \\ \theta^{1.013} &= -\frac{34.6^{1.013}}{\ln 0.6} \\ \theta &= \left(-\frac{34.6^{1.013}}{\ln 0.6}\right)^{1.013^{-1}} \\ &= 67.152\end{aligned}$$

$$\hat{\theta} = \mathbf{67.152}, \quad \hat{\tau} = \mathbf{1.013}$$



An alternative to the interpolation procedure in Step 3 is to sketch a number line. Let's recalculate the 40th sample percentile of Example 2.1.2.1. In this case, $q(n+1) = 6.8$.



We assign **more weight** to $x_{(7)}$ because 6.8 is **closer** to 7, and we assign **less weight** to $x_{(6)}$ because 6.8 is **farther** from 6. Therefore, $x_{(7)}$ receives 0.8 out of the 1 unit, while $x_{(6)}$ receives 0.2 out of the 1 unit. We obtain the same result as before, i.e.

$$\hat{\pi}_{0.4} = 0.2x_{(6)} + 0.8x_{(7)}$$

Smoothed Empirical Percentile — Repeated Values

If any of the observed values from the sample are repeated, then our process of computing $\hat{\pi}_q$ needs a minor adjustment. The index i should be updated such that every unique observed value corresponds to only one i – the **largest** one. To demonstrate, consider the following sample data:

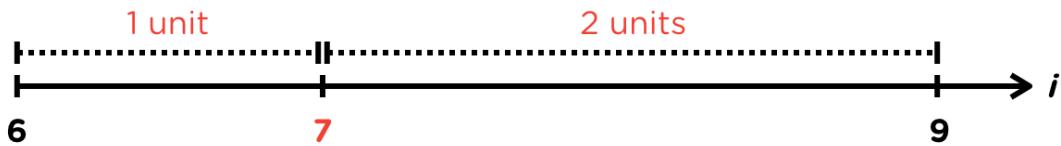
2 4 4 5 7 9 11 11 11

Note that $x_{(2)} = x_{(3)} = 4$ and $x_{(7)} = x_{(8)} = x_{(9)} = 11$. Even though there are 9 observed values, only 6 are unique. Thus, the index i is updated to have the following 6 numbers:

$$i = 1, 3, 4, 5, 6, 9$$

In other words, 2, 7, and 8 are dropped from the index; we keep the indices 3 and 9 because they are the largest ones for the repeated values.

Consequently, linear interpolation occurs between the $x_{(i)}$'s given the updated i . As practice, let's solve for $\hat{\pi}_{0.7}$. First, calculate $q(n+1) = 0.7(9+1) = 7$. Since 7 is not listed in the updated i , we linearly interpolate between the observed values whose updated i 's enclose 7, i.e. interpolate between $x_{(6)}$ and $x_{(9)}$. It is easiest to determine the appropriate weights by sketching a number line.



We assign **more weight** to $x_{(6)}$ because 7 is **closer** to 6, and we assign **less weight** to $x_{(9)}$ because 9 is **farther** from 7. Therefore, $x_{(6)}$ receives 2 out of the 3 units, while $x_{(9)}$ receives 1 out of the 3 units. Therefore,

$$\begin{aligned}\hat{\pi}_{0.7} &= \frac{2}{3}x_{(6)} + \frac{1}{3}x_{(9)} \\ &= \frac{2}{3}(9) + \frac{1}{3}(11) \\ &= 9.67\end{aligned}$$

In contrast, failing to adjust for repeated values in this case would lead to the erroneous conclusion that $\hat{\pi}_{0.7} = x_{(7)} = 11$.

Example 2.1.2.2

In a study with 12 policyholders, the times until the next claim were recorded as follows:

9	3	2	7	3	1
1	9	7	8	9	1

Assume these times follow a Pareto distribution.

Estimate α and θ by matching the 40th and 64th percentiles.

Solution

First, organize the data in ascending order.

1	1	1	2	3	3
7	7	8	9	9	9

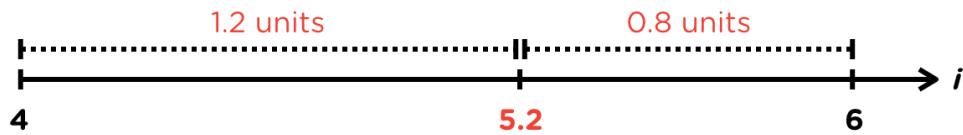
Since there are repeated values, the index is adjusted to

$$i = 3, 4, 6, 8, 9, 12$$

Next, calculate the 40th and 64th sample percentiles.

For the 40th sample percentile,

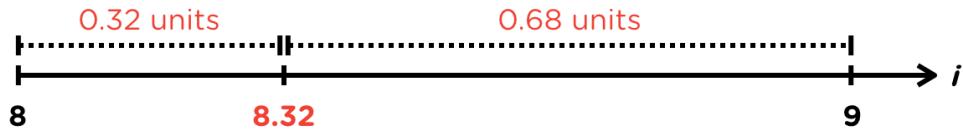
$$q(n+1) = 0.4(12+1) = 5.2$$



$$\begin{aligned}\hat{\pi}_{0.4} &= \frac{0.8}{2}x_{(4)} + \frac{1.2}{2}x_{(6)} \\ &= 0.4(2) + 0.6(3) \\ &= 2.6\end{aligned}$$

For the 64th sample percentile,

$$q(n+1) = 0.64(12+1) = 8.32$$



$$\begin{aligned}\hat{\pi}_{0.64} &= 0.68x_{(8)} + 0.32x_{(9)} \\ &= 0.68(7) + 0.32(8) \\ &= 7.32\end{aligned}$$

With the Pareto CDF from the exam table, the matched equations are

$$1 - \left(\frac{\theta}{2.6 + \theta} \right)^\alpha = 0.4$$

$$1 - \left(\frac{\theta}{7.32 + \theta} \right)^\alpha = 0.64$$

Next, simplify each equation to arrive at

$$\alpha \ln \left(\frac{\theta}{2.6 + \theta} \right) = \ln 0.6$$

$$\alpha \ln \left(\frac{\theta}{7.32 + \theta} \right) = \ln 0.36$$

We do this to obtain two equations that, once combined, cause one of the parameters to drop. Now, we can divide the two equations to drop α and solve for θ first.

$$\alpha \ln \left(\frac{\theta}{2.6 + \theta} \right) / \alpha \ln \left(\frac{\theta}{7.32 + \theta} \right) = \ln 0.6 / \ln 0.36$$

$$\ln \left(\frac{\theta}{2.6 + \theta} \right) = 0.5 \ln \left(\frac{\theta}{7.32 + \theta} \right)$$

$$\left(\frac{\theta}{2.6 + \theta} \right)^2 = \frac{\theta}{7.32 + \theta}$$

$$\theta (7.32 + \theta) = (2.6 + \theta)^2$$

$$\theta = \frac{2.6^2}{7.32 - 2 \cdot 2.6}$$

$$= 3.189$$

$$\alpha \ln \left(\frac{3.189}{2.6 + 3.189} \right) = \ln 0.6$$

$$\alpha = \ln 0.6 / \ln \left(\frac{3.189}{2.6 + 3.189} \right)$$

$$= 0.857$$

$$\hat{\alpha} = \mathbf{0.857}, \quad \hat{\theta} = \mathbf{3.189}$$



2.1.3 Maximum Likelihood Estimation: Complete Data

Maximum likelihood estimation (MLE) is another approach to point estimation. As the name suggests, MLE finds the parameters that maximize the likelihood of the observations.

Let's start with understanding **likelihood**. Informally, "likelihood" is synonymous with "probability". The proper definition for the likelihood of an observation is a probability function evaluated at the observed value. For example, if a random variable with probability function $f(x)$ is observed to be c , then the likelihood of this observation is $f(c)$.

As a result, for independently drawn observations, the **likelihood function** is the product of each observation's likelihood. Assuming that x_1, x_2, \dots, x_n refer to the observed values, the likelihood function is

$$\prod_{i=1}^n f(x_i)$$

Coach's Remarks

In this subsection, we use $f(x)$ to represent a probability function applicable to either discrete or continuous distributions.

Imagine $f(x)$ involves a generic parameter θ that we desire to estimate. Realize this means that the likelihood function is a function of θ . Thus, we denote the likelihood function of θ as

$$L(\theta) = \prod_{i=1}^n f(x_i) \tag{2.1.3.1}$$

To solidify these concepts, consider this short example.

You observed the values 3, 1, and 2, which you believe to have independently come from a Poisson distribution with mean λ . What is the likelihood function of λ ?

The PMF of a Poisson distribution with mean λ is

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

Therefore, the likelihood function of λ is

$$\begin{aligned} L(\lambda) &= \left(\frac{e^{-\lambda} \lambda^3}{3!} \right) \left(\frac{e^{-\lambda} \lambda^1}{1!} \right) \left(\frac{e^{-\lambda} \lambda^2}{2!} \right) \\ &= \frac{e^{-3\lambda} \lambda^6}{12} \end{aligned}$$

As previously hinted, the estimate is the parameter value that **maximizes** the likelihood function. In many cases, this is a just calculus problem: take the first derivative of the likelihood function, and set it equal to 0; the solution of that equation is the MLE estimate. In math notation, the estimate is the value of θ that solves

$$\frac{d}{d\theta} L(\theta) = L'(\theta) = 0 \quad (2.1.3.2)$$

However, it is often the case that taking the derivative of $L(\theta)$ is tedious. A convenient alternative is to maximize $l(\theta) = \ln [L(\theta)]$, the **log-likelihood function**, whose derivative is likely easier to compute. It is guaranteed that maximizing the natural logarithm of any function will result in the same solution as maximizing the original function. The first derivative of the log-likelihood function, $l'(\theta)$, is called a **score function**.

What is the MLE estimate of λ for the previous example?

First, calculate the log-likelihood function:

$$\begin{aligned} l(\lambda) &= \ln [L(\lambda)] \\ &= \ln \left(\frac{e^{-3\lambda} \lambda^6}{12} \right) \\ &= -3\lambda + 6 \ln \lambda - \ln 12 \end{aligned}$$

To maximize the function, set its first derivative equal to 0. Solving for λ produces an estimate of 2.

$$\begin{aligned} l'(\lambda) &= \frac{d}{d\lambda} l(\lambda) \\ &= -3 + \frac{6}{\lambda} = 0 \end{aligned}$$

$$\hat{\lambda} = \frac{6}{3} = 2$$

To obtain MLE estimates for a setup with **more than one** parameter, first realize that the likelihood function would be a multi-variable function of all those parameters. Therefore, find the score functions with respect to each parameter and set them equal to 0. All of those equations are called the **score equations**; the solution to the score equations are the MLE estimates.

Furthermore, the calculus strategy of "take first derivative; set equal to 0; solve" may not always work. It takes for granted that the likelihood function has a global maximum at a critical point. Regardless, remember that the objective is to maximize the likelihood function. If and when the default strategy fails, it may be possible to obtain an estimate by reasoning things through. We will examine such a situation in Section 2.1.5.

Coach's Remarks

The likelihood function appears to be the same as a joint probability function evaluated at all the observed values. You are free to make that association, but the two functions are inherently different. One key distinction is that a likelihood function is a function of the parameter(s), whereas a joint probability function is a function of the random variable outcomes. The point is that likelihood is **not** inherently the same as probability.

Example 2.1.3.1

You observe 4 claims:

20 50 150 380

The claim amounts follow an exponential distribution with mean θ .

Calculate the maximum likelihood estimate of θ .

Solution

Construct the likelihood function by multiplying exponential PDFs, each evaluated at an observed value.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^4 f(x_i) \\ &= \left(\frac{1}{\theta} e^{-20/\theta} \right) \left(\frac{1}{\theta} e^{-50/\theta} \right) \left(\frac{1}{\theta} e^{-150/\theta} \right) \left(\frac{1}{\theta} e^{-380/\theta} \right) \\ &= \frac{1}{\theta^4} e^{-600/\theta} \end{aligned}$$

Then, the log-likelihood function is

$$\begin{aligned} l(\theta) &= \ln [L(\theta)] \\ &= -4 \ln \theta - \frac{600}{\theta} \end{aligned}$$

Finally, take its first derivative, set equal to 0, then solve for θ .

$$\begin{aligned} l'(\theta) &= \frac{d}{d\theta} l(\theta) \\ &= -\frac{4}{\theta} + \frac{600}{\theta^2} = 0 \end{aligned}$$

$$\begin{aligned} -4\theta + 600 &= 0 \\ \theta &= \frac{600}{4} \end{aligned}$$

$$\hat{\theta} = \mathbf{150}$$

Coach's Remarks

We can maximize the likelihood function directly, but this requires applying the product rule.

$$\begin{aligned}
 \frac{d}{d\theta} L(\theta) &= \frac{d}{d\theta} \left(\frac{1}{\theta^4} e^{-600/\theta} \right) \\
 &= -\frac{4}{\theta^5} e^{-600/\theta} + \frac{1}{\theta^4} e^{-600/\theta} \frac{d}{d\theta} \left(-\frac{600}{\theta} \right) \\
 &= -\frac{4}{\theta^5} e^{-600/\theta} + \frac{1}{\theta^4} e^{-600/\theta} \left(\frac{600}{\theta^2} \right) \\
 &= \frac{1}{\theta^5} e^{-600/\theta} \left(-4 + \frac{600}{\theta} \right)
 \end{aligned}$$

Set the first derivative equal to 0. Then, solve for θ .

$$\begin{aligned}
 \frac{1}{\theta^5} e^{-600/\theta} \left(-4 + \frac{600}{\theta} \right) &= 0 \\
 -4 + \frac{600}{\theta} &= 0 \\
 \hat{\theta} &= \frac{600}{4} \\
 &= \mathbf{150}
 \end{aligned}$$

Example 2.1.3.2

Losses follow a distribution with probability density function:

$$f(x) = \frac{x^3 e^{-x/\theta}}{6\theta^4}, \quad x > 0$$

You observe 5 losses totaling 700.

Calculate the maximum likelihood estimate of θ .

Solution

Construct the likelihood function.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^5 \left(\frac{x_i^3 e^{-x_i/\theta}}{6\theta^4} \right) \\ &= \frac{\left(\prod_{i=1}^5 x_i^3 \right) e^{-\sum_{i=1}^5 x_i/\theta}}{6^5 \theta^{20}} \\ &= c \cdot \theta^{-20} e^{-700/\theta} \end{aligned}$$

where $c = \frac{\prod_{i=1}^5 x_i^3}{6^5}$.

Then, the log-likelihood function is

$$l(\theta) = \ln c - 20 \ln \theta - \frac{700}{\theta}$$

Finally, take its first derivative, set equal to 0, then solve for θ .

$$l'(\theta) = 0 - \frac{20}{\theta} + \frac{700}{\theta^2} = 0$$

$$\begin{aligned} -20\theta + 700 &= 0 \\ \theta &= \frac{700}{20} \end{aligned}$$

$$\hat{\theta} = 35$$



Coach's Remarks

c is a multiplicative constant of the likelihood function, as it does not depend on the parameter θ . Notice how it is irrelevant to the process of finding the MLE estimate. This means we may drop c from $L(\theta)$ before proceeding to the rest of the steps.

A word of caution: this is allowed **only** because dropping multiplicative constants does not affect the MLE of the parameter(s). However, if the actual likelihood or log-likelihood function is needed, these constants cannot be dropped.

An appealing property of MLE is its *invariance property*: for a function $g(\cdot)$, the MLE estimate of $g(\theta)$ equals the function evaluated at the MLE estimate of θ , i.e. $g(\hat{\theta})$. It is particularly helpful for determining the MLE estimate of a distribution's mean or variance, as they are functions of distribution parameters.

Example 2.1.3.3

Suppose that a health actuary is analyzing claim sizes and observes the following data:

Loss Amount	Frequency
1.5	43
2	21
2.5	5
3	2
3.5	1
4	1

She assumes every observation was independently drawn from a single-parameter Pareto distribution with $\theta = 1$.

Estimate the mean claim size using maximum likelihood.

Solution

For a single-parameter Pareto with $\theta = 1$, the mean is

$$\mathbb{E}[X] = \frac{\alpha}{\alpha - 1}$$

Hence, first find the MLE estimate of α . Construct the likelihood function.

$$\begin{aligned} L(\alpha) &= \frac{\alpha^{43}}{1.5^{43(\alpha+1)}} \cdot \frac{\alpha^{21}}{2^{21(\alpha+1)}} \cdot \frac{\alpha^5}{2.5^{5(\alpha+1)}} \cdot \frac{\alpha^2}{3^{2(\alpha+1)}} \cdot \frac{\alpha}{3.5^{\alpha+1}} \cdot \frac{\alpha}{4^{\alpha+1}} \\ &= \frac{\alpha^{73}}{1.5^{43(\alpha+1)} \cdot 2^{21(\alpha+1)} \cdot 2.5^{5(\alpha+1)} \cdot 3^{2(\alpha+1)} \cdot 3.5^{\alpha+1} \cdot 4^{\alpha+1}} \\ &= \frac{1}{\underbrace{1.5^{43} \cdot 2^{21} \cdot \dots \cdot 4}_c} \left(\frac{\alpha^{73}}{1.5^{43\alpha} \cdot 2^{21\alpha} \cdot \dots \cdot 4^\alpha} \right) \\ &\propto \frac{\alpha^{73}}{1.5^{43\alpha} \cdot 2^{21\alpha} \cdot \dots \cdot 4^\alpha} \end{aligned}$$

After dropping the multiplicative constant, take its natural log. Then, calculate its first derivative with respect to α , and set equal to 0.

$$\begin{aligned} \ln \left(\frac{\alpha^{73}}{1.5^{43\alpha} \cdot 2^{21\alpha} \cdot \dots \cdot 4^\alpha} \right) &= 73 \ln \alpha - 43\alpha \ln 1.5 - 21\alpha \ln 2 - 5\alpha \ln 2.5 \\ &\quad - 2\alpha \ln 3 - \alpha \ln 3.5 - \alpha \ln 4 \\ \frac{d}{d\alpha} \ln \left(\frac{\alpha^{73}}{1.5^{43\alpha} \cdot 2^{21\alpha} \cdot \dots \cdot 4^\alpha} \right) &= \frac{73}{\alpha} - 43 \ln 1.5 - 21 \ln 2 - 5 \ln 2.5 \\ &\quad - 2 \ln 3 - \ln 3.5 - \ln 4 = 0 \end{aligned}$$

Upon solving for α , its estimate is

$$73 - \alpha (43 \ln 1.5 + 21 \ln 2 + 5 \ln 2.5 + 2 \ln 3 + \ln 3.5 + \ln 4) = 0$$
$$\alpha = \frac{73}{43 \ln 1.5 + 21 \ln 2 + 5 \ln 2.5 + 2 \ln 3 + \ln 3.5 + \ln 4}$$

$$\hat{\alpha} = 1.763$$

Therefore, the MLE estimate of the mean claim size is

$$\frac{\hat{\alpha}}{\hat{\alpha} - 1} = \mathbf{2.31}$$



2.1.4 Maximum Likelihood Estimation: Incomplete Data

🕒 25m

Thus far, we have only considered complete data, i.e. the exact values of a complete dataset are all known. However, we may be given incomplete information instead, such as not knowing the exact values of the observations or only having a partial dataset. Examples of these include truncation, censoring, and grouping.

Truncated Data

Truncated data is an example of an incomplete dataset.

A dataset that is **left-truncated** at d means the dataset does not include data **at or below d** . Thus, the likelihood of such an observation must be the probability function conditioned on being above d , i.e.

$$\frac{f(x)}{\Pr(X > d)}$$

Left truncation is commonly seen in insurance in the form of a **deductible**. When the loss amount is below the deductible, the policyholder will not receive any reimbursement for it, so the loss likely will not be reported. Thus, an insurer will only record a loss when it exceeds the deductible.

Example 2.1.4.1

You are given a sample of losses:

6 7 9 10

No information is available for losses of 5 or less. Assume losses follow an exponential distribution with mean θ .

Determine the maximum likelihood estimate of θ .

Solution

The sample is left-truncated at 5. Thus, the likelihood of each loss is

$$\frac{f(x)}{S(5)}$$

Construct the likelihood function.

$$\begin{aligned} L(\theta) &= \frac{f(6)f(7)f(9)f(10)}{S(5)^4} \\ &= \frac{\frac{1}{\theta^4}e^{-(6+7+9+10)/\theta}}{(e^{-5/\theta})^4} \\ &= \frac{1}{\theta^4}e^{-12/\theta} \end{aligned}$$

Then, calculate the log-likelihood function.

$$l(\theta) = -4 \ln \theta - \frac{12}{\theta}$$

Take its first derivative, set equal to 0, and solve for θ .

$$l'(\theta) = -\frac{4}{\theta} + \frac{12}{\theta^2} = 0$$

$$\begin{aligned} -4\theta + 12 &= 0 \\ \theta &= \frac{12}{4} \end{aligned}$$

$$\hat{\theta} = 3$$

Censored Data

Censored data is an example of observations for which the exact values are not known.

An observation is **right-censored** at m when the value is known to be **at least** m but is only recorded as m . The likelihood of the observation is

$$\Pr(X \geq m)$$

Right censoring is commonly seen in insurance whenever there is a **policy limit**. A policyholder can claim at most the policy limit. Thus, for losses that exceed the maximum covered loss, records will likely only document them as the maximum covered loss. So while the exact amounts of the losses are not known, it is known that those losses are at least the maximum covered loss.

Example 2.1.4.2

For insurance coverage with a policy limit of 10:

- You observed the following claim payments:

2 3 5 8 10 10

- Claim sizes follow a Weibull distribution with $\tau = 2$.

Calculate the maximum likelihood estimate of the 90th percentile of the claim sizes.

Solution

Let $\pi_{0.9}$ be the 90th percentile of a Weibull distribution with $\tau = 2$. Thus,

$$\begin{aligned} 1 - e^{-(\pi_{0.9}/\theta)^2} &= 0.9 \\ -\left(\frac{\pi_{0.9}}{\theta}\right)^2 &= \ln(1 - 0.9) \\ \pi_{0.9} &= \theta \sqrt{-\ln 0.1} \end{aligned}$$

By the invariance property, we can solve for the goal by first computing the MLE estimate of θ , then multiplying it by $\sqrt{-\ln 0.1}$.

Next, note that there are two payments at the policy limit of 10. These payments can be the result of losses at or above 10. Thus, the likelihood for each of those payments is

$$\Pr(X \geq 10) = S(10)$$

since X is continuous. Proceed with constructing the likelihood function. The likelihoods of the first four data points are found using the PDF since the exact values are known. As a result,

$$\begin{aligned} L(\theta) &= f(2) \cdot f(3) \cdot f(5) \cdot f(8) \cdot S(10)^2 \\ &= \frac{2(2/\theta)^2 e^{-(2/\theta)^2}}{2} \cdot \dots \cdot \frac{2(8/\theta)^2 e^{-(8/\theta)^2}}{8} \cdot [e^{-(10/\theta)^2}]^2 \\ &= 2^4 (2 \cdot 3 \cdot 5 \cdot 8) \cdot \frac{1}{\theta^8} e^{-[2^2+3^2+5^2+8^2+2(10^2)]/\theta^2} \\ &\propto \frac{e^{-302/\theta^2}}{\theta^8} \end{aligned}$$

After dropping the multiplicative constant, take its natural log. Then, calculate its first derivative with respect to θ , and set equal to 0.

$$\begin{aligned} \ln \left(\frac{e^{-302/\theta^2}}{\theta^8} \right) &= -\frac{302}{\theta^2} - 8 \ln \theta \\ \frac{d}{d\theta} \ln \left(\frac{e^{-302/\theta^2}}{\theta^8} \right) &= \frac{604}{\theta^3} - \frac{8}{\theta} = 0 \end{aligned}$$

Solving for θ produces

$$604 - 8\theta^2 = 0$$

$$\theta = \sqrt{\frac{604}{8}}$$

$$\hat{\theta} = 8.689$$

Therefore, the MLE estimate of the 90th percentile is

$$\hat{\theta}\sqrt{-\ln 0.1} = 8.689\sqrt{-\ln 0.1} = \mathbf{13.185}$$



A dataset could have both truncated and censored data. Simply use the correct likelihood for each observation to determine the likelihood function.

Example 2.1.4.3

You are given the following information about a group of policies:

Claim Payment	Deductible	Policy Limit
30	-	80
50	10	100
80	10	100
120	20	150
150	30	150

Assume payments at the policy limit resulted from losses above the maximum covered loss.

You are to fit a continuous distribution to the losses using the maximum likelihood method.

Determine the likelihood function.

Solution

Notice that we are given the **claim payments** and are asked for the likelihood function for the **loss** distribution. To translate payments into losses, simply add back the deductible. A similar adjustment is required to translate policy limits into maximum covered losses.

Then, calculate the likelihoods by evaluating the PDF or survival function at the appropriate loss amount and condition it on being greater than the deductible, if any.

Loss Amount	Deductible	Maximum Covered Loss	Likelihood
30	-	80	$f(30)$
60	10	110	$\frac{f(60)}{S(10)}$
90	10	110	$\frac{f(90)}{S(10)}$
140	20	170	$\frac{f(140)}{S(20)}$
> 180	30	180	$\frac{S(180)}{S(30)}$

The likelihood function is the product of the individual likelihoods.

$$L = f(30) \cdot \frac{f(60)}{S(10)} \cdot \frac{f(90)}{S(10)} \cdot \frac{f(140)}{S(20)} \cdot \frac{S(180)}{S(30)}$$



Grouped Data

Similar to censored data, grouped data is an example of observations for which the exact values are not known. Grouped data is presented as the number of observations in a distinct interval. The likelihood of an observation in the interval $(a, b]$ is

$$\begin{aligned} \Pr(a < X \leq b) &= F(b) - F(a) \\ &= S(a) - S(b) \end{aligned}$$

For discrete distributions, the likelihood of grouped data is more easily expressed as the sum of PMFs evaluated at every value within the range, e.g.

$$\Pr(a \leq X \leq b) = p(a) + \dots + p(b)$$

Example 2.1.4.4

The annual number of accidents per car follows a geometric distribution with parameter β where $\beta > 0$.

A sample of 20 cars were observed in the past year:

Number of Accidents	Number of Cars
0	9
[1, 2]	6
[3, 4]	5

Calculate the maximum likelihood estimate of β .

Solution

Let N be the annual number of accidents per car.

The likelihood of a car having 0 accidents is

$$p_N(0) = \frac{1}{1 + \beta}$$

The likelihood of a car having 1 to 2 accidents is

$$\Pr(1 \leq N \leq 2) = p_N(1) + p_N(2) = \frac{\beta}{(1 + \beta)^2} + \frac{\beta^2}{(1 + \beta)^3}$$

The likelihood of a car having 3 to 4 accidents is

$$\Pr(3 \leq N \leq 4) = p_N(3) + p_N(4) = \frac{\beta^3}{(1 + \beta)^4} + \frac{\beta^4}{(1 + \beta)^5}$$

Then, the likelihood function is

$$\begin{aligned}
 L(\beta) &= p_N(0)^9 \cdot [p_N(1) + p_N(2)]^6 \cdot [p_N(3) + p_N(4)]^5 \\
 &= \left(\frac{1}{1+\beta} \right)^9 \left[\frac{\beta}{(1+\beta)^2} + \frac{\beta^2}{(1+\beta)^3} \right]^6 \left[\frac{\beta^3}{(1+\beta)^4} + \frac{\beta^4}{(1+\beta)^5} \right]^5 \\
 &= \left(\frac{1}{1+\beta} \right)^9 \left\{ \left[\frac{\beta}{(1+\beta)^2} \right]^6 \left[1 + \frac{\beta}{1+\beta} \right]^6 \right\} \left\{ \left[\frac{\beta^3}{(1+\beta)^4} \right]^5 \left[1 + \frac{\beta}{1+\beta} \right]^5 \right\} \\
 &= \frac{\beta^{6+3(5)}}{(1+\beta)^{9+2(6)+4(5)}} \left[\frac{1+2\beta}{1+\beta} \right]^{6+5} \\
 &= \frac{\beta^{21}(1+2\beta)^{11}}{(1+\beta)^{52}}
 \end{aligned}$$

The log-likelihood function is

$$l(\beta) = 21 \ln \beta + 11 \ln (1+2\beta) - 52 \ln (1+\beta)$$

Take its first derivative and set equal to 0. Simplify to get

$$l'(\beta) = \frac{21}{\beta} + \frac{22}{1+2\beta} - \frac{52}{1+\beta} = 0$$

$$\begin{aligned}
 \frac{21(1+2\beta)(1+\beta) + 22\beta(1+\beta) - 52\beta(1+2\beta)}{\beta(1+2\beta)(1+\beta)} &= 0 \\
 21(1+2\beta)(1+\beta) + 22\beta(1+\beta) - 52\beta(1+2\beta) &= 0 \\
 21 + 63\beta + 42\beta^2 + 22\beta + 22\beta^2 - 52\beta - 104\beta^2 &= 0 \\
 40\beta^2 - 33\beta - 21 &= 0
 \end{aligned}$$

Solve for β using the quadratic formula.

$$\begin{aligned}
 \beta &= \frac{-(-33) \pm \sqrt{(-33)^2 - 4(40)(-21)}}{2(40)} \\
 &= -0.4213 \quad \text{or} \quad 1.2463
 \end{aligned}$$

Since β cannot be negative, the final answer is $\hat{\beta} = \mathbf{1.2463}$.



2.1.5 Maximum Likelihood Estimation: Special Cases

There are many handy formulas for MLE estimates when certain distributions are assumed. We begin by discussing the following distributions in the context of **complete data**:

- Gamma
- Normal
- Poisson
- Binomial
- Negative binomial
- Uniform
- Laplace

For all but the last two distributions on this list, the MLE estimate is equivalent to the method of moments estimate. As most of the distributions involve estimating one parameter, the first sample moment is particularly important. The first sample moment is commonly referred to as the sample mean and denoted as

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (2.1.5.1)$$

GAMMA

For $X \sim \text{Gamma}(\alpha, \theta)$ where α is fixed, the MLE estimate of θ is the sample mean divided by α .

$$\mathbb{E}[X] = \alpha\theta = \bar{x} \quad \Rightarrow \quad \hat{\theta} = \frac{\bar{x}}{\alpha}$$

The **exponential** distribution is a special case of the gamma distribution with $\alpha = 1$. Thus, the MLE estimate of θ would then be the sample mean.

Note that this shortcut applies to Examples 2.1.3.1 and 2.1.3.2; we did not have to solve them using first principles.

NORMAL

From the list, the normal distribution is the only one where we can match the **first two moments** to obtain the MLE estimates of two parameters. For $X \sim \text{Normal}(\mu, \sigma^2)$, matching the moments produces

$$\mathbb{E}[X] = \mu = \bar{x}$$

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2 = \frac{\sum_{i=1}^n x_i^2}{n}$$

Therefore, the MLE estimates of μ and σ^2 are

$$\hat{\mu} = \bar{x}$$

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

Coach's Remarks

If μ is fixed rather than estimated using MLE, then the MLE estimate of σ^2 is

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

which is **not** equal to $\frac{\sum_{i=1}^n x_i^2}{n} - \mu^2$.

When asked to estimate the **lognormal** parameters, we can convert the lognormal data points to normal and apply the normal shortcuts. For $X \sim \text{Lognormal}(\mu, \sigma^2)$, the MLE estimates are

$$\hat{\mu} = \frac{\sum_{i=1}^n \ln x_i}{n}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (\ln x_i)^2}{n} - \hat{\mu}^2$$

POISSON

For $X \sim \text{Poisson}(\lambda)$, the MLE estimate of λ is the sample mean.

$$\hat{\lambda} = \bar{x}$$

BINOMIAL

For $X \sim \text{Binomial}(m, q)$ where m is fixed, the MLE estimate of q is the sample mean divided by m .

$$\mathbb{E}[X] = mq = \bar{x} \quad \Rightarrow \quad \hat{q} = \frac{\bar{x}}{m}$$

NEGATIVE BINOMIAL

For $X \sim \text{Negative Binomial}(r, \beta)$ where r is fixed, the MLE estimate of β is the sample mean divided by r .

$$\mathbb{E}[X] = r\beta = \bar{x} \quad \Rightarrow \quad \hat{\beta} = \frac{\bar{x}}{r}$$

The **geometric** distribution is a special case of the negative binomial distribution with $r = 1$. Thus, the MLE estimate of β would then be the sample mean.

Coach's Remarks

For the **gamma**, **binomial**, and **negative binomial** distributions, if θ , q , and β are estimated using maximum likelihood estimation, then the fitted mean will equal the sample mean. This is regardless of how α , m , and r are determined (i.e., fixed, MLE, or estimated using another method).

Thus, keep the following in mind:

- If a question provides you a value for α , m or r , you can quickly calculate the MLE estimate of θ , q , or β by equating the fitted mean to \bar{x} .
- If instead a question provides you an MLE estimate for θ , q , or β , then you can quickly estimate the missing parameter by equating the fitted mean to \bar{x} .

UNIFORM

For a uniform distribution on the interval $[0, \theta]$, the MLE estimate of θ is the largest observed value.

$$\hat{\theta} = \max(x_1, x_2, \dots, x_n) = x_{(n)}$$

To see why, recall that the PDF of the uniform distribution is the constant $\frac{1}{\theta}$ regardless of the observed value. With n data points, the likelihood function is simply this constant raised to the n^{th} power. This leads to the score function of

$$\begin{aligned} l'(\theta) &= \frac{d}{d\theta} \ln \left(\frac{1}{\theta^n} \right) \\ &= \frac{d}{d\theta} (-n \ln \theta) \\ &= -\frac{n}{\theta} \end{aligned}$$

However, there is no solution to the score equation, i.e. there is no finite value of θ that makes the score function equal 0.

We also could have realized this by noting that the likelihood function, θ^{-n} , is strictly decreasing for $\theta > 0$. Hence, it does not have a global maximum at a critical point. Nevertheless, since the goal is to maximize the likelihood function, it helps knowing that the likelihood function is strictly decreasing; the likelihood function increases as θ decreases.

$$\theta \downarrow \rightarrow L(\theta) = \frac{1}{\theta^n} \uparrow$$

This means we want θ as small as possible. Meanwhile, the uniform random variable must be at most θ , the upper limit of the distribution's range. Therefore, if the data comes from the uniform distribution, then it should be that

$$x_{(1)} \leq \dots \leq x_{(n)} \leq \theta$$

In satisfying these two requirements, the MLE estimate of θ is the largest observed value, $x_{(n)}$.

LAPLACE

For a Laplace or double exponential distribution with parameter θ , a possible MLE estimate is the sample median.

$$\hat{\theta} = \hat{\pi}_{0.5}$$

To see why, note that the PDF of the Laplace distribution is

$$f(x) = \frac{1}{2}e^{-|x-\theta|}, \quad -\infty < x < \infty$$

With n data points, this leads to the score function of

$$l'(\theta) = \sum_{i=1}^n w_i$$

where

$$w_i = \begin{cases} -1, & x_i < \theta \\ 0, & x_i = \theta \\ 1, & x_i > \theta \end{cases}$$

Thus, the score function will equal 0 when the number of observed values less than θ equals the number of observed values greater than θ . This is achieved when θ is the sample median.

Coach's Remarks

We believe that this is the only concept regarding the Laplace distribution to know and memorize for the exam.

Example 2.1.5.1

You observe the following claims from a dataset:

2 5 8 13 16

You fit a uniform distribution on $[0, \theta]$ to the data.

Estimate the variance of the fitted distribution using maximum likelihood.

Solution

For this uniform distribution, the variance formula is

$$\frac{\theta^2}{12}$$

Recall that the MLE estimate of θ is the largest observed value from the sample. Therefore, $\hat{\theta} = 16$. Using the invariance property, the final answer is

$$\frac{16^2}{12} = \mathbf{21.333}$$

■

Example 2.1.5.2

You observe the following claims from a dataset:

25 70 215 535

You fit a lognormal distribution to the data.

Estimate the mean of the fitted distribution using maximum likelihood.

Solution

Let X represent the claim variable, so $X \sim \text{Lognormal}(\mu, \sigma^2)$. Then,

$$\ln X \sim \text{Normal}(\mu, \sigma^2)$$

We can estimate μ and σ^2 using the normal shortcuts.

$$\begin{aligned}\hat{\mu} &= \frac{\sum_{i=1}^4 \ln x_i}{4} \\ &= \frac{\ln 25 + \ln 70 + \ln 215 + \ln 535}{4} \\ &= 4.7801\end{aligned}$$

$$\begin{aligned}
 \hat{\sigma}^2 &= \frac{\sum_{i=1}^4 (\ln x_i)^2}{4} - \hat{\mu}^2 \\
 &= \frac{(\ln 25)^2 + (\ln 70)^2 + (\ln 215)^2 + (\ln 535)^2}{4} - 4.7801^2 \\
 &= 1.3313
 \end{aligned}$$

Look up the lognormal distribution's mean formula in the exam table and apply the invariance property. The final answer is

$$e^{4.7801+0.5(1.3313)} = \mathbf{231.7657}$$

■

There are other MLE estimates with relatively simple closed form expressions, even when permitting **truncated and/or censored data**. Some are summarized in the table below.

In this table:

- n is the number of uncensored data points
- c is the number of censored data points
- x_i is the i^{th} observed value, or the censoring point for censored data
- d_i is the truncation point for the i^{th} observation

Distribution	Shortcut
Pareto, fixed θ	$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n+c} [\ln(x_i + \theta) - \ln(d_i + \theta)]}$
S-P Pareto, fixed θ	$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n+c} \{\ln x_i - \ln [\max(\theta, d_i)]\}}$
Exponential	$\hat{\theta} = \frac{\sum_{i=1}^{n+c} (x_i - d_i)}{n}$
Weibull, fixed τ	$\hat{\theta} = \left(\frac{\sum_{i=1}^{n+c} x_i^\tau - \sum_{i=1}^{n+c} d_i^\tau}{n} \right)^{1/\tau}$

For complete data, the shortcuts simplify with c and d_i 's all equaling 0. Note that these shortcuts could have been used in Examples 2.1.3.1, 2.1.3.3, 2.1.4.1, and 2.1.4.2.

Coach's Remarks

The shortcuts in the table are optional, since they would likely impact only a few problems on the exam. Having said that, the exponential shortcut is likely the most useful, so you may want to memorize it.

Example 2.1.5.3

An insurance company offers two types of policies:

- Policy I has no deductible and a policy limit of 100.
- Policy II has a deductible of 20 and no policy limit.

You are given the following samples from these two policies. Losses below the deductible are not recorded.

Policy	Observed Losses						
I	50	60	60	70	80	100	100
II	30	50	60	70	90	120	

An actuary fits a ground-up exponential distribution using maximum likelihood estimation.

Estimate the mean of the ground-up distribution.

Solution

Recall that ground-up refers to the loss amount without any coverage modifications. Let x_i be the i^{th} observed loss, and d_i be the deductible that x_i is subject to. In addition, n and c are the numbers of uncensored and censored losses, respectively. The mean of an exponential distribution is θ , and the MLE estimate of θ is

$$\hat{\theta} = \frac{\sum_{i=1}^{n+c} (x_i - d_i)}{n}$$

For Policy I, there are 5 uncensored observations.

x_i	d_i	$x_i - d_i$
50	0	50
60	0	60
60	0	60
70	0	70
80	0	80
100	0	100
100	0	100

For Policy II, there are 6 uncensored observations.

x_i	d_i	$x_i - d_i$
30	20	10
50	20	30
60	20	40
70	20	50
90	20	70
120	20	100

Use the exponential shortcut to calculate the MLE estimate of θ .

$$\begin{aligned}\hat{\theta} &= \frac{(50 + 60 + \dots + 100) + (10 + 30 + \dots + 100)}{5 + 6} \\ &= \frac{820}{11} \\ &= \mathbf{74.5455}\end{aligned}$$

Coach's Remarks

The shortcut can be memorized as

$$\hat{\theta} = \frac{\text{total payment}}{\# \text{ of uncensored observations}}$$

In this example, 820 is the total amount the insurance company will pay out for the given set of observed losses, and there are 11 uncensored observations in total.

2.1.6 Kernel Density Estimation

Rather than assuming a continuous distribution for $f(x)$ and performing point estimation on its parameter(s), we may decide to estimate the PDF directly using ***kernel density estimation*** on sample data.

A ***kernel function***, denoted as $k(\cdot)$, is a probability density function with two parameters:

- The i^{th} observed value, x_i
- The ***bandwidth***, b

For every kernel function we will discuss, $k(\cdot)$ is a symmetric density, and x_i represents the center point of symmetry. In other words, x_i represents the **mean** of the kernel function's distribution. Consequently, we denote $k_i(\cdot)$ as the kernel function where x_i is the center of its domain. As for the bandwidth b , its interpretation will depend on the specific function used for $k(\cdot)$.

The idea of kernel density estimation is to:

1. Choose a density form for $k(\cdot)$.
2. Estimate $f(x)$ as the average of $k_1(x), \dots, k_n(x)$.

Therefore, the kernel density estimate of $f(x)$ is

$$\tilde{f}(x) = \frac{1}{n} \sum_{i=1}^n k_i(x) \quad (2.1.6.1)$$

You may interpret this estimate as a discrete mixture, i.e. compare with Equation 1.1.9.1.

Coach's Remarks

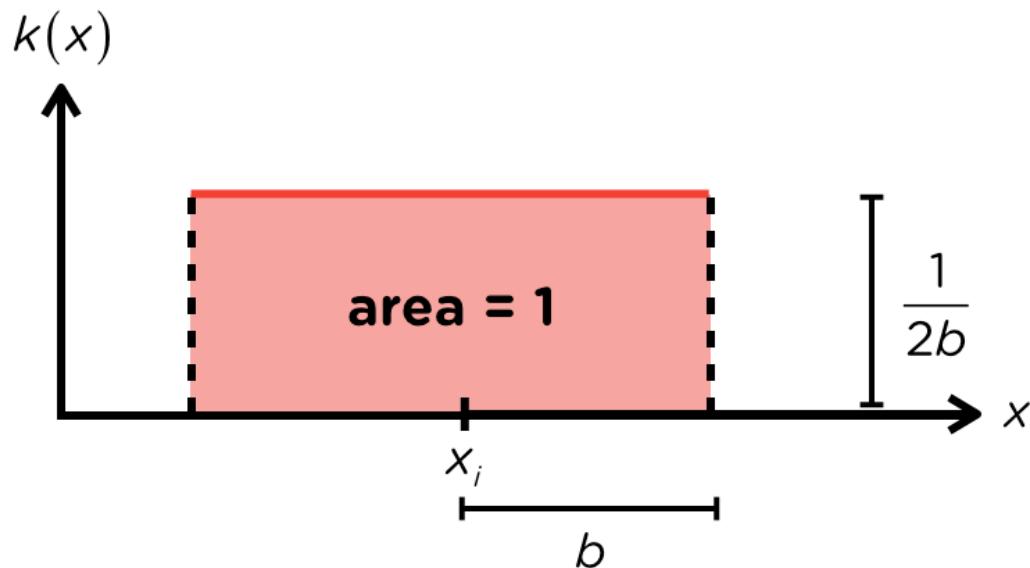
It is very important to not confuse x (the variable of the densities \tilde{f} and k) and x_i (a parameter for density k). To minimize confusion, we will put less emphasis on the math expressions where possible.

For this exam, there are three types of kernel functions to consider:

- Rectangular
- Triangular
- Gaussian

Rectangular Kernels

A **rectangular** or **uniform kernel** assumes a uniform density for the kernel function. The graph below depicts the rectangular kernel for the i^{th} observed value.



For a rectangular kernel, the bandwidth b is the distance from the center point x_i to either end of the domain. Thus, the domain spans a distance of $2b$. Since the kernel function is a density function, the rectangle beneath the function must have an area of 1. This forces the height of the rectangle, i.e. the kernel function itself, to equal $1 / 2b$.

You observe the following claim sizes: 5 2 6

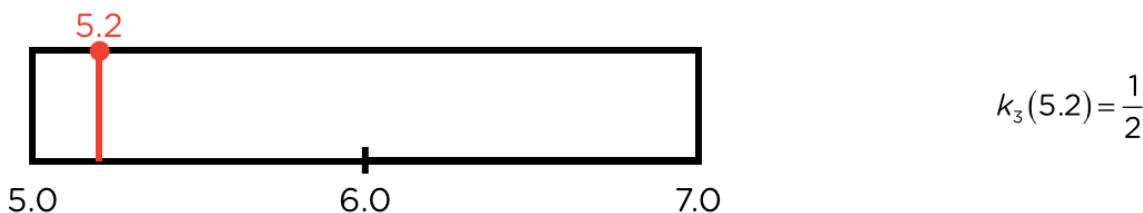
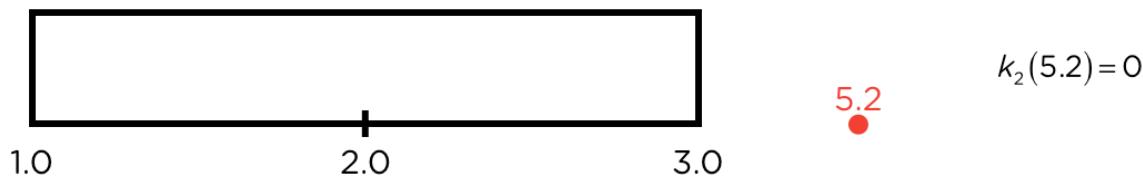
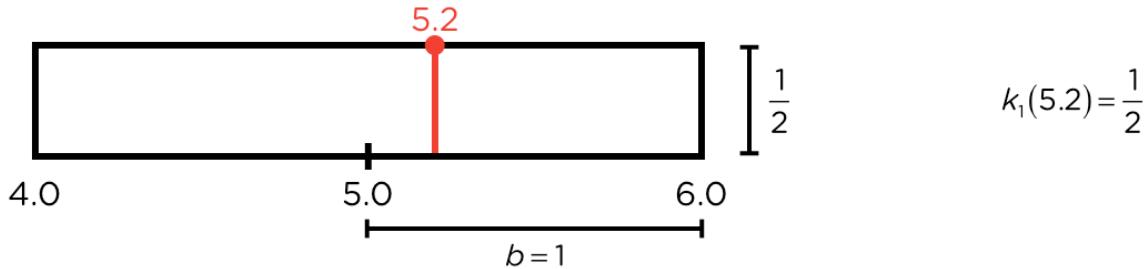
Using a rectangular kernel with bandwidth 1, estimate

1. the PDF of claim sizes evaluated at 5.2.
2. the probability that a claim size is less than 5.2.

Using Equation 2.1.6.1, the goal of Part (1) is

$$\tilde{f}(5.2) = \frac{1}{3}[k_1(5.2) + k_2(5.2) + k_3(5.2)]$$

An easy way to calculate each of the $k_i(5.2)$'s is to sketch the rectangular kernel for each observed value, and notice the heights where $x = 5.2$.



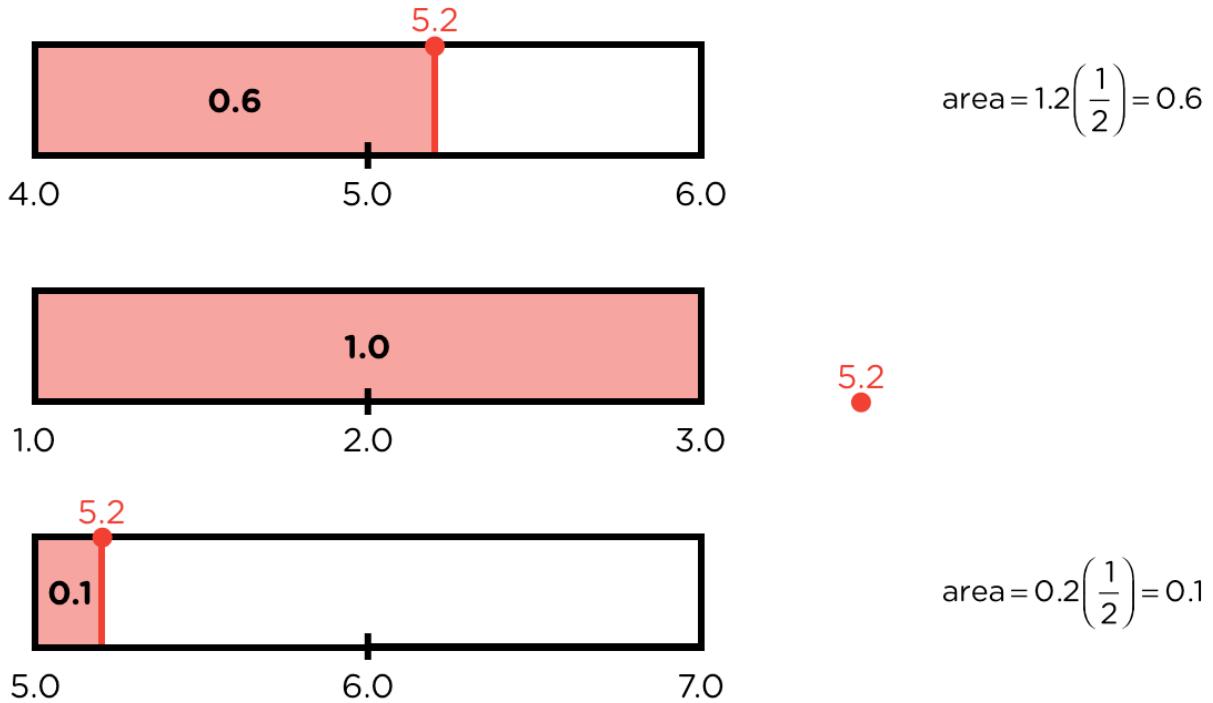
As a result,

$$\tilde{f}(5.2) = \frac{1}{3} \left(\frac{1}{2} + 0 + \frac{1}{2} \right) = \frac{1}{3}$$

Using first principles, Part (2) is solved by

$$\begin{aligned} \int_{-\infty}^{5.2} \tilde{f}(x) dx &= \int_{-\infty}^{5.2} \frac{1}{3}[k_1(x) + k_2(x) + k_3(x)] dx \\ &= \frac{1}{3} \left(\int_{-\infty}^{5.2} k_1(x) dx + \int_{-\infty}^{5.2} k_2(x) dx + \int_{-\infty}^{5.2} k_3(x) dx \right) \end{aligned}$$

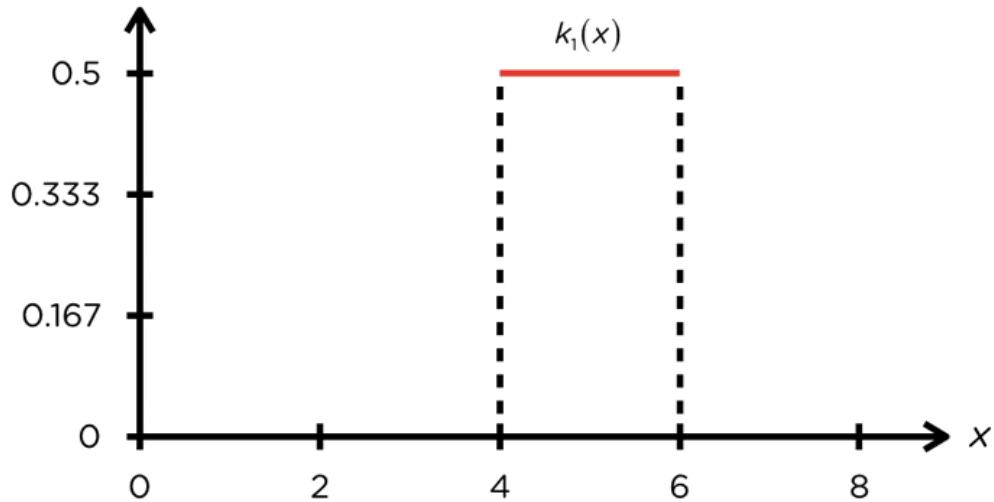
In other words, the strategy to estimate a cumulative (or survival) probability is similar to Equation 2.1.6.1, where the $k_i(x)$'s are replaced with the cumulative (or survival) probabilities from each kernel function; in relation to discrete mixtures, consider Equation 1.1.9.2. Once again, an easy approach is to calculate the areas under the rectangles for the region of the domain that is less than 5.2.



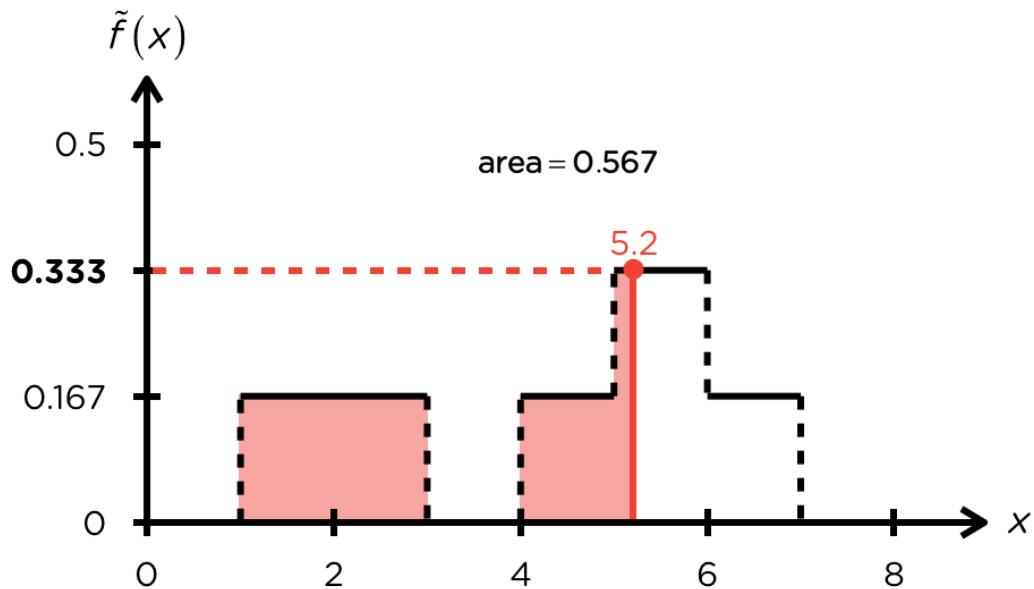
Therefore, the estimated probability that a claim size is less than 5.2 is

$$\frac{1}{3}(0.6 + 1 + 0.1) = \mathbf{0.567}$$

Here is a visual representation of constructing $\tilde{f}(x)$ with the sample data:



From the graph of $\tilde{f}(x)$, we can visualize the answers to this example as well.



Coach's Remarks

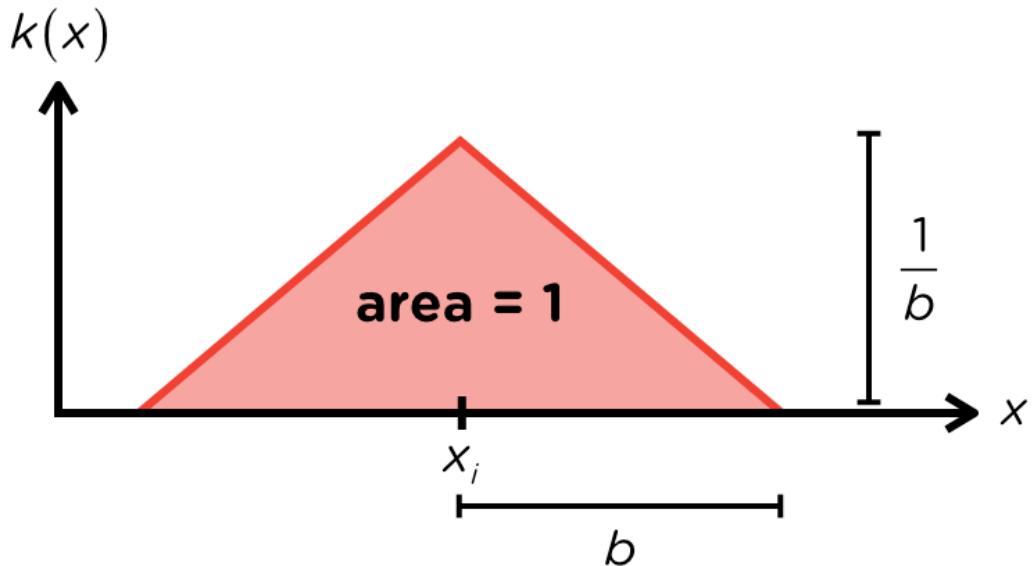
A rectangular kernel centered on x_i can be expressed mathematically as

$$k_i(x) = \begin{cases} \frac{1}{2b}, & x_i - b \leq x \leq x_i + b \\ 0, & \text{otherwise} \end{cases}$$

However, we discourage memorizing this function, as it is more beneficial to understand the concepts intuitively for solving exam problems.

Triangular Kernels

A *triangular kernel* assumes a density in the shape of an isosceles triangle for the kernel function. The graph below depicts the triangular kernel for the i^{th} observed value.



Similar to a rectangular kernel, the bandwidth b is the distance from the center point x_i to either end of the domain. While the domain still spans a distance of $2b$, it is now the triangle beneath the function that must have an area of 1. This forces the height of the triangle to equal $1 / b$.

You observe the following claim sizes: 5 2 6

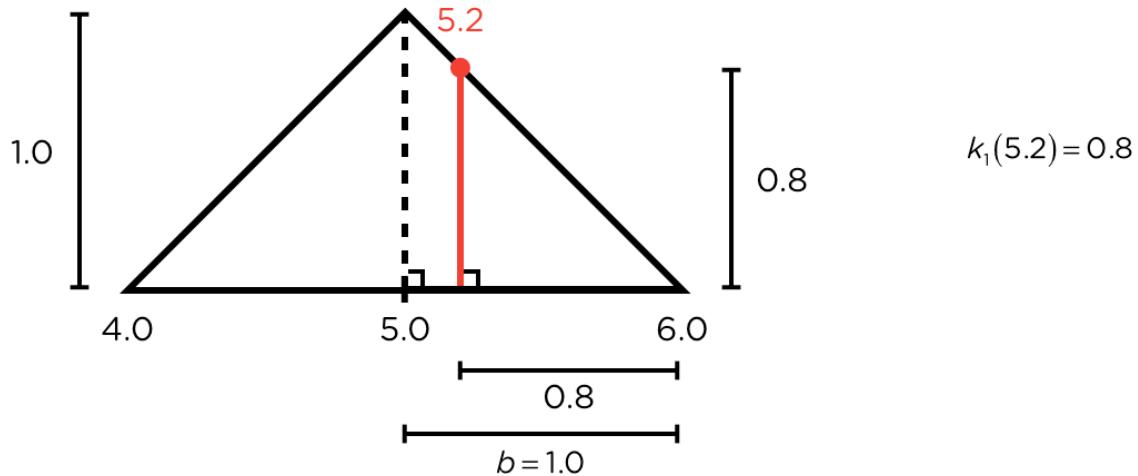
Using a triangular kernel with bandwidth 1, estimate

1. the PDF of claim sizes evaluated at 5.2.
2. the probability that a claim size is less than 5.2.

Using Equation 2.1.6.1, the goal of Part (1) is

$$\tilde{f}(5.2) = \frac{1}{3}[k_1(5.2) + k_2(5.2) + k_3(5.2)]$$

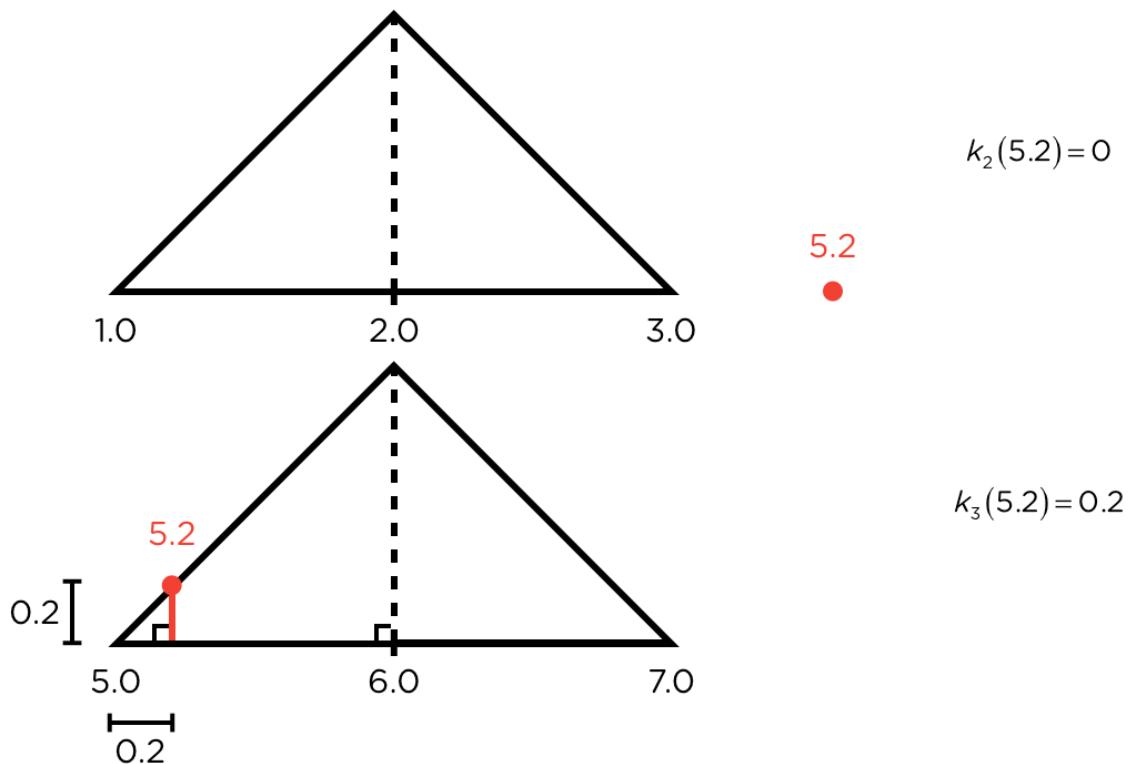
Sketch the triangular kernel for each observed value, and notice the heights where $x = 5.2$.



In the right half of the kernel, notice there are two right triangles that are similar triangles. As a property of similar right triangles, the ratio of triangle heights must equal the ratio of triangle bases. Therefore,

$$\frac{k_1(5.2)}{1/b} = \frac{0.8}{b} \quad \Rightarrow \quad k_1(5.2) = \underbrace{\frac{0.8}{b}}_{\text{ratio}} \times \underbrace{\frac{1}{b}}_{\text{triangle height}} = 0.8$$

Next,



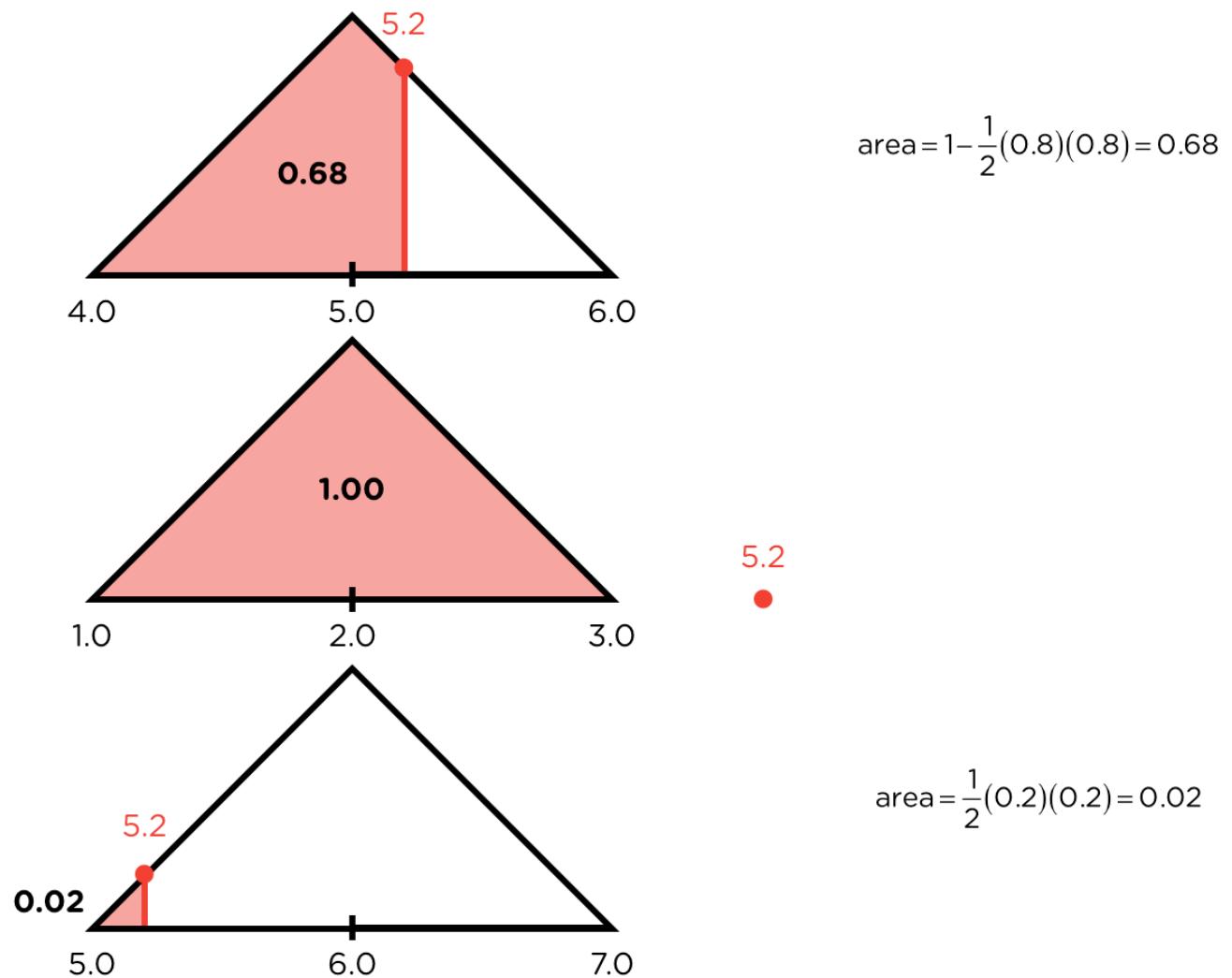
In the left half of the kernel, notice there are two right triangles that are similar triangles. Therefore,

$$k_3(5.2) = \frac{0.2}{b} \left(\frac{1}{b} \right) = 0.2$$

As a result,

$$\tilde{f}(5.2) = \frac{1}{3}(0.8 + 0 + 0.2) = \frac{1}{3}$$

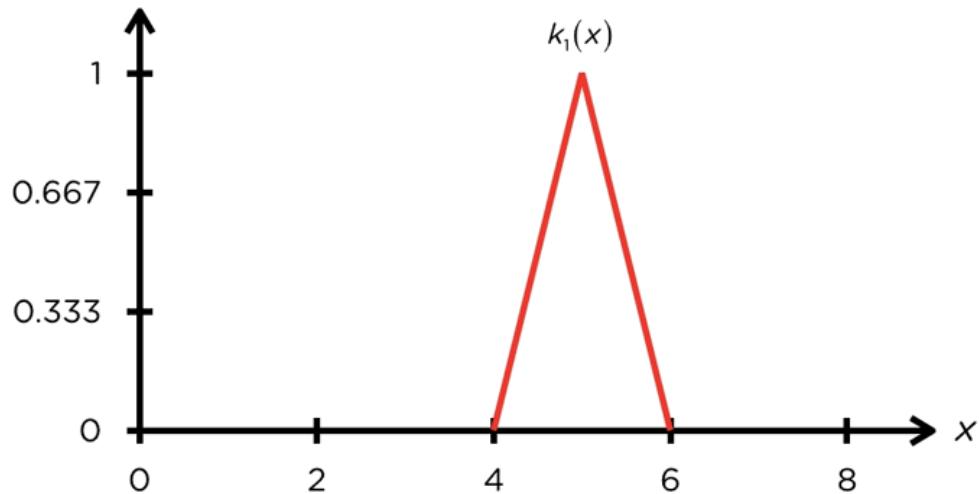
For Part (2), calculate the areas under the triangles for the region of the domain that is less than 5.2, and then take their average.



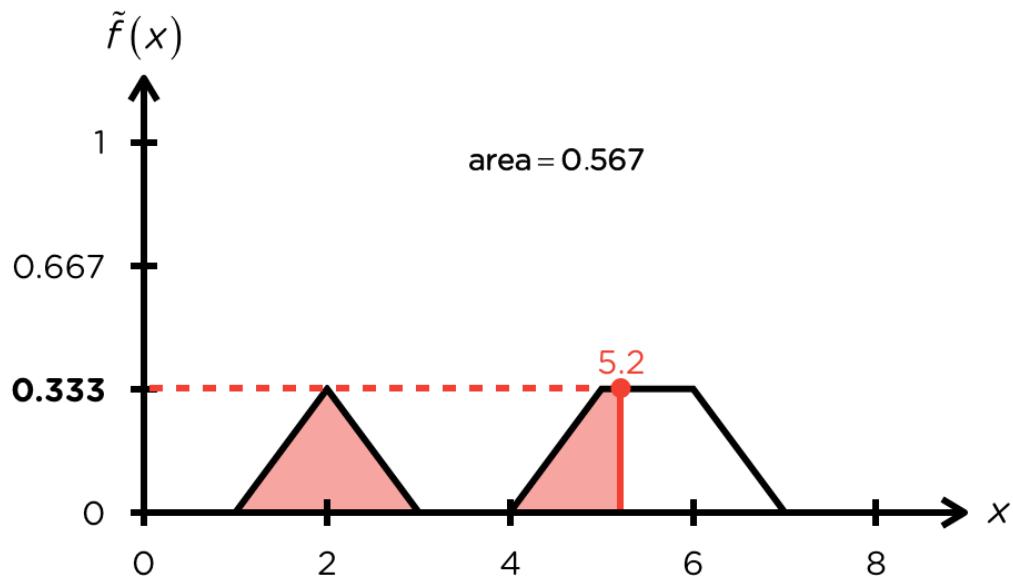
Therefore, the estimated probability that a claim size is less than 5.2 is

$$\frac{1}{3}(0.68 + 1 + 0.02) = \mathbf{0.567}$$

Here is a visual representation of constructing $\tilde{f}(x)$ with the sample data:



As expected, we can visualize the answers to this example from the graph of $\tilde{f}(x)$.



Coach's Remarks

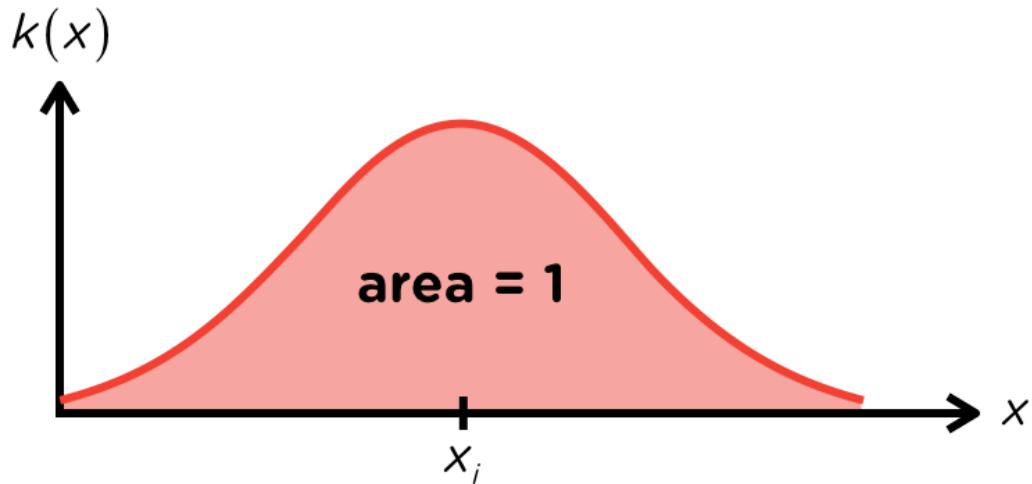
A triangular kernel centered on x_i can be expressed mathematically as

$$\begin{aligned}
 k_i(x) &= \begin{cases} \frac{b - |x - x_i|}{b^2}, & x_i - b \leq x \leq x_i + b \\ 0, & \text{otherwise} \end{cases} \\
 &= \begin{cases} \frac{b - x_i + x}{b^2}, & x_i - b \leq x \leq x_i \\ \frac{b + x_i - x}{b^2}, & x_i < x \leq x_i + b \\ 0, & \text{otherwise} \end{cases}
 \end{aligned}$$

However, we discourage memorizing this function, as it is more beneficial to understand the concepts intuitively for solving exam problems.

Gaussian Kernels

A **Gaussian kernel** assumes a normal density with mean x_i and variance b^2 for the kernel function. Thus, the bandwidth for a Gaussian kernel is the standard deviation of the normal distribution. The graph below depicts the Gaussian kernel for the i^{th} observed value.



Given the PDF in Section 1.1.4, the Gaussian kernel for the i^{th} observed value has the form

$$k_i(x) = \frac{1}{b\sqrt{2\pi}} \exp \left[-\frac{(x - x_i)^2}{2b^2} \right], \quad -\infty < x < \infty$$

You observe the following claim sizes: 5 2 6

Using a Gaussian kernel with bandwidth 1,

1. the PDF of claim sizes evaluated at 5.2.
2. the probability that a claim size is less than 5.2.

To solve Part (1), compute the Gaussian kernel for each observed value, then use Equation 2.1.6.1.

For the first Gaussian kernel,

$$k_1(5.2) = \frac{1}{1\sqrt{2\pi}} \exp \left[-\frac{(5.2 - 5)^2}{2 \cdot 1^2} \right] = 0.3910$$

For the second Gaussian kernel,

$$k_2(5.2) = \frac{1}{1\sqrt{2\pi}} \exp \left[-\frac{(5.2 - 2)^2}{2 \cdot 1^2} \right] = 0.0024$$

For the third Gaussian kernel,

$$k_3(5.2) = \frac{1}{1\sqrt{2\pi}} \exp \left[-\frac{(5.2 - 6)^2}{2 \cdot 1^2} \right] = 0.2897$$

As a result,

$$\tilde{f}(5.2) = \frac{1}{3} (0.3910 + 0.0024 + 0.2897) = \mathbf{0.2277}$$

To solve Part (2), calculate the areas under the normal densities for the region of the domain that is less than 5.2, and then take their average.

For the first Gaussian kernel,

$$\Pr\left(Z \leq \frac{5.2 - 5}{1}\right) = \Phi(0.2) = 0.5793$$

For the second Gaussian kernel,

$$\Pr\left(Z \leq \frac{5.2 - 2}{1}\right) = \Phi(3.2) \approx 1$$

For the third Gaussian kernel,

$$\Pr\left(Z \leq \frac{5.2 - 6}{1}\right) = \Phi(-0.8) = 0.2119$$

Therefore, the estimated probability that the claim size is less than 5.2 is

$$\frac{1}{3}(0.5793 + 1 + 0.2119) = \mathbf{0.597}$$

Example 2.1.6.1

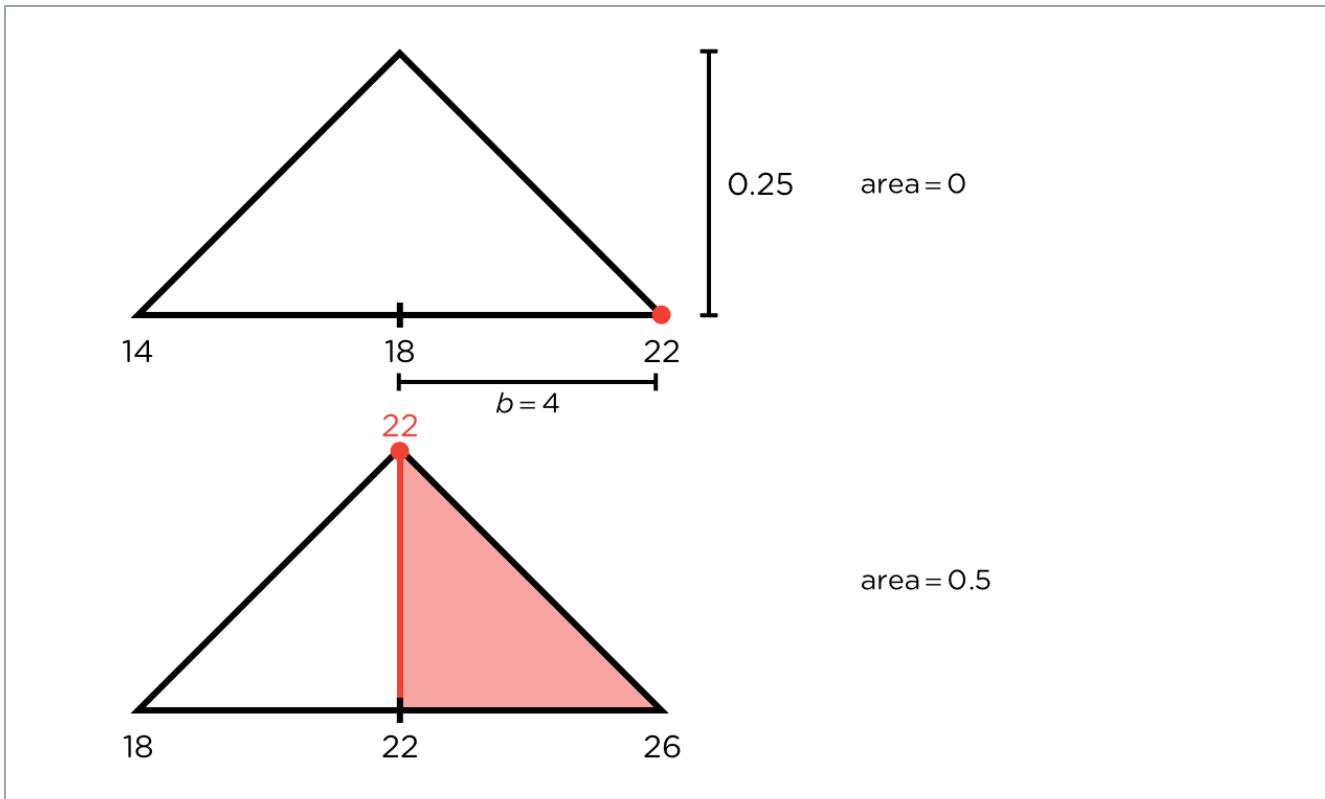
Losses due to fire were observed as:

18 22 25 27 30 30

Using a triangular kernel with bandwidth 4, estimate the probability that a loss due to fire is between 22 and 28.

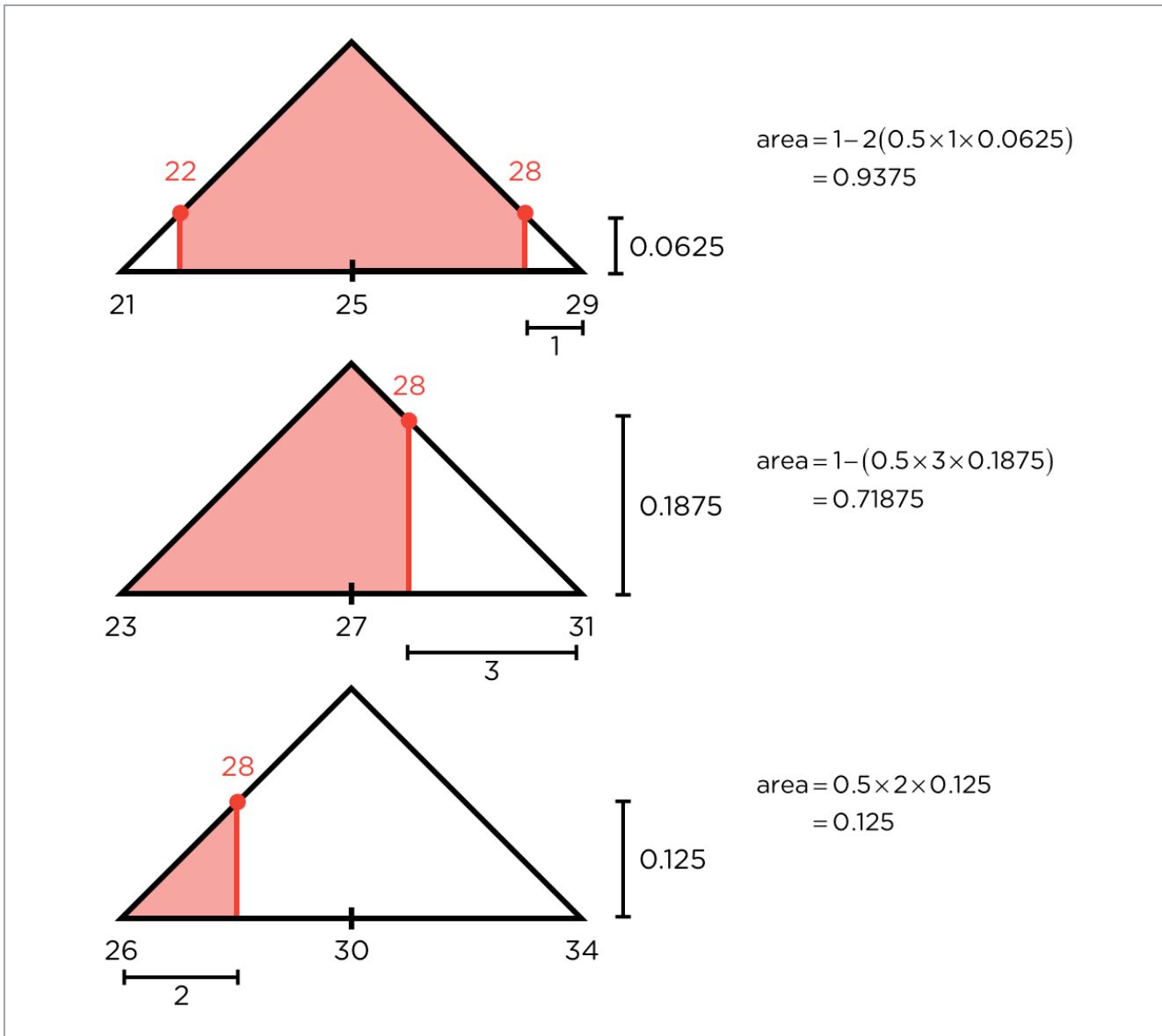
Solution

Over the interval [22, 28], calculate the areas under the triangular kernels. With $b = 4$, the height of the triangular kernels is $1/4 = 0.25$.



For the other three kernels, the following $k_i(x)$'s are needed:

Observed Value	$k_i(22)$	$k_i(28)$
25	$\frac{1}{4}(0.25) = 0.0625$	$\frac{1}{4}(0.25) = 0.0625$
27	-	$\frac{3}{4}(0.25) = 0.1875$
30	-	$\frac{2}{4}(0.25) = 0.125$



Average the areas to compute the final answer.

$$\frac{1}{6}[0 + 0.5 + 0.9375 + 0.71875 + 2(0.125)] = \mathbf{0.401}$$



2.1 Summary

🕒 5m

Method of Moments

Estimate parameters by setting the theoretical moments equal to the sample moments.

$$\mathbb{E}[X^k] = \frac{\sum_{i=1}^n x_i^k}{n}$$

Percentile Matching

Estimate parameters by setting the theoretical percentiles equal to the sample percentiles.

$$\pi_{q_k} = \hat{\pi}_{q_k}$$

where the q_k 's are arbitrarily chosen probabilities. Sample percentiles are calculated using the smoothed empirical percentile approach.

Maximum Likelihood Estimation (MLE)

$$L(\theta) = \prod_{i=1}^n f(x_i)$$

Estimate θ as the value that maximizes $L(\theta)$ or $l(\theta) = \ln [L(\theta)]$. Typically, write the equation $l'(\theta) = 0$, then solve for θ .

- MLE has the invariance property.
- Instead of $f(x)$, the likelihood of
 - an observation from a dataset that is left-truncated at d is $\frac{f(x)}{\Pr(X>d)}$.

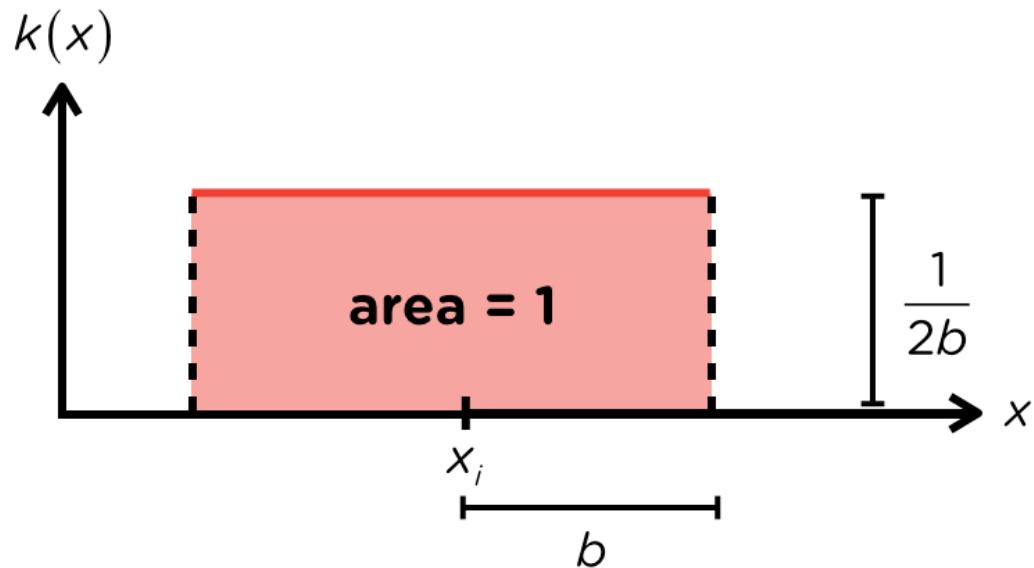
- an observation that is right-censored at m is $\Pr(X \geq m)$.
- an observation recorded as in the interval $(a, b]$ is $\Pr(a < X \leq b)$.
- With complete data, the MLE estimates are the same as the method of moments estimates for:
 - Gamma with fixed α
 - Normal
 - Poisson
 - Binomial with fixed m
 - Negative binomial with fixed r
- With complete data, $\hat{\theta}$ for a uniform distribution on the interval $[0, \theta]$ is the largest observed value, $x_{(n)}$.
- With complete data, a possible $\hat{\theta}$ for a Laplace distribution is the sample median, $\hat{\pi}_{0.5}$.
- Permitting truncated and/or censored data,

Distribution	Shortcut
Pareto, fixed θ	$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n+c} [\ln(x_i + \theta) - \ln(d_i + \theta)]}$
S-P Pareto, fixed θ	$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n+c} \{\ln x_i - \ln [\max(\theta, d_i)]\}}$
Exponential	$\hat{\theta} = \frac{\sum_{i=1}^{n+c} (x_i - d_i)}{n}$
Weibull, fixed τ	$\hat{\theta} = \left(\frac{\sum_{i=1}^{n+c} x_i^\tau - \sum_{i=1}^{n+c} d_i^\tau}{n} \right)^{1/\tau}$

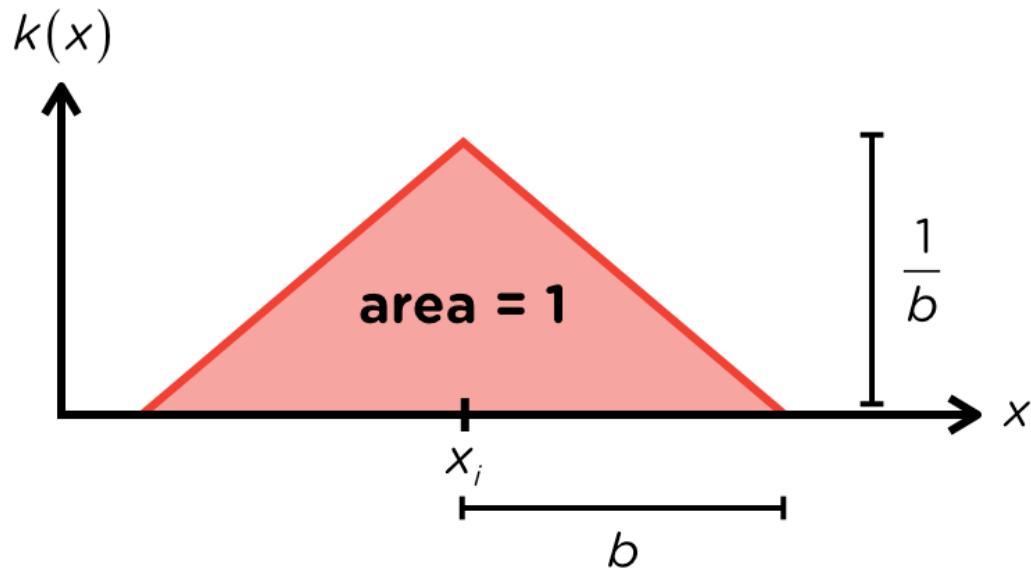
Kernel Density Estimation

$$\tilde{f}(x) = \frac{1}{n} \sum_{i=1}^n k_i(x)$$

RECTANGULAR KERNELS

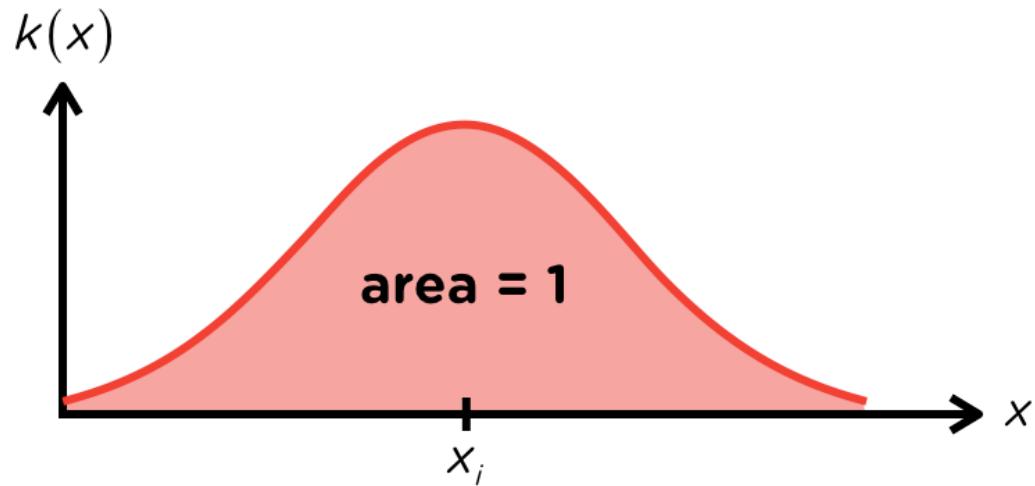


TRIANGULAR KERNELS



Use similar right triangles to calculate $k_i(x)$.

GAUSSIAN KERNELS



b is the standard deviation of the normal distribution.

Appendix

🕒 30m

MLE Matching Moments Shortcuts**Gamma with fixed α**

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \frac{(x_i/\theta)^\alpha e^{-x_i/\theta}}{x_i \Gamma(\alpha)} \\ &= \prod_{i=1}^n \frac{x_i^{\alpha-1} e^{-x_i/\theta}}{\theta^\alpha \Gamma(\alpha)} \\ &= k \cdot \frac{e^{-\sum_{i=1}^n x_i/\theta}}{\theta^{n\alpha}} \end{aligned}$$

where $k = \prod_{i=1}^n x_i^{\alpha-1} \Gamma(\alpha)^{-1}$.

$$l(\theta) = \ln k - \frac{\sum_{i=1}^n x_i}{\theta} - n\alpha \ln \theta$$

$$l'(\theta) = 0 + \frac{\sum_{i=1}^n x_i}{\theta^2} - \frac{n\alpha}{\theta} = 0$$

Thus,

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{i=1}^n x_i}{n\alpha} \\ &= \frac{\bar{x}}{\alpha} \end{aligned}$$



Normal

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= k \cdot (\sigma^2)^{-\frac{n}{2}} e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}} \end{aligned}$$

where $k = (\sqrt{2\pi})^{-n}$.

$$l(\mu, \sigma^2) = \ln k - \frac{n}{2} \ln \sigma^2 - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}$$

Take partial derivatives of the log-likelihood function with respect to μ and σ^2 respectively.

$$\begin{aligned} \frac{\partial}{\partial \mu} l(\mu, \sigma^2) &= 0 - \frac{\sum_{i=1}^n [2(x_i - \mu)(-1)]}{2\sigma^2} \\ &= \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} \\ &= \frac{(\sum_{i=1}^n x_i) - n\mu}{\sigma^2} \end{aligned}$$

$$\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2) = 0 - \frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^4}$$

Set both partial derivatives equal to zero. Now, we have two equations to solve for two parameters.

$$\frac{\sum_{i=1}^n x_i - n\mu}{\sigma^2} = 0 \quad -(1)$$

$$-\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^4} = 0 \quad -(2)$$

Multiply (1) by σ^2 .

$$\begin{aligned} \sum_{i=1}^n x_i - n\mu &= 0 \\ \hat{\mu} &= \frac{\sum_{i=1}^n x_i}{n} \\ &= \bar{x} \end{aligned}$$

Multiply (2) by $2\sigma^4$.

$$\begin{aligned} -n\sigma^2 + \sum_{i=1}^n (x_i - \mu)^2 &= 0 \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{n} \end{aligned}$$



Lognormal

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^n \frac{1}{\sigma x_i \sqrt{2\pi}} e^{-\frac{(\ln x_i - \mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sigma^n (\sqrt{2\pi})^n \prod_{i=1}^n x_i} \exp \left[\sum_{i=1}^n -\frac{(\ln x_i - \mu)^2}{2\sigma^2} \right] \\ &= k \cdot \frac{1}{\sigma^n} \cdot \exp \left[\sum_{i=1}^n -\frac{(\ln x_i - \mu)^2}{2\sigma^2} \right] \end{aligned}$$

where $k = \frac{1}{(\sqrt{2\pi})^n \prod_{i=1}^n x_i}$.

$$l(\mu, \sigma) = \ln k + \sum_{i=1}^n -\frac{(\ln x_i - \mu)^2}{2\sigma^2} - n \ln \sigma$$

Take partial derivatives of the log-likelihood function with respect to μ and σ respectively.

$$\begin{aligned}\frac{\partial}{\partial \mu} l(\mu, \sigma) &= \sum_{i=1}^n -\frac{2(\ln x_i - \mu)}{2\sigma^2} \cdot (-1) \\ &= \sum_{i=1}^n \frac{(\ln x_i - \mu)}{\sigma^2}\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \sigma} l(\mu, \sigma) &= \sum_{i=1}^n -\frac{-2(\ln x_i - \mu)^2}{2\sigma^3} - \frac{n}{\sigma} \\ &= \sum_{i=1}^n \frac{(\ln x_i - \mu)^2}{\sigma^3} - \frac{n}{\sigma}\end{aligned}$$

Set both partial derivatives equal to zero. Now, we have two equations to solve for two parameters.

$$\sum_{i=1}^n \frac{(\ln x_i - \mu)}{\sigma^2} = 0 \quad - (1)$$

$$\sum_{i=1}^n \frac{(\ln x_i - \mu)^2}{\sigma^3} - \frac{n}{\sigma} = 0 \quad - (2)$$

Multiply (1) by σ^2 .

$$\begin{aligned} \sum_{i=1}^n (\ln x_i - \mu) &= 0 \\ \sum_{i=1}^n \ln x_i - \sum_{i=1}^n \mu &= 0 \\ n\mu &= \sum_{i=1}^n \ln x_i \\ \hat{\mu} &= \frac{\sum_{i=1}^n \ln x_i}{n} \end{aligned}$$

Multiply (2) by σ^3 .

$$\begin{aligned} \sum_{i=1}^n (\ln x_i - \mu)^2 - n\sigma^2 &= 0 \\ -n\sigma^2 &= -\sum_{i=1}^n (\ln x_i - \mu)^2 \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^n (\ln x_i - \hat{\mu})^2}{n} \end{aligned}$$



Poisson

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \\ &= k \cdot e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} \end{aligned}$$

where $k = \prod_{i=1}^n (x_i!)^{-1}$.

$$l(\lambda) = \ln k - n\lambda + \left(\sum_{i=1}^n x_i \right) \ln \lambda$$

$$l'(\lambda) = 0 - n + \frac{\sum_{i=1}^n x_i}{\lambda} = 0$$

Thus,

$$\begin{aligned}\hat{\lambda} &= \frac{\sum_{i=1}^n x_i}{n} \\ &= \bar{x}\end{aligned}$$

■

Binomial with fixed m

$$\begin{aligned}L(q) &= \prod_{i=1}^n \binom{m}{x_i} q^{x_i} (1-q)^{m-x_i} \\ &= k \cdot q^{\sum_{i=1}^n x_i} (1-q)^{\sum_{i=1}^n (m-x_i)}\end{aligned}$$

$$\text{where } k = \prod_{i=1}^n \binom{m}{x_i}.$$

$$l(q) = \ln k + \left(\sum_{i=1}^n x_i \right) \ln q + \left[\sum_{i=1}^n (m - x_i) \right] \ln(1 - q)$$

$$\begin{aligned}
 l'(q) &= 0 + \frac{\sum_{i=1}^n x_i}{q} - \frac{\sum_{i=1}^n (m - x_i)}{1-q} = 0 \\
 \sum_{i=1}^n x_i - \left(\sum_{i=1}^n x_i \right) q - nmq + \left(\sum_{i=1}^n x_i \right) q &= 0
 \end{aligned}$$

Thus,

$$\begin{aligned}
 \hat{q} &= \frac{\sum_{i=1}^n x_i}{nm} \\
 &= \frac{\bar{x}}{m}
 \end{aligned}$$

■

Negative Binomial with fixed r

$$\begin{aligned}
 L(\beta) &= \prod_{i=1}^n \frac{r(r+1)\dots(x_i+r-1)}{x_i!} \cdot \frac{\beta^{x_i}}{(1+\beta)^{x_i+r}} \\
 &= k \cdot \frac{\beta^{\sum_{i=1}^n x_i}}{(1+\beta)^{\sum_{i=1}^n (x_i+r)}}
 \end{aligned}$$

$$\text{where } k = \prod_{i=1}^n \frac{r(r+1)\dots(x_i+r-1)}{x_i!}.$$

$$l(\beta) = \ln k + \left(\sum_{i=1}^n x_i \right) \ln \beta - \left[\sum_{i=1}^n (x_i + r) \right] \ln(1 + \beta)$$

$$\begin{aligned}
 l'(\beta) &= 0 + \frac{\sum_{i=1}^n x_i}{\beta} - \frac{\sum_{i=1}^n (x_i + r)}{1 + \beta} = 0 \\
 &\sum_{i=1}^n x_i + \left(\sum_{i=1}^n x_i \right) \beta - \left(\sum_{i=1}^n x_i \right) \beta - nr\beta = 0
 \end{aligned}$$

Thus,

$$\begin{aligned}
 \hat{\beta} &= \frac{\sum_{i=1}^n x_i}{nr} \\
 &= \frac{\bar{x}}{r}
 \end{aligned}$$

■

Other MLE Shortcuts

Pareto with fixed θ

$$\begin{aligned}
 L(\alpha) &= \prod_{i=1}^n \frac{\frac{\alpha\theta^\alpha}{(x_i+\theta)^{\alpha+1}}}{\left(\frac{\theta}{d_i+\theta}\right)^\alpha} \cdot \prod_{i=n+1}^{n+c} \frac{\left(\frac{\theta}{x_i+\theta}\right)^\alpha}{\left(\frac{\theta}{d_i+\theta}\right)^\alpha} \\
 &= \prod_{i=1}^n \frac{\alpha(d_i+\theta)^\alpha}{(x_i+\theta)^{\alpha+1}} \cdot \prod_{i=n+1}^{n+c} \left(\frac{d_i+\theta}{x_i+\theta}\right)^\alpha \\
 &= k \cdot \alpha^n \left(\prod_{i=1}^{n+c} \frac{d_i+\theta}{x_i+\theta} \right)^\alpha
 \end{aligned}$$

$$\text{where } k = \prod_{i=1}^n (x_i + \theta)^{-1}.$$

$$l(\alpha) = \ln k + n \ln \alpha + \alpha \sum_{i=1}^{n+c} [\ln(d_i + \theta) - \ln(x_i + \theta)]$$

$$l'(\alpha) = 0 + \frac{n}{\alpha} + \sum_{i=1}^{n+c} [\ln(d_i + \theta) - \ln(x_i + \theta)] = 0$$

Thus,

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n+c} [\ln(x_i + \theta) - \ln(d_i + \theta)]}$$

■

Single-Parameter Pareto with fixed θ

$$\begin{aligned} L(\alpha) &= \prod_{i=1}^n \frac{\frac{\alpha \theta^\alpha}{x_i^{\alpha+1}}}{\left[\frac{\theta}{\max(\theta, d_i)}\right]^\alpha} \cdot \prod_{i=n+1}^{n+c} \frac{\left(\frac{\theta}{x_i}\right)^\alpha}{\left[\frac{\theta}{\max(\theta, d_i)}\right]^\alpha} \\ &= \prod_{i=1}^n \frac{\alpha [\max(\theta, d_i)]^\alpha}{x_i^{\alpha+1}} \cdot \prod_{i=n+1}^{n+c} \left[\frac{\max(\theta, d_i)}{x_i}\right]^\alpha \\ &= k \cdot \alpha^n \left[\prod_{i=1}^{n+c} \frac{\max(\theta, d_i)}{x_i} \right]^\alpha \end{aligned}$$

where $k = \prod_{i=1}^n x_i^{-1}$.

$$l(\alpha) = \ln k + n \ln \alpha + \alpha \sum_{i=1}^{n+c} \{\ln[\max(\theta, d_i)] - \ln x_i\}$$

$$l'(\alpha) = 0 + \frac{n}{\alpha} + \sum_{i=1}^{n+c} \{\ln[\max(\theta, d_i)] - \ln x_i\} = 0$$

Thus,

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n+c} \{\ln x_i - \ln[\max(\theta, d_i)]\}}$$

■

Exponential

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \frac{\frac{1}{\theta} e^{-x_i/\theta}}{e^{-d_i/\theta}} \cdot \prod_{i=n+1}^{n+c} \frac{e^{-x_i/\theta}}{e^{-d_i/\theta}} \\ &= \prod_{i=1}^n \frac{1}{\theta} e^{-(x_i-d_i)/\theta} \cdot \prod_{i=n+1}^{n+c} e^{-(x_i-d_i)/\theta} \\ &= \frac{1}{\theta^n} e^{-\sum_{i=1}^{n+c} (x_i-d_i)/\theta} \end{aligned}$$

$$l(\theta) = -n \ln \theta - \frac{\sum_{i=1}^{n+c} (x_i - d_i)}{\theta}$$

$$l'(\theta) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n+c} (x_i - d_i)}{\theta^2} = 0$$

Thus,

$$\hat{\theta} = \frac{\sum_{i=1}^{n+c} (x_i - d_i)}{n}$$



Inverse Exponential

Complete data only.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \frac{\theta e^{-\theta/x_i}}{x_i^2} \\ &= k \cdot \theta^n e^{-\sum_{i=1}^n (\theta/x_i)} \end{aligned}$$

where $k = \prod_{i=1}^n x_i^{-2}$.

$$l(\theta) = \ln k + n \ln \theta - \sum_{i=1}^n \frac{\theta}{x_i}$$

$$l'(\theta) = 0 + \frac{n}{\theta} - \sum_{i=1}^n \frac{1}{x_i} = 0$$

Thus,

$$\hat{\theta} = \frac{n}{\sum_{i=1}^n 1/x_i}$$



Weibull with fixed τ

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^n \frac{\frac{\tau (x_i / \theta)^\tau e^{-(x_i / \theta)^\tau}}{x_i}}{e^{-(d_i / \theta)^\tau}} \cdot \prod_{i=n+1}^{n+c} \frac{e^{-(x_i / \theta)^\tau}}{e^{-(d_i / \theta)^\tau}} \\
&= \prod_{i=1}^n \frac{\tau x_i^{\tau-1} e^{-[(x_i / \theta)^\tau - (d_i / \theta)^\tau]}}{\theta^\tau} \cdot \prod_{i=n+1}^{n+c} e^{-[(x_i / \theta)^\tau - (d_i / \theta)^\tau]} \\
&= k \cdot \theta^{-n\tau} e^{-\sum_{i=1}^{n+c} [(x_i / \theta)^\tau - (d_i / \theta)^\tau]}
\end{aligned}$$

where $k = \prod_{i=1}^n \tau \cdot x_i^{\tau-1}$.

$$\begin{aligned}
l(\theta) &= \ln k - n\tau \ln \theta - \sum_{i=1}^{n+c} \left[\left(\frac{x_i}{\theta} \right)^\tau - \left(\frac{d_i}{\theta} \right)^\tau \right] \\
&= \ln k - n\tau \ln \theta - \frac{\sum_{i=1}^{n+c} (x_i^\tau - d_i^\tau)}{\theta^\tau}
\end{aligned}$$

$$l'(\theta) = 0 - \frac{n\tau}{\theta} + \tau \cdot \frac{\sum_{i=1}^{n+c} (x_i^\tau - d_i^\tau)}{\theta^{\tau+1}} = 0$$

Thus,

$$\hat{\theta} = \left[\frac{\sum_{i=1}^{n+c} (x_i^\tau - d_i^\tau)}{n} \right]^{1/\tau}$$



2.2.0 Overview

 5m

Now we take a step back to consider the theoretical implications of estimation. Our focus shifts away from using **observed** data as in Section 2.1. We start by understanding key terms such as statistic, estimator, and estimate.

Then, we will cover the following properties of an estimator that can aid in evaluating its quality:

- Bias
- Variance
- Mean squared error
- Consistency
- Efficiency
- Minimum-variance unbiased estimator (MVUE)

Finally, we study how the exponential class of distributions and the maximum likelihood estimators fit into the discussion.

2.2.1 Statistics and Estimators

Sample of Random Variables

Imagine taking a sample of size n from a population to study some quantity, such as claim amounts. Thus, there are n random variables, one for each observation's claim amount, before any data is observed. Denote the random variables as X_1, \dots, X_n .

A **random sample** describes a collection of random variables that are independent and identically distributed (i.i.d.).

A **statistic** is a function of random variables from a sample. Among many important statistics to consider, let's start with the following two:

- The **sample mean**

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (2.2.1.1)$$

- The **unbiased sample variance**

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \quad (2.2.1.2)$$

A statistic summarizes the n random variables by mapping them to one value. It is important to note that a statistic is also a random variable (try reflecting on this using Example 1.1.8.1 in Section 1.1.8).

Estimators

An **estimator** is a rule (i.e. a function) that describes how to calculate a parameter estimate. In other words, an estimator is a statistic whose purpose is to estimate a parameter.

The distinction between estimator and estimate can be further illustrated by revisiting several old examples.

Revisiting Example 2.1.1.1

The method of matching moments was used to estimate θ for an exponential distribution.

The **estimator** for this setup is

$$\hat{\theta} = \frac{\sum_{i=1}^n X_i}{n}$$

which is the sample mean, \bar{X} .

Given the data of $n = 33$ students, the **estimate** of θ is 78.788.

Revisiting Example 2.1.3.3

The MLE was used to estimate α for a single-parameter Pareto distribution with $\theta = 1$.

The **estimator** for this setup is

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^n \ln X_i}$$

Given the data of $n = 73$ claims, the **estimate** of α is 1.763.

Unless stated otherwise, we will use θ to denote a generic parameter of interest, and $\hat{\theta}$ to denote a generic estimator of θ . In general, it helps to think of $\hat{\theta}$ as an arbitrary function of X_1, \dots, X_n , and consequently, as a random variable as well.

Coach's Remarks

It is common to use the same symbol $\hat{\theta}$ to denote an estimator as well as an estimate of θ ,

leaving the context to clarify which concept is being referenced. While this can initially become a source of confusion, having a clear understanding of the distinction between estimator and estimate should help.

In addition, this manual uses lowercase letters (e.g. x_i) to denote an observed data point, and uppercase letters (e.g. X_i) to denote a random variable where applicable. However, the exam uses lowercase and uppercase letters interchangeably between observed values and random variables.

2.2.2 Some Estimator Properties

Let's begin by discussing these estimator properties:

- Bias
- Variance
- Mean squared error
- Consistency

Bias

The **bias** of an estimator is calculated as

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta \quad (2.2.2.1)$$

This describes how close the estimator's mean is to the true value of the parameter. $\hat{\theta}$ is an **unbiased estimator** when its bias is 0, i.e. when $E[\hat{\theta}] = \theta$. Otherwise, $\hat{\theta}$ is a biased estimator. Clearly, unbiasedness is a favorable quality, but remember this is only one aspect of an estimator.

Since we desire to estimate θ , remember that its value is typically not known. This means calculating a bias may result in an expression instead of a number. Having said that, an exam question could request a numerical answer by providing a value for θ .

Ultimately, bias reveals whether an estimator is unbiased or not. If the bias is not zero, the resulting expression could tell us something about the behavior of the bias. For example, the expression may contain the sample size n , so if

$$\lim_{n \rightarrow \infty} \text{Bias}[\hat{\theta}] = 0 \quad (2.2.2.2)$$

then $\hat{\theta}$ is **asymptotically unbiased**. Intuitively, this means the bias is less of an issue when a large sample is taken.

Let X_1, X_2, \dots, X_n be a random sample with unknown mean μ . Let the sample mean, \bar{X} , be an estimator of μ .

Determine if \bar{X} is an unbiased estimator of μ .

To calculate the bias of \bar{X} , first calculate its mean.

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{\sum_{i=1}^n X_i}{n}\right] \\ &= \frac{E[\sum_{i=1}^n X_i]}{n} \\ &= \frac{E[X_1] + E[X_2] + \dots + E[X_n]}{n} \\ &= \frac{\mu + \mu + \dots + \mu}{n} \\ &= \frac{n\mu}{n} \\ &= \mu \end{aligned}$$

Since $E[\bar{X}] = \mu$,

$$\begin{aligned} \text{Bias}[\bar{X}] &= E[\bar{X}] - \mu \\ &= \mu - \mu \\ &= 0 \end{aligned}$$

Therefore, \bar{X} is an unbiased estimator of μ .

Coach's Remarks

It is important to know the following probability fact:

For a random sample of size n ,

$$\mathbb{E}[\bar{X}] = \mathbb{E}[X], \quad \text{Var}[\bar{X}] = \frac{\text{Var}[X]}{n}$$

From this fact, it is apparent that \bar{X} is an unbiased estimator of the mean for a random sample. We mainly went through the steps to illustrate how to calculate bias in general.

The next two motivating examples may seem challenging, but their primary purpose is to highlight a few key points that will be presented after working through them.

Let X_1, X_2, \dots, X_n be a random sample with **known** mean μ and unknown variance σ^2 . Define $\hat{\sigma}^2$ as an estimator of σ^2 , where

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n}$$

Determine if $\hat{\sigma}^2$ is an unbiased estimator of σ^2 .

Before finding the expected value of $\hat{\sigma}^2$, rewrite the estimator so that it is easier to handle going forward.

$$\begin{aligned} \frac{\sum_{i=1}^n (X_i - \mu)^2}{n} &= \frac{\sum_{i=1}^n (X_i^2 - 2\mu X_i + \mu^2)}{n} \\ &= \frac{\sum_{i=1}^n X_i^2}{n} - \frac{2\mu (\sum_{i=1}^n X_i)}{n} + \frac{n\mu^2}{n} \\ &= \frac{\sum_{i=1}^n X_i^2}{n} - 2\mu \bar{X} + \mu^2 \end{aligned}$$

Moreover, recall that

$$\begin{aligned} \text{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ \Rightarrow \mathbb{E}[X^2] &= \text{Var}[X] + \mathbb{E}[X]^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

As a result,

$$\begin{aligned}
 E[\hat{\sigma}^2] &= E\left[\frac{\sum_{i=1}^n X_i^2}{n} - 2\mu\bar{X} + \mu^2\right] \\
 &= \frac{E[\sum_{i=1}^n X_i^2]}{n} - 2\mu E[\bar{X}] + E[\mu^2] \\
 &= \frac{n(\sigma^2 + \mu^2)}{n} - 2\mu(\mu) + \mu^2 \\
 &= \sigma^2
 \end{aligned}$$

Since $E[\hat{\sigma}^2] = \sigma^2$, the bias of $\hat{\sigma}^2$ is 0. Therefore, **$\hat{\sigma}^2$ is an unbiased estimator of σ^2** .

Notice that we unrealistically assumed that μ was known and proceeded to use it in our estimator. Let's see what happens if we use the sample mean, \bar{X} , as a substitute for an unknown μ .

Let X_1, X_2, \dots, X_n be a random sample with unknown mean μ and unknown variance σ^2 . Define $\hat{\sigma}^2$ as an estimator of σ^2 , where

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

Determine if $\hat{\sigma}^2$ is an unbiased estimator of σ^2 .

Borrowing from our work in the previous example, note that

$$\begin{aligned}
 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} &= \frac{\sum_{i=1}^n X_i^2}{n} - 2\bar{X}\bar{X} + \bar{X}^2 \\
 &= \frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2
 \end{aligned}$$

As a result,

$$\begin{aligned}
E[\hat{\sigma}^2] &= E\left[\frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2\right] \\
&= \frac{E[\sum_{i=1}^n X_i^2]}{n} - E[\bar{X}^2] \\
&= \frac{E[\sum_{i=1}^n X_i^2]}{n} - \left(\text{Var}[\bar{X}] + E[\bar{X}]^2\right) \\
&= \frac{n(\sigma^2 + \mu^2)}{n} - \frac{\sigma^2}{n} - \mu^2 \\
&= \sigma^2 - \frac{\sigma^2}{n} \\
&= \sigma^2 \left(\frac{n-1}{n}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{Bias}[\hat{\sigma}^2] &= E[\hat{\sigma}^2] - \sigma^2 \\
&= \sigma^2 \left(\frac{n-1}{n}\right) - \sigma^2 \\
&= -\frac{\sigma^2}{n} \neq 0
\end{aligned}$$

and $\hat{\sigma}^2$ is a biased estimator of σ^2 .

Here are key conclusions from the two motivating examples:

- The negative bias suggests that the average estimate would be lower than the true value of σ^2 . On the other hand, it is an asymptotically unbiased estimator of σ^2 ; as $n \rightarrow \infty$, the bias converges to 0.
- Notice that the biased estimator would be unbiased if multiplied by $\frac{n}{n-1}$.

$$E\left[\frac{n}{n-1}\hat{\sigma}^2\right] = \frac{n}{n-1}E[\hat{\sigma}^2] = \frac{n}{n-1} \cdot \sigma^2 \left(\frac{n-1}{n}\right) = \sigma^2$$

In addition, this unbiased estimator is

$$\frac{n}{n-1} \cdot \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = S^2$$

i.e. Equation 2.2.1.2, hence the name "**unbiased** sample variance".

- The use of \bar{X} in place of μ is where the bias emerges. This substitution imposes one mathematical constraint on the X_i 's, which relates to how a denominator of $n - 1$ would adjust the bias to 0.

Variance

By definition, variance measures the dispersion of a random variable from its mean. In math notation, the variance of $\hat{\theta}$ is

$$\text{Var}[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

Hence, an estimator with a larger variance suggests that the estimate is prone to be a value further from the estimator's mean. As a result, even an unbiased estimator may be unhelpful when the variance is large.

Similar to calculating the mean, keep in mind that the variance of an estimator could result in an expression instead of a number.

Let X_1, X_2, \dots, X_n be a random sample with unknown mean μ and unknown variance σ^2 . Define $\hat{\mu}$ as an estimator of μ , where

$$\hat{\mu} = \frac{X_1 + X_n}{2}$$

Determine

- if $\hat{\mu}$ is an unbiased estimator of μ .
- the variance of $\hat{\mu}$.

For Part (1),

$$\begin{aligned}
 E[\hat{\mu}] &= E\left[\frac{X_1 + X_n}{2}\right] \\
 &= \frac{E[X_1 + X_n]}{2} \\
 &= \frac{E[X_1] + E[X_n]}{2} \\
 &= \frac{\mu + \mu}{2} \\
 &= \frac{2\mu}{2} \\
 &= \mu
 \end{aligned}$$

Since $E[\hat{\mu}] = \mu$, the bias of $\hat{\mu}$ is 0. Therefore, $\hat{\mu}$ is an unbiased estimator of μ .

For Part (2),

$$\begin{aligned}
 \text{Var}[\hat{\mu}] &= \text{Var}\left[\frac{X_1 + X_n}{2}\right] \\
 &= \frac{\text{Var}[X_1 + X_n]}{2^2} \\
 &= \frac{\text{Var}[X_1] + \text{Var}[X_n]}{4} \\
 &= \frac{\sigma^2 + \sigma^2}{4} \\
 &= \frac{2\sigma^2}{4} \\
 &= \frac{\sigma^2}{2}
 \end{aligned}$$

Coach's Remarks

Recall that the variance of \bar{X} is $\frac{\sigma^2}{n}$ for a random sample. With n exceeding 2 in this example, note that \bar{X} would have a smaller variance than $\hat{\mu}$, even though both are unbiased estimators of μ . In that aspect, \bar{X} would be preferable over $\hat{\mu}$.

This supports the intuition that the average of all n observations should do better at estimating the population mean, compared to the average of only the first and last observations (or any pair of observations for that matter).

Mean Squared Error (MSE)

The **mean squared error** is also commonly used to determine the quality of an estimator. By definition,

$$\text{MSE}[\hat{\theta}] = E\left[\left(\hat{\theta} - \theta\right)^2\right] \quad (2.2.2.3)$$

Notice the definition of a variance is similar to the formula above, where the parameter θ is replaced with the mean of $\hat{\theta}$. Thus, the MSE and variance of an estimator are similar measures; the former measures the average squared deviation **from the true parameter value**, while the latter measures the average squared deviation **from the estimator's mean**.

An alternative formula for the MSE is

$$\text{MSE}[\hat{\theta}] = \text{Var}[\hat{\theta}] + \left(\text{Bias}[\hat{\theta}]\right)^2 \quad (2.2.2.4)$$

The relation between MSE and variance is also seen in this formula. As mentioned, the distinction is whether θ or $E[\hat{\theta}]$ is the "point of reference" from which the squared deviation is measured.

Therefore, when the estimator's bias is 0 (i.e. $E[\hat{\theta}] = \theta$), the MSE will equal the variance; they have the same "point of reference".

$$\text{MSE}[\hat{\theta}] = \text{Var}[\hat{\theta}] + 0^2 = \text{Var}[\hat{\theta}]$$

Example 2.2.2.1

For independent random variables X_1, X_2, X_3 ,

- $E[X_i] = \frac{5-i}{4}\theta, \quad i = 1, 2, 3$
- $\text{Var}[X_i] = (i+1)\theta^2, \quad i = 1, 2, 3$
- An estimator for θ is

$$\hat{\theta} = \frac{4X_1 + 3X_2 + 2X_3}{9}$$

For $\theta > 0$, determine the absolute difference in the lowest and highest values of $\hat{\theta}$ if the MSE of $\hat{\theta}$ is less than 100.

Solution

Start by calculating the mean of $\hat{\theta}$.

$$\begin{aligned} E[\hat{\theta}] &= E\left[\frac{4X_1 + 3X_2 + 2X_3}{9}\right] \\ &= \frac{4 \cdot E[X_1] + 3 \cdot E[X_2] + 2 \cdot E[X_3]}{9} \\ &= \frac{4\theta + 3(0.75\theta) + 2(0.5\theta)}{9} \\ &= \frac{29}{36}\theta \end{aligned}$$

Next, find the variance of $\hat{\theta}$.

$$\begin{aligned} \text{Var}[\hat{\theta}] &= \text{Var}\left[\frac{4X_1 + 3X_2 + 2X_3}{9}\right] \\ &= \frac{4^2 \cdot \text{Var}[X_1] + 3^2 \cdot \text{Var}[X_2] + 2^2 \cdot \text{Var}[X_3]}{9^2} \\ &= \frac{4^2(2\theta^2) + 3^2(3\theta^2) + 2^2(4\theta^2)}{9^2} \\ &= \frac{25}{27}\theta^2 \end{aligned}$$

Therefore,

$$\begin{aligned}
 \text{MSE}[\hat{\theta}] &= \text{Var}[\hat{\theta}] + (\text{Bias}[\hat{\theta}])^2 \\
 &= \frac{25}{27}\theta^2 + \left(\frac{29}{36}\theta - \theta\right)^2 \\
 &= \frac{25}{27}\theta^2 + \left(\frac{29}{36} - 1\right)^2\theta^2 \\
 &= 0.9637\theta^2
 \end{aligned}$$

$$\begin{aligned}
 \text{MSE}[\hat{\theta}] &< 100 \\
 \Rightarrow 0.9637\theta^2 &< 100 \\
 \Rightarrow \theta^2 &< \frac{100}{0.9637} \\
 \Rightarrow -\sqrt{\frac{100}{0.9637}} &< \theta < \sqrt{\frac{100}{0.9637}}
 \end{aligned}$$

Since θ must be positive, the actual interval is $0 < \theta < 10.19$. Thus, the absolute difference in the lowest and highest values of θ when the MSE is under 100 is **10.19**.



Example 2.2.2.2

For independent random variables X_1, X_2, X_3 ,

- θ is a parameter whose true value is 4.

- $E[X_i] = \frac{5-i}{4}\theta, \quad i = 1, 2, 3$

- $\text{Var}[X_i] = (i+1)\theta^2, \quad i = 1, 2, 3$

- Two estimators for θ are

$$\hat{\theta}_1 = \frac{X_1 + X_2}{2}, \quad \hat{\theta}_2 = X_3$$

Consider estimators of θ that have the form: $w\hat{\theta}_1 + (1 - w)\hat{\theta}_2$.

Calculate the w that causes the estimator to have the smallest mean squared error.

Solution

Start by calculating the mean and variance of $w\hat{\theta}_1 + (1 - w)\hat{\theta}_2$. Keep in mind that the covariance of $\hat{\theta}_1$ and $\hat{\theta}_2$ is 0 since the X_i 's are independent.

$$\begin{aligned} E[w\hat{\theta}_1 + (1 - w)\hat{\theta}_2] &= w \cdot E[\hat{\theta}_1] + (1 - w) \cdot E[\hat{\theta}_2] \\ &= w \cdot E\left[\frac{X_1 + X_2}{2}\right] + (1 - w) \cdot E[X_3] \\ &= \frac{w}{2}(\theta + 0.75\theta) + (1 - w)(0.5\theta) \\ &= 3.5w + 2(1 - w) \\ &= 1.5w + 2 \end{aligned}$$

$$\begin{aligned} \text{Var}[w\hat{\theta}_1 + (1 - w)\hat{\theta}_2] &= w^2 \cdot \text{Var}[\hat{\theta}_1] + (1 - w)^2 \cdot \text{Var}[\hat{\theta}_2] \\ &= w^2 \cdot \text{Var}\left[\frac{X_1 + X_2}{2}\right] + (1 - w)^2 \cdot \text{Var}[X_3] \\ &= \frac{w^2}{4}(2\theta^2 + 3\theta^2) + (1 - w)^2(4\theta^2) \\ &= 20w^2 + 64(1 - w)^2 \\ &= 84w^2 - 128w + 64 \end{aligned}$$

Therefore,

$$\begin{aligned} \text{MSE}[\hat{\theta}] &= \text{Var}[\hat{\theta}] + (\text{Bias}[\hat{\theta}])^2 \\ &= 84w^2 - 128w + 64 + (1.5w + 2 - 4)^2 \\ &= 86.25w^2 - 134w + 68 \end{aligned}$$

To minimize the MSE with respect to w , take its first derivative with respect to w and set it equal to 0. Solving for w gives

$$\frac{d}{dw} (86.25w^2 - 134w + 68) = 172.5w - 134 = 0$$

$$w = \frac{134}{172.5} = \mathbf{0.777}$$

■

Consistency

Another way to examine an estimator is to see if it is *consistent*. By definition, $\hat{\theta}$ is a consistent estimator of θ if

$$\lim_{n \rightarrow \infty} \Pr\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) = 0 \quad (2.2.2.5)$$

for all $\varepsilon > 0$. Consistency assesses whether $\hat{\theta}$ converges (in probability) to θ when the sample size approaches ∞ .

Let X_1, X_2, \dots, X_n be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2 . Let the sample mean, \bar{X} , be an estimator of μ .

Using the given definition, determine if \bar{X} is a consistent estimator of μ .

Since the linear combination of independent normal random variables is also normally distributed, \bar{X} follows a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. Therefore,

$$\begin{aligned}
\Pr(|\bar{X} - \mu| > \varepsilon) &= \Pr(\bar{X} - \mu < -\varepsilon) + \Pr(\bar{X} - \mu > \varepsilon) \\
&= \Pr\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < -\frac{\varepsilon}{\sigma/\sqrt{n}}\right) + \Pr\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{\varepsilon}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\right) + \left[1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right] \\
&= 2 \cdot \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\right)
\end{aligned}$$

$$\lim_{n \rightarrow \infty} \left[2 \cdot \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\right) \right] = 0, \text{ for all } \varepsilon > 0$$

Thus, \bar{X} is a consistent estimator of μ .

In the context of this exam, using the definition to prove consistency tends to be challenging. On the other hand, if

- $\hat{\theta}$ is asymptotically unbiased, and
- $\lim_{n \rightarrow \infty} \text{Var}[\hat{\theta}] = 0$

then $\hat{\theta}$ is consistent. Otherwise, it is inconclusive.

Coach's Remarks

The two sufficient conditions for consistency can be described as a single sufficient condition:

$$\lim_{n \rightarrow \infty} \text{MSE}[\hat{\theta}] = 0$$

Furthermore, the method of moments estimators are consistent as long as they are also unique.

Example 2.2.2.3

Let X_1, X_2, \dots, X_n be a random sample with unknown mean β and unknown variance $7\beta^2$. Let the sample mean, \bar{X} , be an estimator of β .

Determine whether \bar{X} is consistent.

Solution

Realize that \bar{X} is also the method of moments estimator for β , i.e. \bar{X} is the solution to β in the equation

$$\mathbb{E}[X] = \frac{\sum_{i=1}^n X_i}{n}$$

Since \bar{X} uniquely solves the equation, conclude that **\bar{X} is consistent.**



Alternative Solution

Given the random sample, we already know that \bar{X} is an unbiased estimator of the mean β , which also implies that \bar{X} is asymptotically unbiased. Mathematically,

$$\lim_{n \rightarrow \infty} \text{Bias}[\bar{X}] = \lim_{n \rightarrow \infty} 0 = 0$$

Also, note that the variance of \bar{X} is $\frac{7\beta^2}{n}$ given the random sample. Thus,

$$\lim_{n \rightarrow \infty} \text{Var}[\bar{X}] = \lim_{n \rightarrow \infty} \frac{7\beta^2}{n} = 0$$

With those two facts established, conclude that \bar{X} is consistent.



Example 2.2.2.4

Determine which statements regarding estimators are true.

- I. The MSE is the expected squared difference between the estimator and its mean.
- II. If $\hat{\theta}$ is unbiased, then it is also consistent if $\lim_{n \rightarrow \infty} \text{Var}[\hat{\theta}] = 0$.
- III. For unbiased estimators, their MSE will equal their variance.
- IV. Generally speaking, an estimator with lower MSE is preferred.

Solution

I is false because the MSE is the expected squared difference between the estimator and the true parameter value. The given description actually refers to the variance.

II is true because if an estimator is asymptotically unbiased and its variance approaches 0 as the sample size approaches ∞ , then it is consistent. If an estimator is unbiased, then it is also asymptotically unbiased.

III is true because $\text{MSE}[\hat{\theta}] = \text{Var}[\hat{\theta}] + (\text{Bias}[\hat{\theta}])^2 = \text{Var}[\hat{\theta}] + 0^2 = \text{Var}[\hat{\theta}]$.

IV is true because it is desirable to expect a low squared difference between the estimator and the true parameter value. In addition, MSE is the sum of the estimator's variance and squared bias, where it is desirable to minimize both quantities.

Therefore, **only II, III, and IV are true.**



2.2.3 Rao-Cramér Lower Bound

🕒 30m

This subsection focuses on the estimator property of efficiency. To that end, we must first learn of the Fisher information.

Fisher Information

ONE-PARAMETER CASE

The *Fisher information*, or simply *information*, of θ is

$$I(\theta) = \text{Var}[l'(\theta)] \quad (2.2.3.1)$$

Loosely speaking, information measures how sensitive $l(\theta)$ is to data. High information suggests that the shape of $l(\theta)$ is sensitive to observable data; this is captured or expressed through a high variation in $l'(\theta)$.

An equivalent way of calculating information is

$$I(\theta) = -\mathbb{E}[l''(\theta)] \quad (2.2.3.2)$$

which is more often preferred over computing a variance. The appendix at the end of this section proves their equivalence.

In this setup, $l''(\theta)$ is a random variable because we have the random sample X_1, \dots, X_n (as opposed to the observed values x_1, \dots, x_n). Additionally, note that

$$\begin{aligned} l(\theta) &= \ln [L(\theta)] \\ &= \ln \left[\prod_{i=1}^n f(X_i) \right] \\ &= \sum_{i=1}^n \ln [f(X_i)] \end{aligned}$$

$$l''(\theta) = \frac{d^2}{d\theta^2} \left\{ \sum_{i=1}^n \ln [f(X_i)] \right\} = \sum_{i=1}^n \frac{d^2}{d\theta^2} \ln [f(X_i)]$$

Therefore, it is convenient to first work with one X , then multiply the result by n , i.e.

$$\begin{aligned} I(\theta) &= -E \left[\sum_{i=1}^n \frac{d^2}{d\theta^2} \ln [f(X_i)] \right] \\ &= - \sum_{i=1}^n E \left[\frac{d^2}{d\theta^2} \ln [f(X_i)] \right] \\ &= -n \cdot E \left[\frac{d^2}{d\theta^2} \ln [f(X)] \right] \end{aligned} \tag{2.2.3.3}$$

Let's see the formula in action. Consider the following example:

A random sample is drawn from a distribution with density function:

$$f(x) = \frac{x^{\alpha-1} e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}, \quad x > 0$$

where α is known.

Determine the Fisher information of θ .

To use Equation 2.2.3.3, first take the natural log of the density function, then find its second derivative with respect to θ .

$$\begin{aligned} \ln [f(x)] &= (\alpha - 1) \ln x - \frac{x}{\theta} - \ln \Gamma(\alpha) - \alpha \ln \theta \\ \frac{d}{d\theta} \ln [f(x)] &= 0 + \frac{x}{\theta^2} - 0 - \frac{\alpha}{\theta} \\ \frac{d^2}{d\theta^2} \ln [f(x)] &= -\frac{2x}{\theta^3} + \frac{\alpha}{\theta^2} \end{aligned}$$

Then,

$$\begin{aligned}
 I(\theta) &= -n \cdot E \left[\frac{d^2}{d\theta^2} \ln [f(X)] \right] \\
 &= -n \cdot E \left[-\frac{2X}{\theta^3} + \frac{\alpha}{\theta^2} \right] \\
 &= n \left(\frac{2}{\theta^3} E[X] - \frac{\alpha}{\theta^2} \right)
 \end{aligned}$$

Recognize the density function as belonging to a gamma distribution. Since gamma's mean is $\alpha\theta$, solve for the Fisher information as

$$\begin{aligned}
 I(\theta) &= n \left(\frac{2}{\theta^3} [\alpha\theta] - \frac{\alpha}{\theta^2} \right) \\
 &= \frac{n\alpha}{\theta^2}
 \end{aligned}$$

Coach's Remarks

Formulating the information relies on a set of assumptions commonly referred to as "regularity conditions". Hence, $I(\theta)$ is not valid if one of the regularity conditions is violated. We do not expect this exam to cover these conditions.

However, one of the conditions is that the domain of $f(x)$ must be the same for all values of θ . So, note that all concepts related to information do not apply to such cases, e.g. a uniform distribution valid on the interval $[0, \theta]$.

Coach's Remarks

You can solve for the information of $g(\theta)$ as

$$I(\theta) \cdot g'(\theta)^{-2}$$

MULTI-PARAMETER CASE

If X 's distribution involves more than one parameter, then the Fisher information of the parameters refers to a matrix instead. Let's start by considering the case with two parameters, generically denoted as θ_1 and θ_2 .

The entries of the information matrix mimic the form of Equation 2.2.3.3. Specifically, the i^{th} row, j^{th} column entry of the information matrix is

$$-n \cdot E \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln [f(X)] \right]$$

where i and j are either 1 or 2. Consequently, the information matrix of θ_1 and θ_2 is

$$\mathbf{I}(\theta_1, \theta_2) = \begin{bmatrix} -n \cdot E \left[\frac{\partial^2}{\partial \theta_1^2} \ln [f(X)] \right] & -n \cdot E \left[\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \ln [f(X)] \right] \\ -n \cdot E \left[\frac{\partial^2}{\partial \theta_2 \partial \theta_1} \ln [f(X)] \right] & -n \cdot E \left[\frac{\partial^2}{\partial \theta_2^2} \ln [f(X)] \right] \end{bmatrix}$$

As a reminder, order is interchangeable when taking partial derivatives, i.e. $\partial \theta_1 \partial \theta_2$ is equivalent to $\partial \theta_2 \partial \theta_1$. Therefore, the two off-diagonal entries of $\mathbf{I}(\theta_1, \theta_2)$ are the same.

It is then straightforward to extend to the case with p parameters. The information matrix of $\theta_1, \dots, \theta_p$ is a $p \times p$ matrix with entries that mimic the form of Equation 2.2.3.3. The i^{th} row, j^{th} column entry is found by taking partial derivatives with respect to θ_i and θ_j , where i and j are any integers from 1 to p .

Rao-Cramér Lower Bound

The Fisher information is useful in several ways; we now discuss how it relates to the **Rao-Cramér lower bound**. For the case of one parameter θ with an unbiased estimator $\hat{\theta}$, it is known that the variance of $\hat{\theta}$ has a lower bound, i.e.

$$\text{Var}[\hat{\theta}] \geq \frac{1}{I(\theta)} \tag{2.2.3.4}$$

The same idea holds when there are p parameters $\theta_1, \dots, \theta_p$. Let $\hat{\theta}_j$ be an unbiased estimator of θ_j , for $j = 1, \dots, p$. Then, the Rao-Cramér inequality is

$$\text{Var}[\hat{\theta}_j] \geq j^{\text{th}} \text{ diagonal entry of } \mathbf{I}^{-1} \quad (2.2.3.5)$$

where \mathbf{I}^{-1} is the inverse of the information matrix $\mathbf{I}(\theta_1, \dots, \theta_p)$. In addition, we encourage you to think about how Equation 2.2.3.4 is the special case of $p = 1$ which operates under exactly the same mechanics.

Keep in mind that the information matrix can be expressed as n times the information matrix for a sample size of 1. As a result, \mathbf{I}^{-1} can be found by inverting the information matrix for a sample size of 1, then dividing each entry by n .

As the variance of an estimator cannot be smaller than the Rao-Cramér lower bound, it is desirable if the variance of an unbiased $\hat{\theta}$ **equals** the Rao-Cramér lower bound. If the variance attains the lower bound, then $\hat{\theta}$ is known as an **efficient estimator**.

The **efficiency** of an unbiased estimator of θ is

$$\text{Eff}[\hat{\theta}] = \frac{1/I(\theta)}{\text{Var}[\hat{\theta}]} \quad (2.2.3.6)$$

or more generally, the Rao-Cramér lower bound divided by the estimator's variance. Hence, an efficient estimator has an efficiency of 1.

Example 2.2.3.1

A distribution has three parameters: α, β, γ . You are given:

- For a single random variable, the inverse of the information matrix of α, β, γ is
$$\begin{bmatrix} 25 & 0 & -12 \\ 0 & 6 & 9 \\ -12 & 9 & 13 \end{bmatrix}.$$
- Based on a random sample of size 20, $\hat{\gamma}$ is an unbiased estimator of γ with a variance of 2.

Calculate the efficiency of $\hat{\gamma}$.

Solution

Since we are given the inverse of the information matrix for a sample size of 1, divide it by $n = 20$ in order to find the inverse of $\mathbf{I}(\alpha, \beta, \gamma)$.

$$\mathbf{I}^{-1} = \frac{1}{20} \begin{bmatrix} 25 & 0 & -12 \\ 0 & 6 & 9 \\ -12 & 9 & 13 \end{bmatrix} = \begin{bmatrix} 1.25 & 0 & -0.60 \\ 0 & 0.30 & 0.45 \\ -0.60 & 0.45 & 0.65 \end{bmatrix}$$

Because we are interested in an unbiased estimator of γ (i.e. the third parameter in sequence), we want the 3rd diagonal entry of \mathbf{I}^{-1} . This means that 0.65 is the Rao-Cramér lower bound for the variance of an unbiased estimator of γ .

Thus, the answer is

$$\begin{aligned} \text{Eff}[\hat{\gamma}] &= \frac{\text{Rao-Cramér lower bound}}{\text{Var}[\hat{\gamma}]} \\ &= \frac{0.65}{2} \\ &= \mathbf{0.325} \end{aligned}$$



Example 2.2.3.2

A random sample of size n is taken from a normal distribution with mean μ and variance σ^2 .

Determine the variance of an efficient estimator of σ^2 .

Solution

By definition, an efficient estimator's variance equals the Rao-Cramér lower bound. Therefore, we seek the Rao-Cramér lower bound for the variance of an unbiased estimator of σ^2 . This lower bound is the 2nd diagonal entry of \mathbf{I}^{-1} , i.e. the inverse of the information matrix $\mathbf{I}(\mu, \sigma^2)$. The immediate goal is to determine the information matrix. Recall that the normal density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

First take the natural log of the density, then find the relevant partial derivatives. Since our interest is in σ^2 rather than σ , we should express $f(x)$ in terms of σ^2 before proceeding.

$$\ln[f(x)] = -\frac{1}{2}\ln(\sigma^2) - \ln(\sqrt{2\pi}) - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\begin{aligned}\frac{\partial}{\partial\mu}\ln[f(x)] &= -0 - 0 + \frac{x-\mu}{\sigma^2} \\ \frac{\partial}{\partial\sigma^2}\ln[f(x)] &= -\frac{1}{2\sigma^2} - 0 + \frac{(x-\mu)^2}{2\sigma^4}\end{aligned}$$

$$\begin{aligned}\frac{\partial^2}{\partial\mu^2}\ln[f(x)] &= -\frac{1}{\sigma^2} \\ \frac{\partial^2}{\partial\sigma^2\partial\mu}\ln[f(x)] &= -\frac{x-\mu}{\sigma^4} \\ \frac{\partial^2}{(\partial\sigma^2)^2}\ln[f(x)] &= \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}\end{aligned}$$

Next, calculate the negative expected values of the second partial derivatives.

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\mu^2}\ln[f(X)]\right] = \frac{1}{\sigma^2}$$

$$-\mathbb{E} \left[\frac{\partial^2}{\partial \sigma^2 \partial \mu} \ln [f(X)] \right] = \frac{\overbrace{\mathbb{E}[X] - \mu}^{\mu}}{\sigma^4} = 0$$

$$-\mathbb{E} \left[\frac{\partial^2}{(\partial \sigma^2)^2} \ln [f(X)] \right] = -\frac{1}{2\sigma^4} + \frac{\overbrace{\mathbb{E}[(X - \mu)^2]}^{\sigma^2}}{\sigma^6} = \frac{1}{2\sigma^4}$$

As a result, the information matrix is

$$\begin{aligned} \mathbf{I}(\mu, \sigma^2) &= \begin{bmatrix} -n \cdot \mathbb{E} \left[\frac{\partial^2}{\partial \mu^2} \ln [f(X)] \right] & -n \cdot \mathbb{E} \left[\frac{\partial^2}{\partial \sigma^2 \partial \mu} \ln [f(X)] \right] \\ -n \cdot \mathbb{E} \left[\frac{\partial^2}{\partial \sigma^2 \partial \mu} \ln [f(X)] \right] & -n \cdot \mathbb{E} \left[\frac{\partial^2}{(\partial \sigma^2)^2} \ln [f(X)] \right] \end{bmatrix} \\ &= n \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \end{aligned}$$

Now we must invert the information matrix. First note that

$$\begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}^{-1} = \frac{1}{\frac{1}{2\sigma^4} - 0} \begin{bmatrix} \frac{1}{2\sigma^4} & 0 \\ 0 & \frac{1}{\sigma^2} \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}$$

For a refresher on inverting a 2×2 matrix, review Section 1.6.5.

So the inverse of the information matrix is

$$\mathbf{I}^{-1} = \frac{1}{n} \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

In conclusion, the Rao-Cramér lower bound for the variance of an unbiased estimator of σ^2 is $\frac{2\sigma^4}{n}$.



Coach's Remarks

For extra credit, consider the fact that S^2 is an unbiased estimator of σ^2 . Without proving it, the variance of this unbiased estimator is

$$\text{Var}[S^2] = \frac{2\sigma^4}{n-1}$$

Therefore, S^2 is **not** an efficient estimator because its variance did not attain the Rao-Cramér lower bound we just found; its efficiency is $\frac{n-1}{n}$.

On the other hand, we have also shown that \bar{X} is an efficient estimator of μ ; spend some time to think that through.

2.2.4 Minimum Variance Unbiased Estimator

If we prioritize the properties of unbiasedness and low variance, then we would desire an efficient estimator. However, an efficient estimator may not exist for certain contexts, i.e. it may be impossible for the variance of an unbiased estimator to attain the ideal Rao-Cramér lower bound.

Hence, the **minimum variance unbiased estimator (MVUE)** would be the next best option. It is an unbiased estimator with the smallest variance among all unbiased estimators, regardless of the true parameter value.

To be clear, the MVUE **does not** necessarily have the smallest variance of all estimators; it is only smallest among the **unbiased** ones. Therefore, the variance of the MVUE could possibly be larger than the variance of a biased estimator.

With Y denoting a statistic of the random sample X_1, \dots, X_n , the Lehmann-Scheffé theorem identifies three conditions that lead to a unique MVUE:

1. Y is a sufficient statistic for θ .
2. The distribution of Y comes from a complete family of distributions.
3. There is a function of Y , $\varphi(Y)$, that is an unbiased estimator of θ .

When all three conditions are met, $\varphi(Y)$ is the MVUE of θ . Collectively, the first two conditions can be described as Y being a **complete sufficient statistic**. We will study these conditions in detail.

Sufficiency

Y is a **sufficient statistic** for θ if and only if

$$f(x_1, \dots, x_n | y) = h(x_1, \dots, x_n) \quad (2.2.4.1)$$

where $h(x_1, \dots, x_n)$ does not depend on θ . The equation says that conditioning on the value of Y leads to a joint distribution of the sample that is unaffected by θ . Intuitively, this means knowing $Y = y$ would thoroughly or sufficiently capture how θ impacts the sample, making it unnecessary to know each of the n observed values.

Coach's Remarks

Similar to the approach in Section 2.1.3, we use $f(x)$ as a probability function applicable to

either discrete or continuous distributions. Furthermore, since we have independent X_i 's, the conditional probability function can be written as

$$f(x_1, \dots, x_n | y) = \frac{\prod_{i=1}^n f(x_i)}{f_Y(y)}$$

with x_1, \dots, x_n subject to y . But note that the definition of sufficiency does not require the i.i.d. assumption.

Also, it is important note that sufficiency alone does not imply unbiasedness.

Example 2.2.4.1

A random sample of size n is taken from a Bernoulli distribution with success probability q , where $n > 2$.

Determine which of the following statistics are sufficient for q :

1. The sum of the entire sample.
2. The sum of the first two random variables of the sample.

Solution to (1)

Let $Y = \sum_{i=1}^n X_i$. Since Y is the sum of n i.i.d. Bernoulli random variables, Y follows a binomial distribution, i.e.

$$p_Y(y) = \binom{n}{y} q^y (1-q)^{n-y}, \quad y = 0, 1, \dots, n$$

As a result,

$$\begin{aligned}
p(x_1, \dots, x_n | y) &= \frac{\prod_{i=1}^n p(x_i)}{p_Y(y)} \\
&= \frac{\prod_{i=1}^n q^{x_i} (1-q)^{1-x_i}}{\binom{n}{y} q^y (1-q)^{n-y}} \\
&= \frac{q^{\sum_{i=1}^n x_i} (1-q)^{n-\sum_{i=1}^n x_i}}{\left(\sum_{i=1}^n x_i\right) q^{\sum_{i=1}^n x_i} (1-q)^{n-\sum_{i=1}^n x_i}} \\
&= \left(\frac{n}{\sum_{i=1}^n x_i}\right)^{-1} \\
&= h(x_1, \dots, x_n)
\end{aligned}$$

Keep in mind that $y = \sum_{i=1}^n x_i$. Since $\left(\frac{n}{\sum_{i=1}^n x_i}\right)^{-1}$ does not depend on q , conclude that $\sum_{i=1}^n X_i$ is a sufficient statistic for q .



Solution to (2)

Let $Y = X_1 + X_2$. Since Y is the sum of two i.i.d. Bernoulli random variables, Y follows a binomial distribution, i.e.

$$p_Y(y) = \binom{2}{y} q^y (1-q)^{2-y}, \quad y = 0, 1, 2$$

As a result,

$$\begin{aligned}
p(x_1, \dots, x_n | y) &= \frac{\prod_{i=1}^n p(x_i)}{p_Y(y)} \\
&= \frac{\prod_{i=1}^n q^{x_i} (1-q)^{1-x_i}}{\binom{2}{y} q^y (1-q)^{2-y}} \\
&= \frac{q^{\sum_{i=1}^n x_i} (1-q)^{n-\sum_{i=1}^n x_i}}{\binom{2}{x_1+x_2} q^{x_1+x_2} (1-q)^{2-(x_1+x_2)}} \\
&= \binom{2}{x_1+x_2}^{-1} q^{\sum_{i=3}^n x_i} (1-q)^{n-2-\sum_{i=3}^n x_i} \\
&\neq h(x_1, \dots, x_n)
\end{aligned}$$

Keep in mind that $y = x_1 + x_2$. Because the conditional PMF depends on q , conclude that **$X_1 + X_2$ is not a sufficient statistic for q** .



Example 2.2.4.2

A random sample of size n is taken from an exponential distribution with mean θ .

Determine whether the sample mean is a sufficient statistic for θ .

Solution

Let $Y = \bar{X}$. Since Y is the average of n i.i.d. exponential random variables, Y follows a gamma distribution. We use the fact that

- the sum of i.i.d. exponential random variables is a gamma random variable, and
- a scaled gamma random variable is also a gamma random variable.

Therefore,

$$Y \sim \text{Gamma} \left(n, \frac{\theta}{n} \right)$$

$$f_Y(y) = \frac{y^{n-1} e^{-yn/\theta}}{\Gamma(n) \left(\frac{\theta}{n}\right)^n}, \quad y > 0$$

As a result,

$$\begin{aligned} f(x_1, \dots, x_n | y) &= \frac{\prod_{i=1}^n f(x_i)}{f_Y(y)} \\ &= \left[\prod_{i=1}^n \frac{e^{-x_i/\theta}}{\theta} \right] \div \left[\frac{y^{n-1} e^{-yn/\theta}}{\Gamma(n) \left(\frac{\theta}{n}\right)^n} \right] \\ &= \left[\frac{e^{-\sum_{i=1}^n x_i/\theta}}{\theta^n} \right] \div \left[\frac{\bar{x}^{n-1} e^{-\sum_{i=1}^n x_i/\theta}}{\Gamma(n) \left(\frac{\theta}{n}\right)^n} \right] \\ &= \frac{\Gamma(n)}{\bar{x}^{n-1} n^n} \\ &= h(x_1, \dots, x_n) \end{aligned}$$

Keep in mind that $y = \bar{x}$. Therefore, \bar{X} is a sufficient statistic for θ .



Coach's Remarks

For extra credit, prove that $\sum_{i=1}^n X_i$ is also a sufficient statistic for θ . You should obtain

$$h(x_1, \dots, x_n) = \frac{\Gamma(n)}{\left(\sum_{i=1}^n x_i\right)^{n-1}}$$

FACTORIZATION THEOREM

The definition of sufficiency requires knowing $f_Y(y)$, but this may not be feasible for any imaginable statistic Y . Hence, the **factorization theorem** provides a helpful alternative to proving sufficiency. It states that Y is a sufficient statistic for θ if and only if

$$f(x_1, \dots, x_n) = h_1(y, \theta) \cdot h_2(x_1, \dots, x_n) \quad (2.2.4.2)$$

holds for some non-negative functions h_1 and h_2 , where $h_2(x_1, \dots, x_n)$ does not depend on θ .

Let's see this theorem in action by revisiting the Bernoulli random sample from Example 2.2.4.1.

$$\begin{aligned} p(x_1, \dots, x_n) &= \prod_{i=1}^n q^{x_i} (1-q)^{1-x_i} \\ &= q^{\sum_{i=1}^n x_i} (1-q)^{n-\sum_{i=1}^n x_i} \\ &= q^{\sum_{i=1}^n x_i} (1-q)^{n-\sum_{i=1}^n x_i} \cdot 1 \\ &= h_1\left(\sum_{i=1}^n x_i, q\right) \cdot h_2(x_1, \dots, x_n) \end{aligned}$$

Thus, $\sum_{i=1}^n X_i$ is a sufficient statistic for q by the factorization theorem. In the same vein, $X_1 + X_2$ is not a sufficient statistic for q , given that it is impossible to write the joint PMF as $h_1(x_1 + x_2, q)$ times an expression that does not depend on q .

Using the factorization theorem, show that for a random sample of size n from a Poisson distribution with mean λ , the statistic $\sum_{i=1}^n X_i$ is sufficient for λ .

$$\begin{aligned} p(x_1, \dots, x_n) &= \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \\ &= \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \\ &= e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} \cdot \left(\prod_{i=1}^n x_i! \right)^{-1} \\ &= h_1\left(\sum_{i=1}^n x_i, \lambda\right) \cdot h_2(x_1, \dots, x_n) \end{aligned}$$

This proves that $\sum_{i=1}^n X_i$ is a sufficient statistic for λ .

Here is some additional information:

- When the random sample's distribution has a range that depends on θ , the factorization theorem should be used carefully; this is covered in Example 2.2.4.5.
- For a one-to-one function $g(\cdot)$, Y being a sufficient statistic for θ implies that
 - $g(Y)$ is also a sufficient statistic for θ .
 - Y is also a sufficient statistic for $g(\theta)$.

Example 2.2.4.3

A random sample X_1, X_2, \dots, X_{22} is taken from a single-parameter Pareto distribution with a lower limit of 3. A chosen estimator for α is

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^n \ln\left(\frac{X_i}{3}\right)}$$

Determine which statements are true.

- I. $\hat{\alpha}$ is the maximum likelihood estimator.
- II. $\hat{\alpha}$ is a function of a sufficient statistic for α .
- III. If $\prod_{i=1}^{22} x_i = 6.243 \times 10^{16}$, then α is estimated as 0.03.

Solution

The PDF of this single-parameter Pareto is

$$f(x) = \frac{\alpha \cdot 3^\alpha}{x^{\alpha+1}}, \quad x > 3$$

I is true. While this can be proven by finding the MLE from first principles, recall the formula provided in Section 2.1.5. The estimator $\hat{\alpha}$ matches the formula when c and all d_i 's are 0.

$$\begin{aligned}\hat{\alpha} &= \frac{n}{\sum_{i=1}^n \ln \left(\frac{X_i}{3} \right)} \\ &= \frac{n}{\sum_{i=1}^n (\ln X_i - \ln 3)}\end{aligned}$$

II is true.

$$\begin{aligned}f(x_1, \dots, x_n) &= \prod_{i=1}^n \frac{\alpha \cdot 3^\alpha}{x_i^{\alpha+1}} \\ &= \frac{(\alpha \cdot 3^\alpha)^n}{(\prod_{i=1}^n x_i)^{\alpha+1}} \\ &= \frac{(\alpha \cdot 3^\alpha)^n}{(\prod_{i=1}^n x_i)^{\alpha+1}} \cdot 1 \\ &= h_1 \left(\prod_{i=1}^n x_i, \alpha \right) \cdot h_2(x_1, \dots, x_n)\end{aligned}$$

By the factorization theorem, $\prod_{i=1}^n X_i$ is a sufficient statistic for α . Finally, note that

$$\begin{aligned}\hat{\alpha} &= \frac{n}{\sum_{i=1}^n (\ln X_i - \ln 3)} \\ &= \frac{n}{\sum_{i=1}^n \ln X_i - \sum_{i=1}^n \ln 3} \\ &= \frac{\ln \left(\prod_{i=1}^n X_i \right) - n \ln 3}{\ln \left(\prod_{i=1}^n X_i \right)} \\ &= g \left(\prod_{i=1}^n X_i \right)\end{aligned}$$

which demonstrates that $\hat{\alpha}$ is a function of a sufficient statistic.

III is false. Given $n = 22$ and $\prod_{i=1}^{22} x_i = 6.243 \times 10^{16}$, the estimate of α is

$$\hat{\alpha} = \frac{22}{\ln(6.243 \times 10^{16}) - 22 \ln 3} = 1.517$$

Therefore, **only I and II are true.**



Completeness

The distribution of Y comes from a **complete** family of distributions when $E[g(Y)] = 0$ implies that $g(y) = 0$ for every possible value of θ . More intuitively, completeness means that any function $g(\cdot)$ that causes the mean of $g(Y)$ to equal 0 must be a function that maps to 0.

Additionally,

- proving that a family is complete typically requires more rigorous math than necessary for the exam, so proofs will not be emphasized.
- a more accurate description of " $g(y) = 0$ " is " $Pr[g(Y) = 0] = 1$ ", but the former is enough for our purposes.

Completeness establishes that there is only one function of Y that is an unbiased estimator of θ , i.e. $\varphi(Y)$ is unique. To demonstrate this, assume there is another function of Y , $\psi(Y)$, that is also an unbiased estimator of θ . As a result, $\varphi(Y) - \psi(Y)$ is a function of Y whose mean equals 0.

$$\begin{aligned} E[\varphi(Y) - \psi(Y)] &= E[\varphi(Y)] - E[\psi(Y)] \\ &= \theta - \theta \\ &= 0 \end{aligned}$$

So if Y comes from a complete family of distributions, then we can conclude that

$$\begin{aligned} E[\varphi(Y) - \psi(Y)] = 0 &\Rightarrow \varphi(y) - \psi(y) = 0 \\ &\Rightarrow \varphi(y) = \psi(y) \end{aligned}$$

i.e. a function of Y that is an unbiased estimator of θ is unique.

Rao-Blackwell

The Rao-Blackwell theorem touches on the "minimum variance" aspect of the MVUE. First, let

- Y be a sufficient statistic for θ , and
- Z be an unbiased estimator of θ

such that Y is not a function of Z only, nor vice versa. Y 's sufficiency guarantees that the distribution of $(Z | Y)$ does not involve θ , and so $E_Z[Z | Y]$ is another statistic. Importantly,

1. $E_Z[Z | Y]$ is an unbiased estimator of θ .
2. $\text{Var}[E_Z[Z | Y]] \leq \text{Var}[Z]$.

These can be easily proven using the Law of Total Expectation and the Law of Total Variance; see the appendix at the end of this section.

In conclusion, this theorem asserts that, when compared to **any** unbiased estimator Z , the unbiased estimator $E_Z[Z | Y]$ is better in terms of having a lower or equal variance. In other words, $E_Z[Z | Y]$ fits the description of the MVUE.

The Lehmann-Scheffé theorem goes a step further. It is important to realize that $E_Z[Z | Y]$ is a function of Y . (Similarly, Example 1.1.9.2 in Section 1.1.9 has $E_X[X | q] = 30q$, which is a function of q .) Given that Y comes from a complete family of distributions, there is only one function of Y that is an unbiased estimator of θ . Altogether, no other function of Y can be unbiased besides $E_Z[Z | Y]$, making it the **unique** MVUE.

We may be tempted to think that one should find a Z (i.e. any unbiased estimator) and subsequently $E_Z[Z | Y]$ in order to obtain the MVUE. However, Lehmann-Scheffé shows that it is not necessary. Simply, if

- Y is a complete sufficient statistic for θ , and
- $\varphi(Y)$ is unbiased, i.e. $E[\varphi(Y)] = \theta$,

then $\varphi(Y)$ must be the unique $E_Z[Z | Y]$ based on completeness, and in turn, has the lowest variance based on the Rao-Blackwell theorem. So, our attention narrows to finding a function of Y (i.e. a complete sufficient statistic) that is unbiased.

Example 2.2.4.4

A random sample of size n is taken from an exponential distribution with mean θ . You are given that $\sum_{i=1}^n X_i$ is a complete sufficient statistic for θ . Determine the MVUE of θ .

Solution

Given a complete sufficient statistic, we simply need to determine the function of $\sum_{i=1}^n X_i$ that is unbiased, $\varphi(\sum_{i=1}^n X_i)$. We start by taking the expected value of the complete sufficient statistic.

$$E\left[\sum_{i=1}^n X_i\right] = n\theta$$

The objective is to find an equation in the form of "expected value equals parameter". Notice that

$$\begin{aligned} \frac{1}{n} E\left[\sum_{i=1}^n X_i\right] &= \theta \\ \Rightarrow E\left[\frac{\sum_{i=1}^n X_i}{n}\right] &= \theta \\ \Rightarrow E\left[\varphi\left(\sum_{i=1}^n X_i\right)\right] &= \theta \end{aligned}$$

Therefore, the MVUE of θ is $\frac{\sum_{i=1}^n X_i}{n}$, i.e. \bar{X} .



Example 2.2.4.5

A random sample of size n is taken from a uniform distribution valid on the interval $[0, \theta]$. Let $X_{(n)} = \max(X_1, \dots, X_n)$.

1. Show that $X_{(n)}$ is a sufficient statistic for θ .
2. Given that $X_{(n)}$ is a complete sufficient statistic, determine the MVUE of θ .

Solution to (1)

The PDF of this uniform distribution is

$$f(x) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta$$

When the parameter of interest is part of the domain of $f(x)$, it is information that should be captured using an indicator function, $I(\cdot)$, which equals 1 when the statement inside the parentheses is true, but equals 0 otherwise. In this case,

$$f(x) = \frac{1}{\theta} \cdot I(0 \leq x \leq \theta)$$

Writing the PDF in this form helps to properly apply the factorization theorem. Then, the joint PDF is

$$\begin{aligned} f(x_1, \dots, x_n) &= \prod_{i=1}^n \frac{I(0 \leq x_i \leq \theta)}{\theta} \\ &= \frac{\prod_{i=1}^n I(0 \leq x_i \leq \theta)}{\theta^n} \end{aligned}$$

The numerator equals 1 when all the x_i 's are within the interval $[0, \theta]$, but equals 0 otherwise. This is equivalent to producing

- a 1 if the smallest x_i is at least 0 **and** the largest x_i is at most θ , or
- a 0 otherwise.

$$\begin{aligned}
f(x_1, \dots, x_n) &= \frac{I(\min[x_1, \dots, x_n] \geq 0) \cdot I(\max[x_1, \dots, x_n] \leq \theta)}{\theta^n} \\
&= \frac{I(\max[x_1, \dots, x_n] \leq \theta)}{\theta^n} \cdot I(\min[x_1, \dots, x_n] \geq 0) \\
&= h_1(\max[x_1, \dots, x_n], \theta) \cdot h_2(x_1, \dots, x_n)
\end{aligned}$$

Therefore, $X_{(n)}$ is a sufficient statistic for θ .

■

Solution to (2)

Given a complete sufficient statistic, we simply need to determine the $\varphi(X_{(n)})$ that is unbiased. We start by taking the expected value of the n^{th} order statistic; use Equation 1.4.3.1 here.

$$\mathbb{E}[X_{(k)}] = a + \frac{k(b-a)}{n+1}$$

$$\begin{aligned}
\mathbb{E}[X_{(n)}] &= 0 + \frac{n(\theta-0)}{n+1} \\
&= \frac{n}{n+1}\theta
\end{aligned}$$

In finding an equation in the form of "expected value equals parameter", we get

$$\begin{aligned}
\frac{n+1}{n}\mathbb{E}[X_{(n)}] &= \theta \\
\Rightarrow \mathbb{E}\left[\frac{n+1}{n}X_{(n)}\right] &= \theta \\
\Rightarrow \mathbb{E}[\varphi(X_{(n)})] &= \theta
\end{aligned}$$

Therefore, the MVUE of θ is $\frac{n+1}{n}X_{(n)}$.



If we instead want the MVUE of a function of θ , then a minor adjustment is needed. The only change is to find $\varphi(Y)$ that is an unbiased estimator of $g(\theta)$. Let's see this in action.

Example 2.2.4.6

A random sample of size n is taken from an exponential distribution with mean θ . You are given that $\sum_{i=1}^n X_i$ is a complete sufficient statistic for θ .

Determine the MVUE of θ^{-1} .

Solution

From the work done in Example 2.2.4.4, we see that

$$\frac{1}{E[\bar{X}]} = \theta^{-1}$$

but this is **not** in the form of "expected value equals parameter". However, this hints at how $E[\bar{X}^{-1}]$ could bring us a step closer. Recall from Example 2.2.4.2 that

$$\bar{X} \sim \text{Gamma}\left(n, \frac{\theta}{n}\right)$$

which means

$$\bar{X}^{-1} \sim \text{Inverse Gamma}\left(n, \frac{n}{\theta}\right)$$

From the exam table,

$$\begin{aligned} \mathbb{E}\left[\bar{X}^{-1}\right] &= \frac{\left(\frac{n}{\theta}\right)^1 \Gamma(n-1)}{\Gamma(n)} \\ &= \frac{n(n-2)!}{(n-1)!} \theta^{-1} \\ &= \frac{n}{n-1} \theta^{-1} \end{aligned}$$

Consequently,

$$\begin{aligned} \frac{n-1}{n} \mathbb{E}\left[\bar{X}^{-1}\right] &= \theta^{-1} \\ \Rightarrow \mathbb{E}\left[\frac{n-1}{n} \left(\frac{n}{\sum_{i=1}^n X_i}\right)\right] &= \theta^{-1} \\ \Rightarrow \mathbb{E}\left[\varphi\left(\sum_{i=1}^n X_i\right)\right] &= \theta^{-1} \end{aligned}$$

which means the MVUE of θ^{-1} is $\frac{n-1}{\sum_{i=1}^n X_i}$.



2.2.5 Exponential Class of Distributions

When working with a random sample drawn from a distribution that belongs to the exponential class or family, a helpful result emerges in relation to sufficiency and completeness.

Random variable X has a distribution that belongs to the *exponential class of distributions* if its probability function can be written in the form

$$f(x) = \exp [a(x) \cdot b(\theta) + c(\theta) + d(x)] \quad (2.2.5.1)$$

where X 's domain cannot depend on parameter θ , and $a(\cdot)$, $b(\cdot)$, $c(\cdot)$, and $d(\cdot)$ are functions of their corresponding arguments.

Here are some distributions that belong to the exponential family:

- Binomial (fixed m)
- Normal (fixed σ^2 **or** fixed μ)
- Poisson
- Gamma (fixed α)
- Inverse Gaussian (fixed θ)
- Negative binomial (fixed r)

Show that the Poisson distribution belongs to the exponential class of distributions.

$$\begin{aligned} p(x) &= \frac{e^{-\lambda} \lambda^x}{x!} \\ &= \exp [-\lambda] \cdot \exp \left[\ln \left(\frac{\lambda^x}{x!} \right) \right] \\ &= \exp [-\lambda + \ln (\lambda^x) - \ln (x!)] \\ &= \exp [x \ln \lambda - \lambda - \ln (x!)] \end{aligned}$$

In this form, note that

- $a(x) = x$
- $b(\lambda) = \ln \lambda$
- $c(\lambda) = -\lambda$
- $d(x) = -\ln (x!)$

which proves that **the Poisson distribution belongs to the exponential class of distributions.**

Much can be said about the exponential family in general (e.g. more details are found in Section 3.8), but we want to focus on the joint probability function of the random sample for such a distribution. Note that

$$\begin{aligned}
 f(x_1, \dots, x_n) &= \prod_{i=1}^n f(x_i) \\
 &= \exp \left[\left\{ \sum_{i=1}^n a(x_i) \right\} b(\theta) + n \cdot c(\theta) + \sum_{i=1}^n d(x_i) \right] \\
 &= \exp \left[\left\{ \sum_{i=1}^n a(x_i) \right\} b(\theta) + n \cdot c(\theta) \right] \cdot \exp \left[\sum_{i=1}^n d(x_i) \right] \\
 &= h_1 \left[\sum_{i=1}^n a(x_i), \theta \right] \cdot h_2(x_1, \dots, x_n)
 \end{aligned}$$

Therefore, by the factorization theorem, this proves that $\sum_{i=1}^n a(X_i)$ is a sufficient statistic of θ . Moreover, it can be proven that the distribution of $\sum_{i=1}^n a(X_i)$ comes from a complete family of distributions. This is a powerful result, as a random sample drawn from any member of the exponential family leads to $\sum_{i=1}^n a(X_i)$ being a complete sufficient statistic for θ .

Based on the given list of exponential family distributions, the following table summarizes their key results.

Distribution	Parameter of Interest		MVUE
Binomial	q	$\sum_{i=1}^n X_i$	$\frac{1}{m} \bar{X}$
Normal	μ	$\sum_{i=1}^n X_i$	\bar{X}
Normal	σ^2	$\sum_{i=1}^n (X_i - \mu)^2$	$\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$
Poisson	λ	$\sum_{i=1}^n X_i$	\bar{X}
Gamma	θ	$\sum_{i=1}^n X_i$	$\frac{1}{\alpha} \bar{X}$
Inverse Gaussian	μ	$\sum_{i=1}^n X_i$	\bar{X}

Distribution	Parameter of Interest	$\sum_{i=1}^n a(X_i)$	MVUE
Negative binomial	β	$\sum_{i=1}^n X_i$	$\frac{1}{r}\bar{X}$

Memorizing this table is optional, as the results are not hard to derive from the probability functions.

Example 2.2.5.1

A random sample of size n is taken from a normal distribution with mean 0 and variance σ^2 .

Determine the MVUE of σ^2 .

Solution

The summary table states that the MVUE is

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

Since $\mu = 0$ for this problem, the answer is

$$\frac{1}{n} \sum_{i=1}^n (X_i - 0)^2 = \frac{\sum_{i=1}^n X_i^2}{n}$$



Alternative Solution

Here are the steps without using the summary table. First, determine a complete sufficient statistic for σ^2 . We use the fact that a normal distribution with a fixed mean is a member of

the exponential family. Recall the normal PDF, evaluate it at $\mu = 0$, and express it in the exponential family form.

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-0)^2}{2\sigma^2}\right] = \exp\left[x^2\left(-\frac{1}{2\sigma^2}\right) - \frac{1}{2}\ln(2\pi\sigma^2)\right]$$

hence

- $a(x) = x^2$
- $b(\sigma^2) = -\frac{1}{2\sigma^2}$
- $c(\sigma^2) = -\frac{1}{2}\ln(2\pi\sigma^2)$
- $d(x) = 0$

Therefore, a complete sufficient statistic for σ^2 is

$$\sum_{i=1}^n a(X_i) = \sum_{i=1}^n X_i^2$$

Since the mean is 0,

$$\text{Var}[X] = \mathbb{E}[X^2] - 0^2 = \mathbb{E}[X^2]$$

and in turn,

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^n X_i^2\right] &= \sum_{i=1}^n \mathbb{E}[X_i^2] \\ &= \sum_{i=1}^n \text{Var}[X_i] \\ &= n\sigma^2 \end{aligned}$$

Consequently,

$$\frac{1}{n} E \left[\sum_{i=1}^n X_i^2 \right] = \sigma^2 \quad \Rightarrow \quad E \left[\frac{\sum_{i=1}^n X_i^2}{n} \right] = \sigma^2$$

which means the MVUE of σ^2 is $\frac{1}{n} \sum_{i=1}^n X_i^2$ based on Lehmann-Scheffé (i.e. we have found an unbiased estimator of σ^2 that is a function of the complete sufficient statistic $\sum_{i=1}^n X_i^2$). ■

Coach's Remarks

So far, the concepts pertaining to MVUE and the exponential family were presented assuming there is only one unknown parameter, θ . These concepts have extensions that cover situations with more than one unknown parameter.

To promote efficient studying, we believe that candidates need only know the scenario where the random sample is taken from a normal distribution with unknown μ and unknown σ^2 . The MVUEs of μ and σ^2 are \bar{X} and S^2 , respectively.

2.2.6 Maximum Likelihood Estimators

We now bolster the usefulness of MLE by considering three desirable attributes that are sometimes found in maximum likelihood estimators. In this subsection, we denote $\hat{\theta}$ as a maximum likelihood estimator for θ . The three attributes are:

1. $\hat{\theta}$ is a consistent estimator of θ .
2. $\hat{\theta}$ is asymptotically normally distributed.
3. If a sufficient statistic Y for θ exists, then $\hat{\theta}$ is a function of Y .

There are theorems that specify when these attributes are guaranteed to hold. We will not go over the theorems for the first two attributes; they cover a broader scope than what we feel is needed for the exam. Instead, we may replace those theorems with the following conditions:

- X_1, \dots, X_n form a random sample
- Typical regularity conditions hold
- $\hat{\theta}$ uniquely solves the score equation

These conditions are true for many cases common to the exam.

As for the third attribute, its theorem is much simpler: it expects a random sample, and $\hat{\theta}$ needs only to exist uniquely (i.e. not required to solve the score equation).

While the theorems guarantee the attributes when their conditions are met, they do not suggest what happens when the conditions are not met.

The first and third attributes require no further explanation, so let's expand on the second attribute.

Asymptotic Normality

$\hat{\theta}$ asymptotically follows a normal distribution with mean θ and variance $1/I(\theta)$, i.e. the reciprocal of the Fisher information. Recall that this variance is the Rao-Cramér lower bound. In other words, $\hat{\theta}$ is also asymptotically efficient.

It was mentioned that $I(\theta)$ equals n times the information for a sample size of 1. This means the asymptotic variance of $\hat{\theta}$ is proportional to $1/n$, and thus shrinks as n increases.

EXACT VARIANCE

In Section 2.2.2, we considered the variance of generic estimators of θ . The same rules apply if we wish to find the **exact** variance of a maximum likelihood estimator, $\text{Var}[\hat{\theta}]$. This is noteworthy for a few reasons:

- There are many situations where $\hat{\theta}$ is a simple function of \bar{X} (recall Section 2.1.5), which makes calculating the exact variance straightforward.
- For the cases in Section 2.1.5 where the maximum likelihood and method of moments estimators are the same, the exact variance is equal to the asymptotic variance.

Let's demonstrate with the following example:

A random sample is drawn from a distribution with density function:

$$f(x) = \frac{x^{\alpha-1} e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}, \quad x > 0$$

where α is known.

Determine the variance of the maximum likelihood estimator of θ .

The density function belongs to a gamma distribution. For a gamma distribution where α is known, $\hat{\theta}$ can be determined by matching the fitted and sample means, as discussed in Section 2.1.5.

$$\alpha\theta = \bar{X} \quad \Rightarrow \quad \hat{\theta} = \frac{\bar{X}}{\alpha}$$

Therefore,

$$\begin{aligned}
\text{Var}[\hat{\theta}] &= \text{Var}\left[\frac{\bar{X}}{\alpha}\right] \\
&= \frac{1}{\alpha^2} \cdot \frac{\text{Var}[X]}{n} \\
&= \frac{1}{\alpha^2} \cdot \frac{\alpha\theta^2}{n} \\
&= \frac{\theta^2}{n\alpha}
\end{aligned}$$

Furthermore, recall from Section 2.2.3 that we found the Fisher information for the same scenario; it is

$$I(\theta) = \frac{n\alpha}{\theta^2}$$

Notice that the asymptotic variance and the exact variance are the same in this case.

$$\frac{1}{I(\theta)} = \frac{\theta^2}{n\alpha} = \text{Var}[\hat{\theta}]$$

Consequently, the variance of $\hat{\theta}$ attains the Rao-Cramér lower bound.

Coach's Remarks

The asymptotic normality of maximum likelihood estimators extends to the multi-parameter case. The maximum likelihood estimators $\hat{\theta}_1, \dots, \hat{\theta}_p$ asymptotically follow a multivariate normal distribution with means $\theta_1, \dots, \theta_p$ and variance-covariance matrix \mathbf{I}^{-1} . This concept will be used in Section 3.8, but we do not expect it to be tested in the current, broader context.

Before ending with some examples, let's revisit Example 2.2.4.6 to tie together some core ideas. The example is repeated here for convenience:

Example 2.2.4.6

A random sample of size n is taken from an exponential distribution with mean θ . You are given that $\sum_{i=1}^n X_i$ is a complete sufficient statistic for θ .

Determine the MVUE of θ^{-1} .

The solution proposes to find the expected value of \bar{X}^{-1} ; perhaps this seems to be nothing more than a guess. However, given the aforementioned concepts, this is not the case.

First, establish that there exists a sufficient statistic for θ^{-1} . Use the fact that the exponential distribution is a member of the exponential family.

$$\begin{aligned}\frac{1}{\theta} e^{-x/\theta} &= \theta^{-1} e^{-x\theta^{-1}} \\ &= \exp[x(-\theta^{-1}) + \ln(\theta^{-1})]\end{aligned}$$

Given that $a(x) = x$, a complete sufficient statistic for θ^{-1} is $\sum_{i=1}^n X_i$. This is the same complete sufficient statistic for θ provided in the example. This illustrates that it is possible for a statistic to be sufficient and complete for both a generic parameter θ and a function of θ .

Next, determine the maximum likelihood estimator of θ^{-1} , which we will denote as $\widehat{\theta}^{-1}$. While it can be found by the usual "solve the score equation" strategy, let's use the fact that $\widehat{\theta}$ is \bar{X} (from Section 2.1.5) and apply the invariance property (the maximum likelihood estimator of $g(\theta)$ is $g(\widehat{\theta})$). Altogether,

$$\widehat{\theta}^{-1} = (\widehat{\theta})^{-1} = \bar{X}^{-1}$$

Therefore, \bar{X}^{-1} is not an arbitrary statistic; it is the maximum likelihood estimator of θ^{-1} . So, in search of the MVUE, why is taking the expected value of a maximum likelihood estimator useful? According to a maximum likelihood estimator attribute, since a sufficient statistic for θ^{-1} exists, then \bar{X}^{-1} must be a function of that sufficient statistic. Because the sufficient statistic in question is a complete sufficient statistic, there are two possible outcomes:

- If \bar{X}^{-1} is unbiased, then \bar{X}^{-1} is the MVUE, i.e. $\varphi(\sum_{i=1}^n X_i) = \bar{X}^{-1}$.
- If \bar{X}^{-1} is biased, then \bar{X}^{-1} should be close to the MVUE, since maximum likelihood estimators are asymptotically unbiased under asymptotic normality.

Hence, calculating $E\left[\bar{X}^{-1}\right]$ is the next step. As noted in the solution, \bar{X}^{-1} is asymptotically unbiased, and the MVUE is $\frac{n-1}{n}\bar{X}^{-1}$.

Example 2.2.6.1

A random sample of size 5 is drawn from a geometric distribution with mean β .

4 5 5 11 20

You want to estimate β using maximum likelihood estimation.

Estimate the variance of the estimator, $\hat{\beta}$.

Solution

Recall that a geometric distribution is a negative binomial distribution with $r = 1$. Thus, $\hat{\beta}$ can be calculated by matching the fitted and sample means.

$$\hat{\beta} = \bar{X}$$

Therefore, the variance of $\hat{\beta}$ is

$$\begin{aligned}\text{Var}\left[\hat{\beta}\right] &= \text{Var}\left[\bar{X}\right] \\ &= \frac{\text{Var}[X]}{n} \\ &= \frac{\beta(1+\beta)}{n}\end{aligned}$$

We estimate the variance based on the estimate of the parameter. The MLE estimate of β is the sample mean of the data.

$$\hat{\beta} = \bar{x} = \frac{4 + 5 + 5 + 11 + 20}{5} = 9$$

As a result, the estimated variance is

$$\begin{aligned}\widehat{\text{Var}}[\hat{\beta}] &= \frac{\hat{\beta}(1 + \hat{\beta})}{n} \\ &= \frac{9 \cdot 10}{5} \\ &= \mathbf{18}\end{aligned}$$



Example 2.2.6.2

Claim sizes of a group health insurance policy have the following density function:

$$f(x) = \frac{\alpha \cdot 500^\alpha}{(x + 500)^{\alpha+1}}, \quad x > 0$$

A sample of claims is given below:

36 107 234 601 644

Estimate the asymptotic variance of the maximum likelihood estimator of α .

Solution

The asymptotic variance of the maximum likelihood estimator $\hat{\alpha}$ is the reciprocal of $I(\alpha)$. So, first solve for the information using Equation 2.2.3.3.

Begin by taking the natural log of the density function, and then calculate its second derivative with respect to α .

$$\begin{aligned}\ln [f(x)] &= \ln \alpha + \alpha \ln 500 - (\alpha + 1) \ln (x + 500) \\ \frac{d}{d\alpha} \ln [f(x)] &= \frac{1}{\alpha} + \ln 500 - \ln (x + 500) \\ \frac{d^2}{d\alpha^2} \ln [f(x)] &= -\frac{1}{\alpha^2} + 0 - 0\end{aligned}$$

Using Equation 2.2.3.3, we get

$$\begin{aligned}I(\alpha) &= -n \cdot E \left[\frac{d^2}{d\alpha^2} \ln [f(X)] \right] \\ &= -5 \cdot E \left[-\frac{1}{\alpha^2} \right] \\ &= \frac{5}{\alpha^2}\end{aligned}$$

Therefore, the asymptotic variance of $\hat{\alpha}$ is

$$\frac{1}{I(\alpha)} = \frac{\alpha^2}{5}$$

We estimate this variance based on the MLE estimate of α . We may use the MLE formula in Section 2.1.5 since the density provided is a Pareto density with $\theta = 500$. Instead, let's determine the score function. Rather than deriving it from scratch, consider using an expression we have already found, i.e. $\frac{d}{d\alpha} \ln [f(x)]$.

$$\begin{aligned}
l'(\alpha) &= \frac{d}{d\alpha} \left\{ \ln \left[\prod_{i=1}^5 f(x_i) \right] \right\} \\
&= \frac{d}{d\alpha} \left\{ \sum_{i=1}^5 \ln [f(x_i)] \right\} \\
&= \sum_{i=1}^5 \frac{d}{d\alpha} \ln [f(x_i)] \\
&= \sum_{i=1}^5 \left[\frac{1}{\alpha} + \ln 500 - \ln (x_i + 500) \right] \\
&= \frac{5}{\alpha} + 5 \ln 500 - \sum_{i=1}^5 \ln (x_i + 500) \\
&= \frac{5}{\alpha} + 5 \ln 500 - (\ln 536 + \dots + \ln 1,144)
\end{aligned}$$

Set the score function equal to 0 and solve for α . The MLE estimate is

$$\begin{aligned}
\frac{5}{\alpha} + 5 \ln 500 - (\ln 536 + \dots + \ln 1,144) &= 0 \\
5 + \alpha [5 \ln 500 - (\ln 536 + \dots + \ln 1,144)] &= 0 \\
\alpha &= \frac{5}{(\ln 536 + \dots + \ln 1,144) - 5 \ln 500}
\end{aligned}$$

$$\hat{\alpha} = 2.2081$$

Finally, estimate the asymptotic variance as

$$\frac{\hat{\alpha}^2}{5} = \frac{2.2081^2}{5} = \mathbf{0.9751}$$



Example 2.2.6.3

A random sample X_1, \dots, X_n is drawn from a Bernoulli distribution with success probability q . The sample mean is used to estimate q .

Determine which statements are true.

- I. The sample mean is the maximum likelihood estimator of q .
- II. The sample mean is not a consistent estimator of q .
- III. The sample mean is the MVUE.
- IV. The sample mean is not an efficient estimator of q , but is asymptotically efficient.
- V. No other estimator of q can have a lower variance than the sample mean's variance.

Solution

I is true. A Bernoulli distribution is a binomial distribution with $m = 1$. According to Section 2.1.5, the maximum likelihood estimator of q is the same as the method of moments estimator, thus

$$\hat{q} = \bar{X}$$

where \hat{q} is the maximum likelihood estimator of q .

II is false. For the cases in Section 2.1.5 where the maximum likelihood and method of moments estimators are the same, you may memorize that those maximum likelihood estimators uniquely solve their respective score equations. Therefore, $\hat{q} = \bar{X}$ is a consistent estimator of q .

III is true. The Bernoulli distribution is a member of the exponential family. As a result, there exists a sufficient statistic for q , which is also a complete sufficient statistic. As the maximum likelihood estimator, \bar{X} must be a function of this complete sufficient statistic. Since \bar{X} is also an unbiased estimator of q , conclude that \bar{X} is the MVUE.

IV is false. For the cases in Section 2.1.5 where the maximum likelihood and method of moments estimators are the same, the exact variance and the asymptotic variance of the maximum likelihood estimator are equal. Said differently, the variance of $\hat{q} = \bar{X}$ attains the Rao-Cramér lower bound. Hence, \bar{X} is an efficient estimator of q .

V is false. \bar{X} is the MVUE, meaning it has the lowest variance among all unbiased estimators of q . However, biased estimators could have a lower variance than \bar{X} .

Therefore, **only I and III are true.**



2.2 Summary

For a sample of n random variables, the sample mean and the unbiased sample variance are

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}, \quad S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

For a random sample,

$$E[\bar{X}] = E[X], \quad \text{Var}[\bar{X}] = \frac{\text{Var}[X]}{n}$$

Estimator Properties

- Bias:

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

- Mean squared error:

$$\text{MSE}[\hat{\theta}] = E\left[\left(\hat{\theta} - \theta\right)^2\right] = \text{Var}[\hat{\theta}] + \left(\text{Bias}[\hat{\theta}]\right)^2$$

- Consistency:

If $\hat{\theta}$ is asymptotically unbiased, and $\lim_{n \rightarrow \infty} \text{Var}[\hat{\theta}] = 0$, then $\hat{\theta}$ is consistent.

- Efficiency:

$$\text{Eff}[\hat{\theta}] = \frac{1/I(\theta)}{\text{Var}[\hat{\theta}]}$$

$\hat{\theta}$ is efficient when $\text{Eff}[\hat{\theta}] = 1$, i.e. its variance attains the Rao-Cramér lower bound.

Fisher Information

For a random sample X_1, \dots, X_n ,

- One-parameter case:

$$\begin{aligned} I(\theta) &= \text{Var}[l'(\theta)] \\ &= -\mathbb{E}[l''(\theta)] \\ &= -n \cdot \mathbb{E}\left[\frac{d^2}{d\theta^2} \ln[f(X)]\right] \end{aligned}$$

For an unbiased $\hat{\theta}$, the Rao-Cramér lower bound for $\text{Var}[\hat{\theta}]$ is $\frac{1}{I(\theta)}$.

- Multi-parameter case:

$\mathbf{I}(\theta_1, \dots, \theta_p)$ is a $p \times p$ matrix whose i^{th} row and j^{th} column entry is

$$-n \cdot \mathbb{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln[f(X)]\right]$$

For an unbiased $\hat{\theta}_j$, the Rao-Cramér lower bound for $\text{Var}[\hat{\theta}_j]$ is the j^{th} diagonal entry of \mathbf{I}^{-1} .

Minimum Variance Unbiased Estimator (MVUE)

- The MVUE is an unbiased estimator with the smallest variance among all unbiased estimators, regardless of the true parameter value.
- For a random sample X_1, \dots, X_n and statistic Y , the MVUE of θ is $\varphi(Y)$ when
 - Y is a complete sufficient statistic for θ , and
 - $\varphi(Y)$ is an unbiased estimator of θ .

- Sufficiency:

- Y is a sufficient statistic for θ if and only if

$$f(x_1, \dots, x_n | y) = h(x_1, \dots, x_n)$$

where $h(x_1, \dots, x_n)$ does not depend on θ .

- By the factorization theorem, sufficiency is also attained if and only if

$$f(x_1, \dots, x_n) = h_1(y, \theta) \cdot h_2(x_1, \dots, x_n)$$

for non-negative functions h_1 and h_2 , where $h_2(x_1, \dots, x_n)$ does not depend on θ .

- For a one-to-one function $g(\cdot)$, Y being a sufficient statistic for θ implies that

- $g(Y)$ is also a sufficient statistic for θ
- Y is also a sufficient statistic for $g(\theta)$

- Y is from a complete family of distributions when $E[g(Y)] = 0$ implies that $g(y) = 0$ for every possible value of θ .
- By the Rao-Blackwell theorem, the variance of the unbiased estimator $E_Z[Z | Y]$ is at most the variance of any unbiased estimator Z , where Y is a sufficient statistic. Together with completeness, the MVUE $\varphi(Y)$ is $E_Z[Z | Y]$.

Exponential Class of Distributions

- Distributions in the exponential family have probability functions in the form of

$$f(x) = \exp [a(x) \cdot b(\theta) + c(\theta) + d(x)]$$

- $\sum_{i=1}^n a(X_i)$ is a complete sufficient statistic for θ .

Maximum Likelihood Estimators

Under specific circumstances, the maximum likelihood estimator of θ

- is a consistent estimator.
- asymptotically follows a normal distribution with mean θ and variance $\frac{1}{I(\theta)}$; its exact variance may equal the asymptotic variance.
- is a function of a sufficient statistic Y .

Appendix

Fisher Information's Equivalent Forms

Take note of the following calculus and algebra facts:

- $\frac{d}{dt} \ln [g(t)] = \frac{\frac{d}{dt} g(t)}{g(t)}$
- $\left(\sum_{i=1}^n y_i \right)^2 = \sum_{i=1}^n y_i^2 + 2 \left(\sum_{1 \leq i < j \leq n} y_i y_j \right)$

In addition, note that for a random sample with $f(x)$ as the PMF/PDF,

$$l(\theta) = \ln \left[\prod_{i=1}^n f(X_i) \right] = \sum_{i=1}^n \ln [f(X_i)]$$

We want to show that the following equation holds:

$$\text{Var}[l'(\theta)] = -\mathbb{E}[l''(\theta)]$$

To streamline the steps, here are three convenient results (which are proven at the very end):

- $\mathbb{E} \left[\frac{d}{d\theta} \ln [f(X_i)] \right] = 0$
- $\mathbb{E} \left[\frac{d}{d\theta} \ln [f(X_i)] \cdot \frac{d}{d\theta} \ln [f(X_j)] \right] = 0, \quad i \neq j$
- $\mathbb{E} \left[\frac{\frac{d^2}{d\theta^2} f(X_i)}{f(X_i)} \right] = 0$

Let's start by simplifying the left side of the equation.

$$\begin{aligned} \text{Var}[l'(\theta)] &= \mathbb{E} \left[\{l'(\theta)\}^2 \right] - \{\mathbb{E}[l'(\theta)]\}^2 \\ &= \mathbb{E} \left[\{l'(\theta)\}^2 \right] - 0^2 \\ &= \mathbb{E} \left[\{l'(\theta)\}^2 \right] \end{aligned}$$

since

$$\begin{aligned}
 E[l'(\theta)] &= E\left[\frac{d}{d\theta}l(\theta)\right] \\
 &= E\left[\frac{d}{d\theta}\left\{\sum_{i=1}^n \ln[f(X_i)]\right\}\right] \\
 &= E\left[\sum_{i=1}^n \frac{d}{d\theta} \ln[f(X_i)]\right] \\
 &= \underbrace{\sum_{i=1}^n E\left[\frac{d}{d\theta} \ln[f(X_i)]\right]}_{=0} \\
 &= 0
 \end{aligned}$$

Continuing further produces

$$\begin{aligned}
 E[\{l'(\theta)\}^2] &= E\left[\left\{\sum_{i=1}^n \frac{d}{d\theta} \ln[f(X_i)]\right\}^2\right] \\
 &= E\left[\sum_{i=1}^n \left\{\frac{d}{d\theta} \ln[f(X_i)]\right\}^2 + 2 \left\{\sum_{1 \leq i < j \leq n} \frac{d}{d\theta} \ln[f(X_i)] \cdot \frac{d}{d\theta} \ln[f(X_j)]\right\}\right] \\
 &= E\left[\sum_{i=1}^n \left\{\frac{d}{d\theta} \ln[f(X_i)]\right\}^2\right] + 2 \left(\underbrace{\sum_{1 \leq i < j \leq n} E\left[\frac{d}{d\theta} \ln[f(X_i)] \cdot \frac{d}{d\theta} \ln[f(X_j)]\right]}_{=0} \right) \\
 &= E\left[\sum_{i=1}^n \left\{\frac{d}{d\theta} \ln[f(X_i)]\right\}^2\right]
 \end{aligned}$$

This concludes simplifying the left side of the equation, $\text{Var}[l'(\theta)]$. Now we simplify the right side of the equation by first rewriting $l''(\theta)$.

$$\begin{aligned}
l''(\theta) &= \frac{d^2}{d\theta^2} \left\{ \sum_{i=1}^n \ln [f(X_i)] \right\} \\
&= \sum_{i=1}^n \frac{d^2}{d\theta^2} \ln [f(X_i)] \\
&= \sum_{i=1}^n \frac{d}{d\theta} \left[\frac{\frac{d}{d\theta} f(X_i)}{f(X_i)} \right] \\
&= \sum_{i=1}^n \left[\frac{\frac{d^2}{d\theta^2} f(X_i)}{f(X_i)} - \frac{\frac{d}{d\theta} f(X_i)}{f(X_i)^2} \cdot \frac{d}{d\theta} f(X_i) \right] \\
&= \sum_{i=1}^n \frac{\frac{d^2}{d\theta^2} f(X_i)}{f(X_i)} - \sum_{i=1}^n \left\{ \frac{\frac{d}{d\theta} f(X_i)}{f(X_i)} \right\}^2 \\
&= \sum_{i=1}^n \frac{\frac{d^2}{d\theta^2} f(X_i)}{f(X_i)} - \sum_{i=1}^n \left\{ \frac{d}{d\theta} \ln [f(X_i)] \right\}^2
\end{aligned}$$

Consequently,

$$\begin{aligned}
-E[l''(\theta)] &= -E \left[\sum_{i=1}^n \frac{\frac{d^2}{d\theta^2} f(X_i)}{f(X_i)} - \sum_{i=1}^n \left\{ \frac{d}{d\theta} \ln [f(X_i)] \right\}^2 \right] \\
&= - \underbrace{\sum_{i=1}^n E \left[\frac{\frac{d^2}{d\theta^2} f(X_i)}{f(X_i)} \right]}_{=0} + E \left[\sum_{i=1}^n \left\{ \frac{d}{d\theta} \ln [f(X_i)] \right\}^2 \right] \\
&= E \left[\sum_{i=1}^n \left\{ \frac{d}{d\theta} \ln [f(X_i)] \right\}^2 \right] \\
&= E[\{l'(\theta)\}^2] \\
&= \text{Var}[l'(\theta)]
\end{aligned}$$

This concludes the proof.

We now prove the aforementioned convenient results. Recall these probability facts when given a random sample:

- $\int_{-\infty}^{\infty} f(x_i) dx_i = 1$

- $f(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i)$

Although a continuous distribution was assumed, the same ideas apply for a discrete distribution with sums replacing the integrals. Furthermore, for n functions $g_1(x_1), \dots, g_n(x_n)$,

$$\int \cdots \int \prod_{i=1}^n g_i(x_i) dx_1 \dots dx_n = \prod_{i=1}^n \left[\int g_i(x_i) dx_i \right]$$

For the first convenient result:

$$\begin{aligned} E\left[\frac{d}{d\theta} \ln [f(X_i)]\right] &= \int_{-\infty}^{\infty} \cdots \int \frac{d}{d\theta} \ln [f(x_i)] \cdot f(x_1, \dots, x_n) dx_1 \dots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int \frac{d}{d\theta} \ln [f(x_i)] \cdot \prod_{k=1}^n f(x_k) dx_1 \dots dx_n \\ &= \left(\int_{-\infty}^{\infty} \frac{d}{d\theta} \ln [f(x_i)] \cdot f(x_i) dx_i \right) \cdot \prod_{\substack{k=1 \\ k \neq i}}^n \left[\int_{-\infty}^{\infty} f(x_k) dx_k \right] \\ &= \left(\int_{-\infty}^{\infty} \frac{\frac{d}{d\theta} f(x_i)}{f(x_i)} \cdot f(x_i) dx_i \right) \cdot \prod_{\substack{k=1 \\ k \neq i}}^n 1 \\ &= \int_{-\infty}^{\infty} \frac{d}{d\theta} f(x_i) dx_i \\ &= \frac{d}{d\theta} \int_{-\infty}^{\infty} f(x_i) dx_i \\ &= \frac{d}{d\theta} 1 \\ &= 0 \end{aligned}$$

assuming the regularity conditions permit us to swap the order of the differential and integral operators.

The proof of the second convenient result is much like the first:

$$\begin{aligned}
E \left[\frac{d}{d\theta} \ln [f(X_i)] \cdot \frac{d}{d\theta} \ln [f(X_j)] \right] &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{d}{d\theta} \ln [f(x_i)] \cdot \frac{d}{d\theta} \ln [f(x_j)] \cdot \prod_{k=1}^n f(x_k) dx_1 \dots dx_n \\
&= \prod_{k=i, j} \left(\int_{-\infty}^{\infty} \frac{d}{d\theta} \ln [f(x_k)] \cdot f(x_k) dx_k \right) \cdot \prod_{\substack{k=1 \\ k \neq i, j}}^n \left(\int_{-\infty}^{\infty} f(x_k) dx_k \right) \\
&= \prod_{k=i, j} \left(\frac{d}{d\theta} 1 \right) \cdot \prod_{\substack{k=1 \\ k \neq i, j}}^n 1 \\
&= 0
\end{aligned}$$

The third convenient result follows likewise:

$$\begin{aligned}
E \left[\frac{\frac{d^2}{d\theta^2} f(X_i)}{f(X_i)} \right] &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\frac{d^2}{d\theta^2} f(x_i)}{f(x_i)} \cdot \prod_{k=1}^n f(x_k) dx_1 \dots dx_n \\
&= \left(\int_{-\infty}^{\infty} \frac{\frac{d^2}{d\theta^2} f(x_i)}{f(x_i)} \cdot f(x_i) dx_i \right) \cdot \prod_{\substack{k=1 \\ k \neq i}}^n \left[\int_{-\infty}^{\infty} f(x_k) dx_k \right] \\
&= \int_{-\infty}^{\infty} \frac{d^2}{d\theta^2} f(x_i) dx_i \cdot \prod_{\substack{k=1 \\ k \neq i}}^n 1 \\
&= \frac{d^2}{d\theta^2} \int_{-\infty}^{\infty} f(x_i) dx_i \\
&= \frac{d^2}{d\theta^2} 1 \\
&= 0
\end{aligned}$$

Rao-Blackwell Key Results

Two key results of the Rao-Blackwell theorem are:

1. $E_Z[Z | Y]$ is an unbiased estimator of θ
2. $\text{Var}[E_Z[Z | Y]] \leq \text{Var}[Z]$

To prove the first result, recall that the Law of Total Expectation gives

$$E[E_Z[Z | Y]] = E[Z]$$

Since Z is an unbiased estimator (i.e. its mean equals θ), then $E_Z[Z | Y]$ is also unbiased since it shares the same mean with Z .

To prove the second result, recall that the Law of Total Variance gives

$$E[\text{Var}_Z[Z | Y]] + \text{Var}[E_Z[Z | Y]] = \text{Var}[Z]$$

Since a variance must be non-negative, conclude that $E[\text{Var}_Z[Z | Y]] \geq 0$, and dropping it from the Law of Total Variance formula produces

$$\text{Var}[E_Z[Z | Y]] \leq \text{Var}[Z]$$

2.3.0 Overview

 5m

In our study of parameters so far, we use point estimation to estimate the value of a parameter. To justify the estimate, we investigate the qualities of the estimator used.

Hypothesis testing takes a very different approach to examine a parameter. Imagine that we wish to assess whether a coin is fair. Using a Bernoulli distribution, this is equivalent to assessing whether $q = 0.5$. If the coin was tossed 10 times, there are a few clear cases. If 5 heads were observed, it seems reasonable to say that the coin is fair; if 0 or 10 heads were observed, it seems reasonable to say that the coin is biased. What about the other possibilities, such as 3 heads?

Hypothesis tests aim to answer similar questions by applying probability concepts. We begin by learning the basic terminology and elements of hypothesis tests. Then, we will discuss specific scenarios where these tests are commonly used to study parameters such as means, proportions, and variances.

2.3.1 Terminology

A hypothesis test takes two opposing possibilities and checks which one is better supported by the available data. Specifically, the data is summarized by a single value, which is judged to be either plausible or implausible using probability. A plausible value supports one possibility, while an implausible value supports the other. This logic is consistent with the intuition expressed in the introductory coin example.

To define the boundary that separates "plausible" from "implausible", we need to be familiar with the terminology used in hypothesis testing.

Null and Alternative Hypotheses

The two "opposing possibilities" mentioned are called the **null hypothesis** and the **alternative hypothesis**, denoted as H_0 and H_1 , respectively. These hypotheses are usually mathematical statements about parameters of interest. For example, "a coin is fair" can be expressed as the hypothesis: Bernoulli's $q = 0.5$.

The null hypothesis often takes a "status quo" position, meaning it is the statement assumed to be true by default. In turn, the alternative hypothesis is typically the statement that a researcher has interest in proving.

In conducting a hypothesis test, the calculations are performed assuming the null hypothesis is true. After weighing the evidence, the researcher decides to either

- **fail to reject** the null hypothesis, or
- **reject** the null hypothesis in favor of the alternative hypothesis.

In other words, without sufficient evidence supporting H_1 , we keep assuming the default of H_0 . Otherwise, sufficient evidence favoring H_1 would suggest that H_0 ought to be rejected.

Coach's Remarks

Other resources may phrase the first decision as "accept the null hypothesis". However, we prefer "fail to reject" because we see it as more precise. A hypothesis test does not prove which hypothesis is true; it only shows which hypothesis is preferred by the evidence.

Test Statistic

The **test statistic** is a statistic that is used in determining whether the null hypothesis should be rejected. Thus, it summarizes the sample observations while assuming the null hypothesis is true. Using the test statistic's distribution, we can determine whether the test statistic calculated from the observed data is considered "plausible" or "implausible".

Coach's Remarks

You may have noticed that the literature on statistics tends to use certain terms loosely. For example, both \bar{X} and \bar{x} are often simply referred to as "sample mean" in spite of the inherent difference between the two, as previously discussed.

The term "test statistic" is no different. However, there is a lesser emphasis on a test statistic being a random variable in hypothesis testing. Therefore, we will proceed with "test statistic" referring only to the value calculated from the data.

Consider this typical scenario: if a test statistic is near either tail of the distribution, then the data appears to be a rare occurrence (i.e. implausible). Conversely, a test statistic that is closer to the center of the distribution suggests that the data appears to be a typical occurrence (i.e. plausible). Keep in mind that the distribution is based on the null hypothesis being true.

When a test statistic is in either tail of the distribution, perhaps it is not true that the data was just a rare occurrence. Instead, the data could have come from a different distribution altogether. This implies that the null hypothesis is actually incorrect. In other words, an extreme test statistic would support the alternative hypothesis more than the null hypothesis.

To assist with learning the rest of the jargon, we will assume the following setup:

- There is one sample observation, X , that is normally distributed with mean μ and known variance σ^2 .
- The hypothesis test investigates the value of μ . Specifically,

$$H_0 : \mu = h \quad H_1 : \mu \neq h$$

for some constant h .

- The test statistic is x , the observed value of X .

To summarize, x is a realization from the normal distribution specified by H_0 . So, if x is very small or large (i.e. in either tail of the normal distribution), that unusual result implies that a mean of h is not convincing; we become suspicious of H_0 and are more inclined to reject it.

Critical Region and Critical Value

The ***critical region*** is the range of test statistic values that we consider "too extreme" and thus decide to reject H_0 in favor of H_1 . A ***critical value*** is a value that separates the critical region from the rest of the possible test statistic values.

Before continuing, it is important to distinguish between ***two-tailed tests*** and ***one-tailed tests***.

- Two-tailed: both tails of the distribution are included in the critical region.
- One-tailed: only one tail of the distribution is included in the critical region.

Based on the H_0 and H_1 we are currently considering, we limit our discussion to two-tailed tests for now. The critical region can be written as

$$[x \leq a] \cup [x \geq b]$$

meaning the test statistic x is "too extreme" if it is smaller than a **or** greater than b , in which case H_0 would be rejected. Therefore, a and b are critical values; they are chosen such that both tails are symmetrical. Since X is normally distributed, the critical region can also be written in terms of the standard normal distribution, i.e.

$$[z \leq -c] \cup [z \geq c] \quad \Leftrightarrow \quad |z| \geq c$$

where the test statistic is now $z = \frac{x-h}{\sigma}$. In this case, notice that we may avoid keeping track of two critical values, $-c$ and c , by taking the absolute value of z .

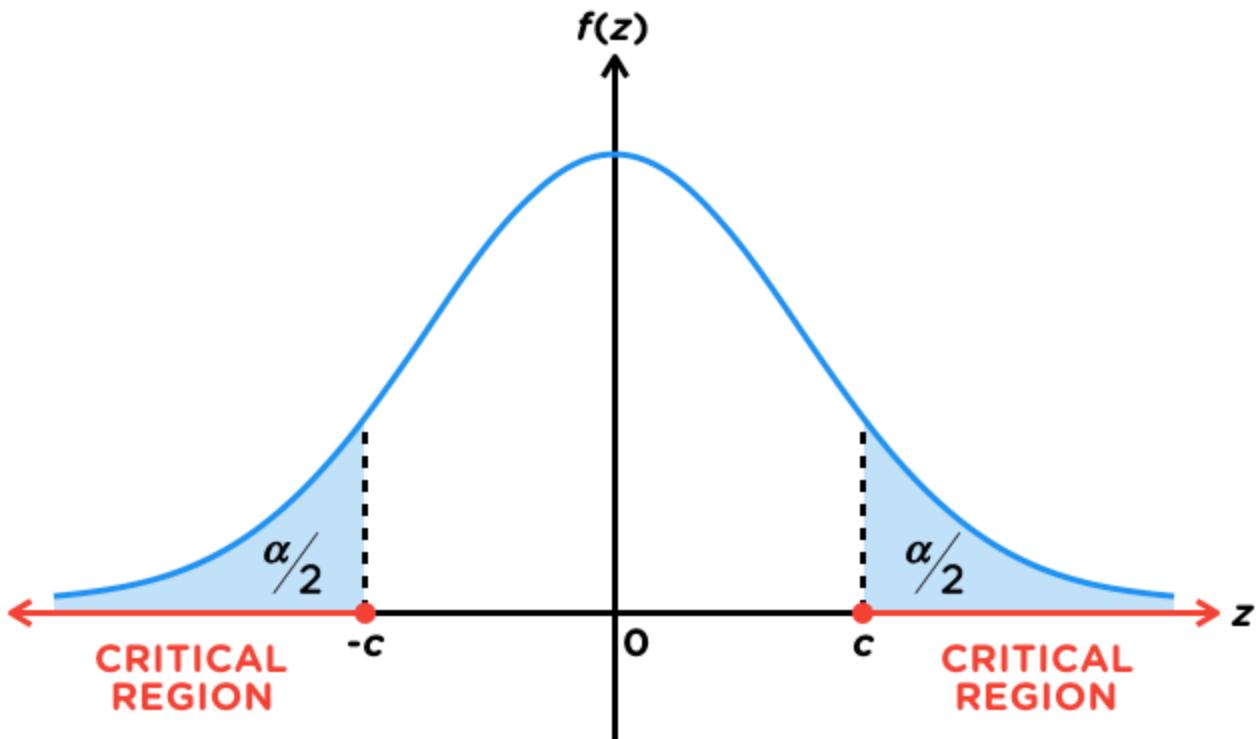
Significance Level

Critical values set the boundary for how extreme the test statistic must be in order to reject the null hypothesis. For our assumed setup, the critical value c is determined by setting a ***significance level*** or ***size*** denoted by α , where

$$\Pr(|Z| \geq c \mid H_0 \text{ is true}) = \alpha$$

Here, the size of the test is the probability of rejecting H_0 , assuming it is true. Clearly, we would prefer not to reject H_0 if it is true; hence, α is typically a small percent. The closer α is to 0, the less likely that H_0 will be rejected; a test statistic would need to be more extreme to provide evidence against H_0 .

The following graph illustrates the concepts. It shows the standard normal distribution assuming H_0 is true.



More generally, for a test investigating a generic parameter θ , the definition of size α is

$$\alpha = \max_{\theta} \Pr(\text{critical region} \mid H_0 \text{ is true}) \quad (2.3.1.1)$$

which we will use and explain later.

From start to finish, a hypothesis test follows this outline:

1. Determine an appropriate significance level for the test. A common value is $\alpha = 0.05$.
2. Based on the distribution which assumes H_0 is true, identify the critical region and determine the critical value that corresponds to the chosen α .

3. Collect data, and use it to calculate the test statistic.
4. Compare the test statistic to the critical value based on the critical region. For the two-tailed test of our assumed setup:

- If $|z| \geq c$, then it is in the critical region; reject H_0 .
- If $|z| < c$, then it is not in the critical region; do not reject H_0 .

Let's perform a complete hypothesis test for a scenario that matches our assumed setup.

Example 2.3.1.1

It is believed that the mean age of licensed drivers in 2017 is 43.7. To test whether this is the case, one licensed driver's age is observed to be 60 in 2017. These ages are normally distributed with variance 80.

Test whether the mean age of licensed drivers in 2017 differs from 43.7 at the 5% significance level.

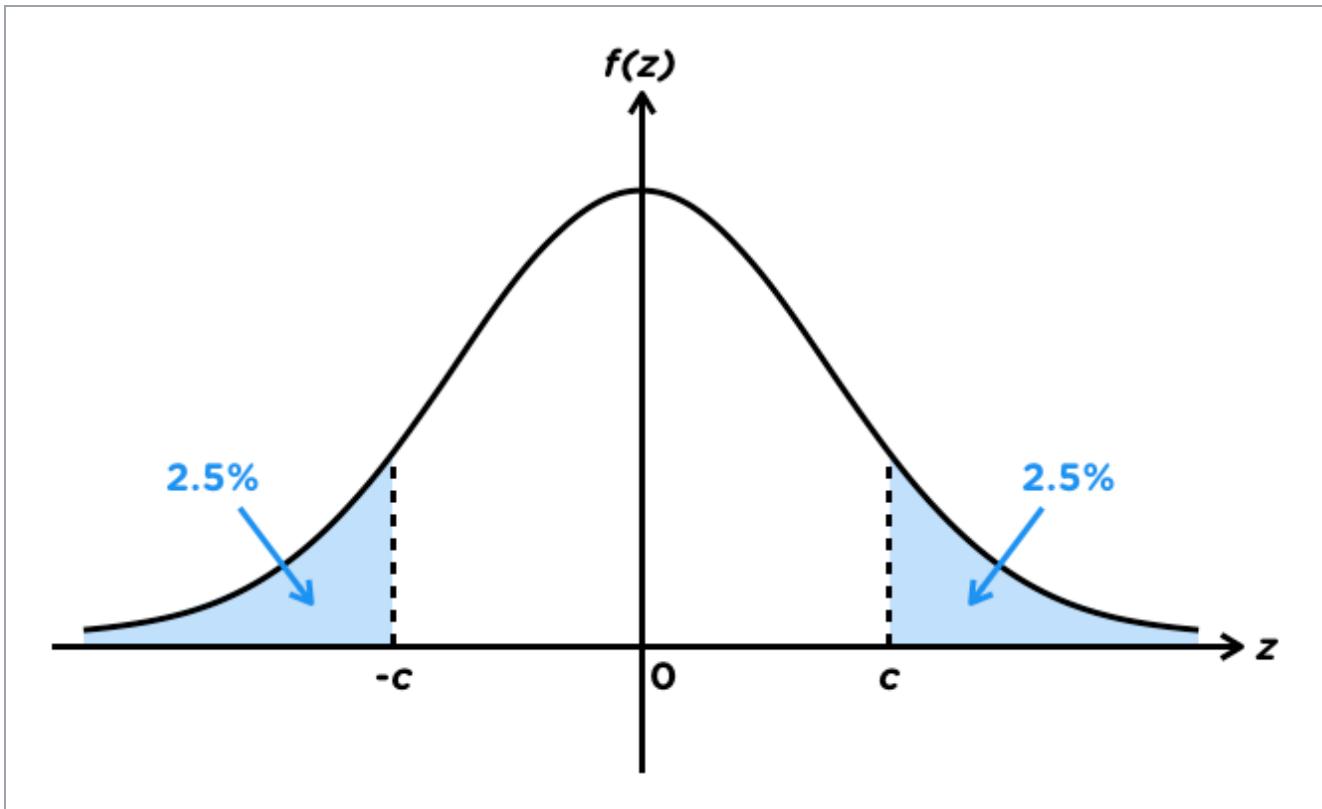
Solution

Start by formally stating the null and alternative hypotheses. By letting μ represent the mean age of licensed drivers in 2017,

$$H_0 : \mu = 43.7 \quad H_1 : \mu \neq 43.7$$

The critical region is $|z| \geq c$. Next, determine the critical value. Since $\alpha = 0.05$,

$$\Pr(|Z| \geq c \mid H_0 \text{ is true}) = 0.05$$



The graph indicates that c is the $2.5 + 95 = 97.5^{\text{th}}$ percentile of Z , i.e. $z_{0.975}$. From the normal distribution table, obtain $c = z_{0.975} = 1.96$.

Next, calculate the test statistic.

$$z = \frac{x - h}{\sigma} = \frac{60 - 43.7}{\sqrt{80}} = 1.82$$

Since $|1.82| < 1.96$, the test statistic does not fall in the critical region, and thus we **do not reject the null hypothesis of $\mu = 43.7$ at the 5% significance level**. This means the observation of a 60-year-old licensed driver is plausible if the mean age is in fact 43.7. ■

p-Value

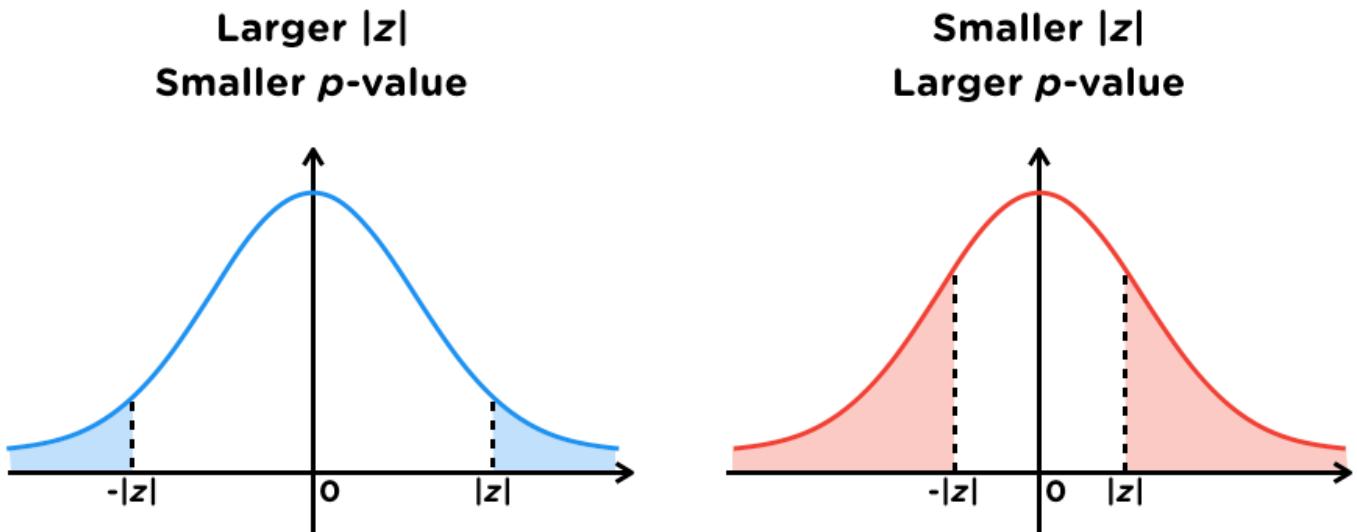
A ***p-value*** is the probability of observing the test statistic or a more extreme value, assuming H_0 is true. Thus, for our assumed setup, the *p*-value is

$$\Pr(|Z| \geq |z| \mid H_0 \text{ is true})$$

This resembles the significance level formula (i.e. the critical region inequality), with $|z|$ replacing the critical value c . Therefore, instead of comparing $|z|$ with c to make a decision, we may compare the p -value with the significance level α .

- If p -value $\leq \alpha$, then reject H_0 .
- If p -value $> \alpha$, then do not reject H_0 .

It should be evident that the two comparisons are equivalent. A larger $|z|$ results in a smaller p -value, and vice versa.



What is the p -value for Example 2.3.1.1?

$$\begin{aligned}
 \Pr(|Z| \geq |z| \mid H_0 \text{ is true}) &= 2 \cdot \Pr(Z \geq |1.82|) \\
 &= 2 \cdot \Pr(Z \geq 1.82) \\
 &= 2 [1 - \Phi(1.82)] \\
 &= 2 (1 - 0.9656) \\
 &= \mathbf{0.0688}
 \end{aligned}$$

As expected, we arrive at the same conclusion to not reject H_0 , since the p -value of 6.88% is greater than the 5% significance level.

One-Tailed Tests

As mentioned, a two-tailed test differs from a one-tailed test by the number of tails included to form the critical region. This is driven by a researcher's specific interest. In the case of a fair coin, **both** extremes of too few or too many heads would suggest the coin may not be fair. Thus, a two-tailed test is appropriate for that scenario. However, there are situations where **only one** extreme is of concern: too extreme to the left, or too extreme to the right, but not both.

For example, consider a factory claiming that the mean number of defective manufactured products per day is 4. The factory would typically be concerned if the mean is in fact much higher than 4, but not concerned if it is lower than 4. A one-tailed test is more suitable in this scenario.

Moreover, note that different forms of the null and alternative hypotheses could result in the same one-tailed test. Despite the nuances in the hypotheses, the core steps of a one-tailed test are the same. In general, to determine which type of test applies, we examine the two hypotheses to identify the appropriate critical region.

RIGHT-TAILED TESTS

Let's alter our assumed setup by only changing the hypotheses. A possible set of hypotheses involving parameter μ and constant h is

$$H_0 : \mu \leq h \quad H_1 : \mu > h$$

An extreme x in the right tail of the normal distribution would support H_1 more than H_0 , but an extreme x in the left tail would not. This means the critical region is in the right tail, given by

$$x \geq b \quad \Rightarrow \quad z \geq c$$

However, there is ambiguity as to how x should be standardized to z , since H_0 now does not assume a fixed value for μ . This is resolved through the formal definition of the significance level, i.e. Equation 2.3.1.1.

$$\max_{\mu} \Pr(Z \geq c \mid H_0 \text{ is true}) = \alpha$$

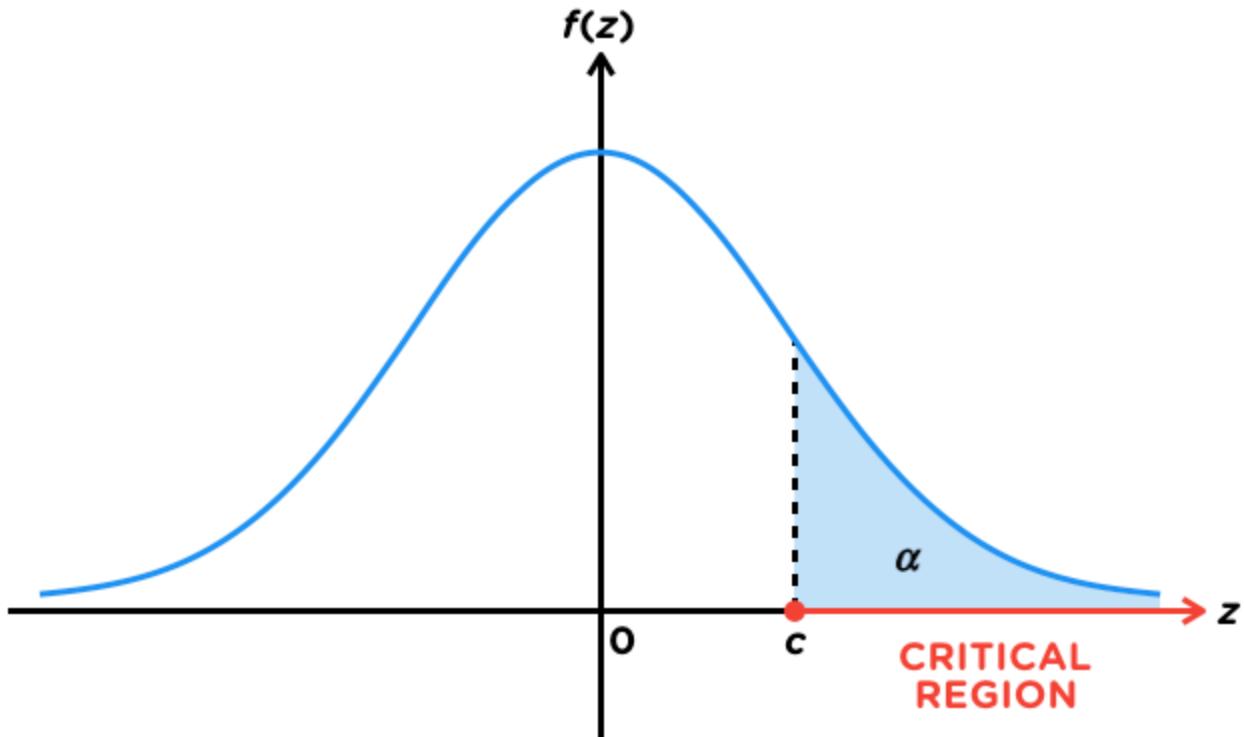
So, the test size is the largest probability for a right-tailed critical region, given H_0 is true. The probability is maximized by changing the value of μ while limited to $\mu \leq h$ (i.e. H_0). It can be

shown that

$$\max_{\mu} \Pr(Z \geq c \mid \mu \leq h) = \Pr(Z \geq c \mid \mu = h)$$

In effect, we operate as though H_0 is $\mu = h$; this is used to standardize x to z .

As seen in the graph below, α is **not** split into two regions. Therefore, for the same α , the critical value here will be different from the critical value in the two-tailed case.



In summary, here are the changes for this case:

- The critical region only involves the right tail of the normal distribution.
- The significance level is $\Pr(Z \geq c \mid H_0 \text{ is true})$.
- The decision logic is:
 - If $z \geq c$ for $\mu = h$, then $z \geq c$ will hold for all other possible μ values in H_0 . Therefore, reject H_0 .
 - If $z < c$ for $\mu = h$, then the test statistic is not extreme enough for some subset of possible μ values in H_0 ; there is not enough evidence to deem H_0 unreasonable. Therefore, do not reject H_0 .

- The p -value is $\Pr(Z \geq z \mid H_0 \text{ is true})$.

The next example is a variation of Example 2.3.1.1.

Example 2.3.1.2

It is believed that the mean age of licensed drivers in 2017 is 43.7. A researcher thinks that the mean age is actually higher. One licensed driver's age is observed to be 60 in 2017. These ages are normally distributed with variance 80.

Test whether the researcher is justified at the 5% significance level.

Solution

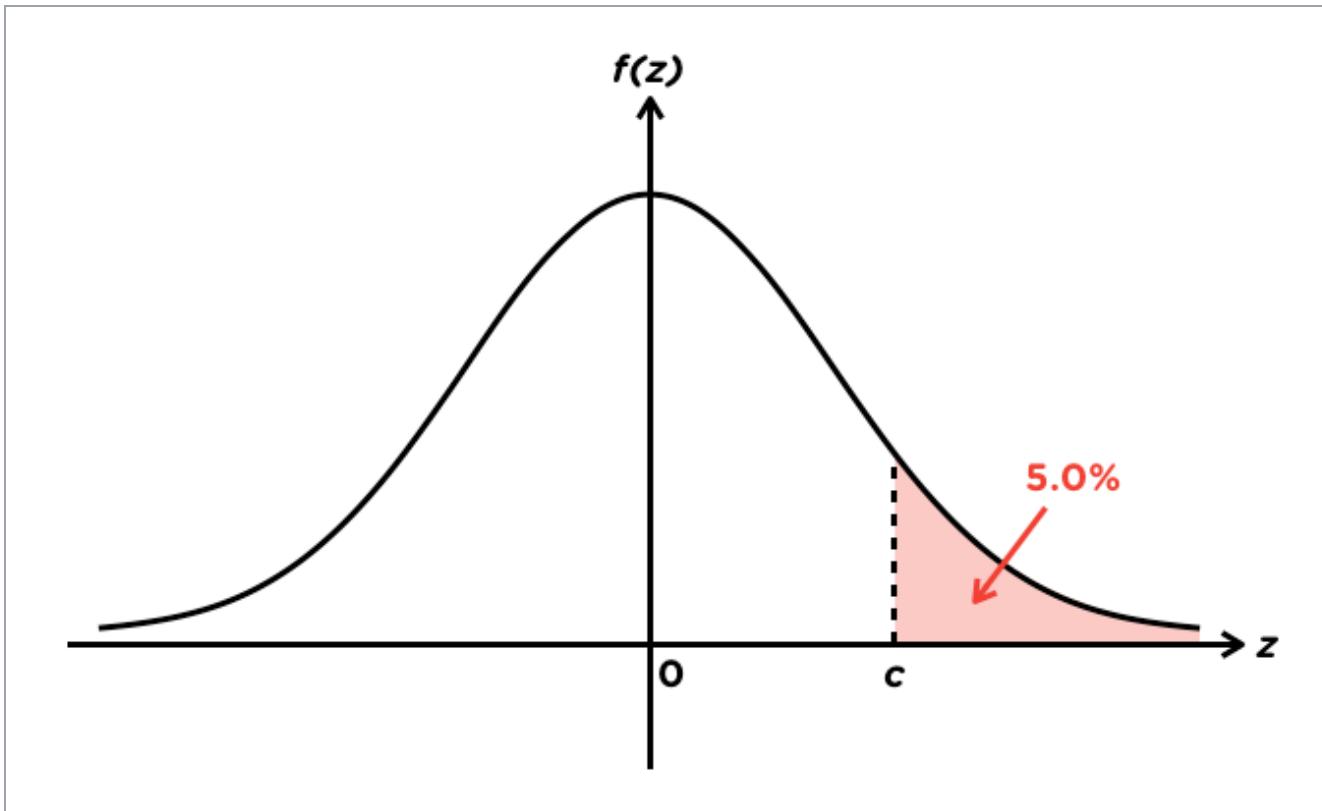
By letting μ represent the mean age of licensed drivers in 2017, the null and alternative hypotheses are

$$H_0 : \mu = 43.7 \quad H_1 : \mu > 43.7$$

Although H_0 is not $\mu \leq 43.7$, this is still a right-tailed test. This is because H_1 would be supported by observing an extreme **older** age, thus justifying the researcher's claim. On the other hand, observing an extreme **younger** age would not justify the researcher's claim.

The critical region is $z \geq c$. Next, determine the critical value. Since $\alpha = 0.05$,

$$\Pr(Z \geq c \mid H_0 \text{ is true}) = 0.05$$



The graph indicates that c is the 95th percentile of Z . From the normal distribution table, $c = z_{0.95} = 1.645$.

Next, calculate the test statistic.

$$z = \frac{x - h}{\sigma} = \frac{60 - 43.7}{\sqrt{80}} = 1.82$$

Since $1.82 > 1.645$, the test statistic falls in the critical region, and thus we **reject the null hypothesis of $\mu = 43.7$ in favor of $\mu > 43.7$ at the 5% significance level.**



Coach's Remarks

Recall that for Example 2.3.1.1, H_0 was not rejected at the same 5% significance level. This is because when including both tails, the driver's age of 60 is not among the "top 5% of extreme values".

However, when including only the right tail, 60 is among the "top 5% of extreme values", so H_0 is rejected here.

LEFT-TAILED TESTS

Now consider the hypotheses

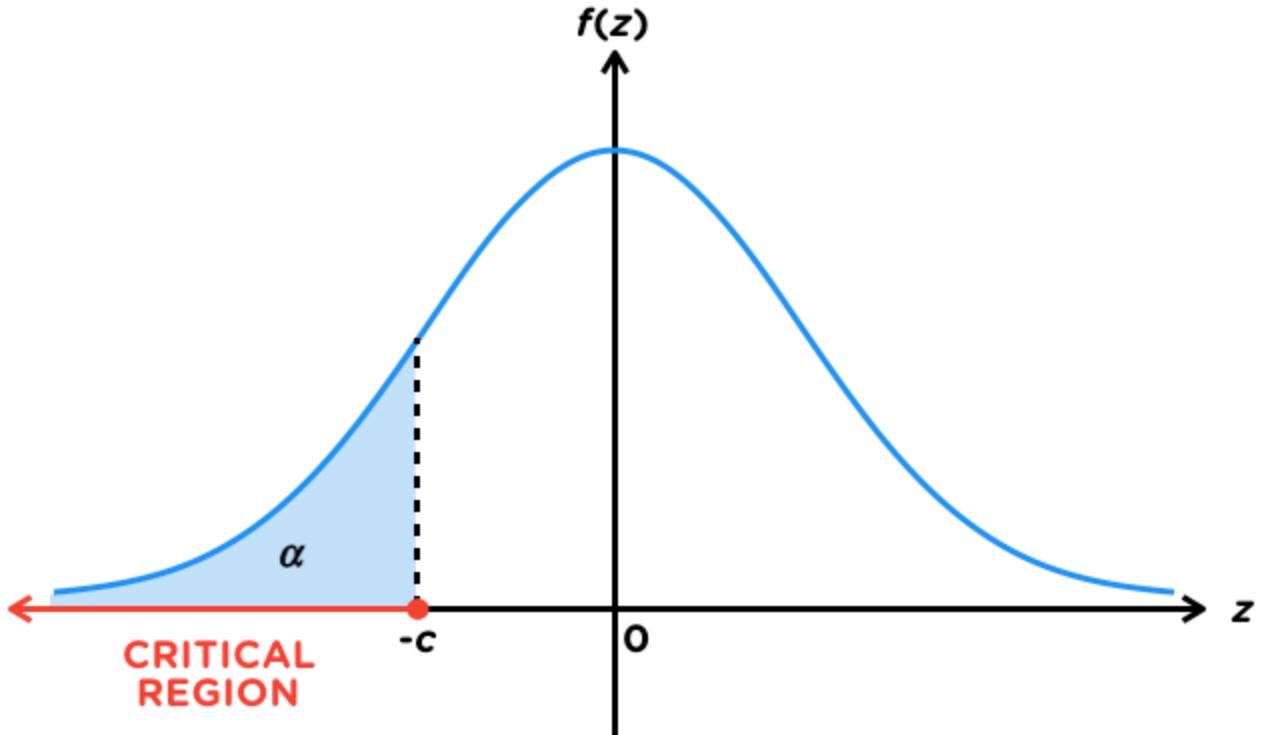
$$H_0 : \mu \geq h \quad H_1 : \mu < h$$

An extreme x in the left tail of the normal distribution would support H_1 more than H_0 , but an extreme x in the right tail would not. This means the critical region is in the left tail, given by

$$x \leq a \quad \Rightarrow \quad z \leq -c$$

where the significance level is

$$\max_{\mu} \Pr(Z \leq -c \mid H_0 \text{ is true}) = \Pr(Z \leq -c \mid \mu = h) = \alpha$$



In summary, here are the changes for this case:

- The critical region only involves the left tail of the normal distribution.

- The significance level is $\Pr(Z \leq -c \mid H_0 \text{ is true})$.
- The decision logic is:
 - If $z \leq -c$ for $\mu = h$, then $z \leq -c$ will hold for all other possible μ values in H_0 . Therefore, reject H_0 .
 - If $z > -c$ for $\mu = h$, then the test statistic is not extreme enough for some subset of possible μ values in H_0 ; there is not enough evidence to deem H_0 unreasonable. Therefore, do not reject H_0 .
- The p -value is $\Pr(Z \leq z \mid H_0 \text{ is true})$.

Example 2.3.1.3

The mean total cholesterol for overweight individuals is said to be at least 250mg/dL. A drug chemist thinks the mean total cholesterol is actually lower. An overweight individual is observed to have a total cholesterol of 208mg/dL. Assume total cholesterol for these individuals is distributed normally with variance v .

At the 3% significance level, determine the largest value of v that would result in supporting the chemist.

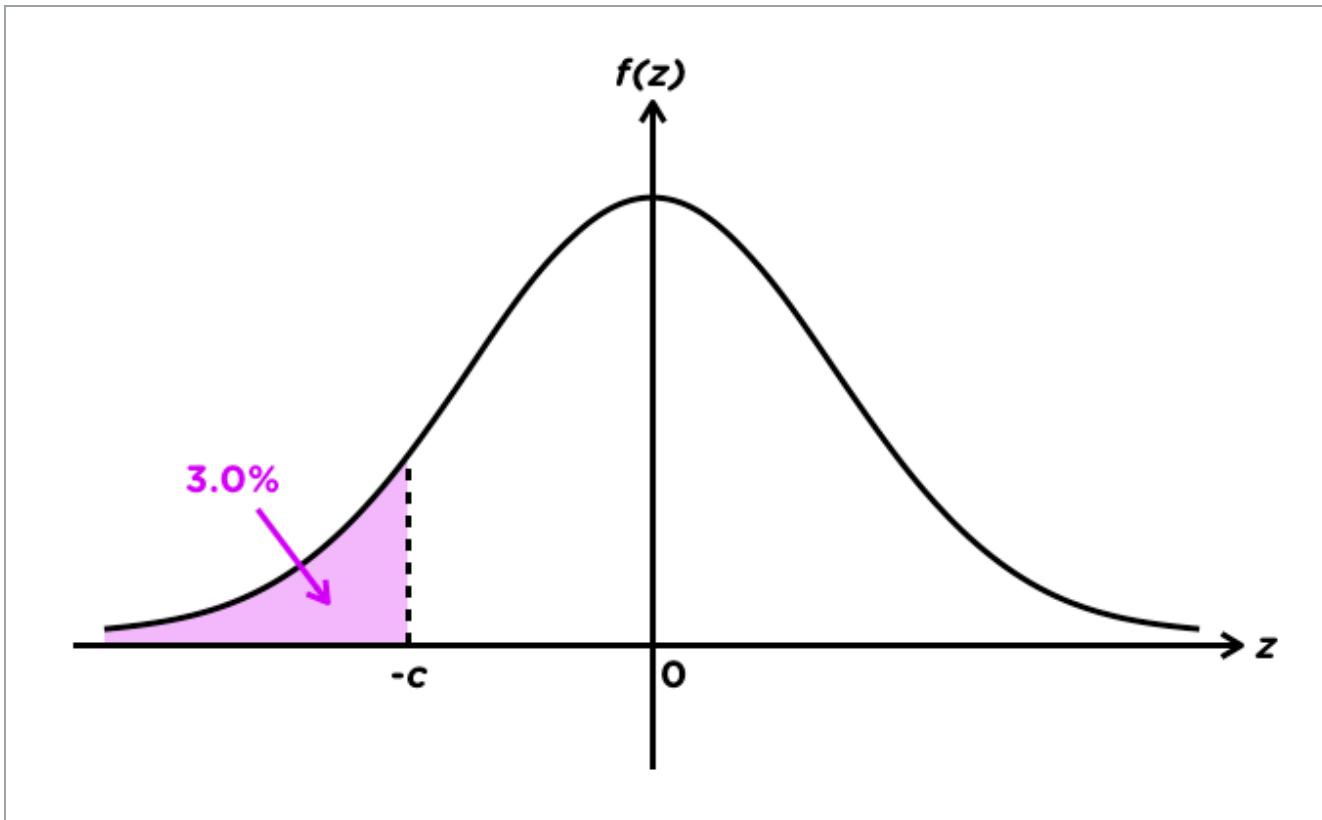
Solution

By letting μ represent the mean total cholesterol of overweight individuals, the null and alternative hypotheses are

$$H_0 : \mu \geq 250 \quad H_1 : \mu < 250$$

Since $\alpha = 0.03$,

$$\Pr(Z \leq -c \mid \mu = 250) = 0.03$$



The graph indicates that c (as opposed to $-c$) would be the 97th percentile of Z . From the normal distribution table, $c = z_{0.97} = 1.88$. In turn, the critical value is $-c = -1.88$.

To support the chemist (i.e. reject H_0), the test statistic must be less than or equal to the critical value. Remember that $\mu = 250$ is used to calculate the test statistic.

$$\begin{aligned}
 \frac{208 - 250}{\sqrt{v}} &\leq -1.88 \\
 -\frac{42}{\sqrt{v}} &\leq -1.88 \\
 \frac{42}{\sqrt{v}} &\geq 1.88 \\
 \sqrt{v} &\leq \frac{42}{1.88} \\
 v &\leq \left(\frac{42}{1.88}\right)^2 \\
 &= 499.0946
 \end{aligned}$$

The largest value of v that would result in rejecting H_0 is approximately **499**.

The following table summarizes the test decisions for the three setups presented thus far, including the critical values in terms of z_q , the $100q^{\text{th}}$ percentile of Z . Remember that z without a subscript is the test statistic, and α is the significance level.

	Left-Tailed	Two-Tailed	Right-Tailed
Critical Region	$z \leq -c$	$ z \geq c$	$z \geq c$
Critical Value	$-z_{1-\alpha}$	$z_{1-\frac{\alpha}{2}}$	$z_{1-\alpha}$

Coach's Remarks

For left-tailed tests, we denote the critical value in terms of the standard normal as $-c$ rather than c . Other resources may use a different notation. For example, a single symbol like c could be used to represent the critical value for all three variants. Ultimately, it is simply a different way to express the same idea.

Note that the three setups were simplistic so that the focus could be the procedure and logic behind hypothesis tests. To further generalize hypothesis tests, consider these possibilities:

- Use a random sample rather than only one random variable.
- The test statistic comes from a distribution that is non-normal.

Let's demonstrate with a few examples.

Example 2.3.1.4

X_1, X_2, \dots, X_{10} is a random sample drawn from a Bernoulli distribution with success probability q . We wish to test the hypotheses

$$H_0 : q = 0.5 \quad H_1 : q > 0.5$$

The critical region is $\sum_{i=1}^{10} x_i \geq c$. The data records 8 successes.

Calculate the p -value of this test.

Solution

Given H_0 , notice that $\sum_{i=1}^{10} X_i$ has a binomial distribution with $m = 10$ and $q = 0.5$. As the test statistic equals 8, the p -value of this test is

$$\Pr\left(\sum_{i=1}^{10} X_i \geq 8 \middle| H_0 \text{ is true}\right) = \binom{10}{8}(0.5)^8(0.5)^2 + \binom{10}{9}(0.5)^9(0.5) + \binom{10}{10}(0.5)^{10}$$

$$= \mathbf{0.055}$$



Coach's Remarks

To be clear, 8 is **not** the critical value c . Instead, 8 is the test statistic because 8 successes is the result observed/calculated from the data. We hope this highlights the distinction between α and the p -value.

Furthermore, we do not take the absolute value of the test statistic because

- this is a right-tailed test, and
- the binomial distribution is not symmetric about 0.

Example 2.3.1.5

For a random sample of size 5 drawn from a Poisson distribution with mean λ , we wish to test the hypotheses

$$H_0 : \lambda = 1 \quad H_1 : \lambda < 1$$

The critical region is $\sum_{i=1}^5 x_i \leq 2$.

Calculate the size of this test.

Solution

Given H_0 , notice that $\sum_{i=1}^5 X_i$ has a Poisson distribution with mean $5 \cdot 1 = 5$. Therefore, the test size is

$$\Pr\left(\sum_{i=1}^5 X_i \leq 2 \mid H_0 \text{ is true}\right) = e^{-5} + e^{-5} \cdot 5 + \frac{e^{-5} \cdot 5^2}{2!} \\ = \mathbf{0.125}$$



2.3.2 Test Errors and Power

Type I and Type II Errors

While we are unable to know for certain if a hypothesis is true, we try to make an informed decision with a hypothesis test. This means there is no certainty that hypothesis test decisions are always right. However, we can assume H_0 to be either true or false, and then consider the impact of making a wrong decision.

A **Type I error** occurs when H_0 is rejected while it is true. The probability of making this error equals the significance level, α . In other words, we are willing to make a wrong decision $100\alpha\%$ of the time when H_0 is true, balanced by the possibility that H_0 is actually false.

A **Type II error** occurs when H_0 fails to be rejected while it is false.

	H_0 is True	H_0 is False
Reject H_0	Type I Error	Correct Decision
Fail to Reject H_0	Correct Decision	Type II Error

In layman's terms, a Type I error is a false positive, while a Type II error is a false negative.

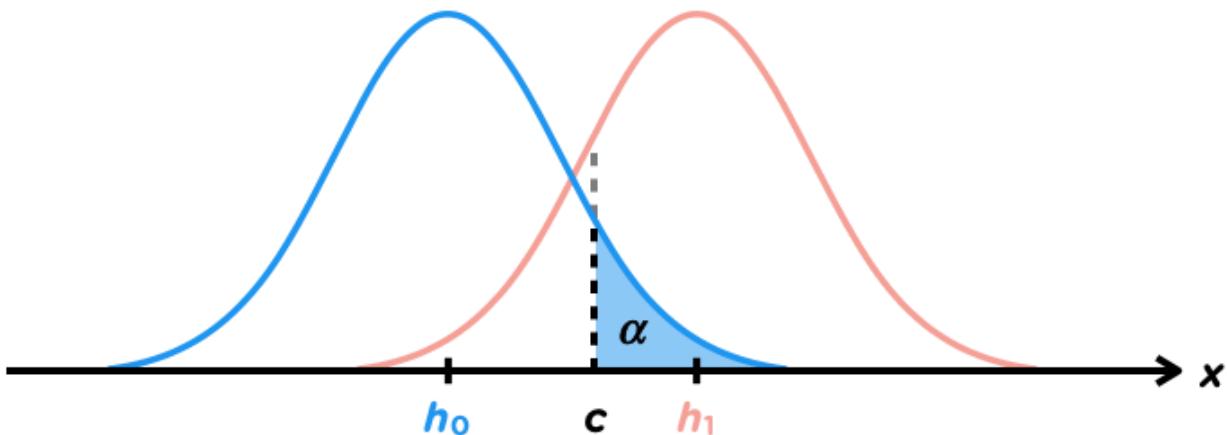
Power of a Test

We would like to make the right decision of rejecting H_0 when it is false. The probability of making this right decision is called the **power of a test**. Hence, the complement to the power is the probability of committing a Type II error.

Note that calculating the power requires an explicit value of the parameter such that H_0 is false. For example, imagine a right-tailed test with hypotheses

$$H_0 : \mu = h_0 \quad H_1 : \mu = h_1$$

for a normal distribution with mean μ . Then, the following animation illustrates the test size and the power of the test:



Let's see some examples on these concepts, beginning with the same scenario from Example 2.3.1.1.

Example 2.3.2.1

It is believed that the mean age of licensed drivers in 2017 is 43.7. To test whether this is the case, one licensed driver's age is observed to be 60 in 2017. These ages are normally distributed with variance 80.

We reject the null hypothesis at the 5% significance level if there is evidence that the mean age of licensed drivers in 2017 differs from 43.7.

If the mean age of licensed drivers in 2017 is in fact 48, calculate the power of the test.

Solution

Recall that, at the 5% significance level, the critical region for this two-tailed test is

$$|z| \geq 1.96$$

where 1.96 is the critical value in the unit of the standardized normal variable. To calculate the power, we need to find the critical values in the original unit of age.

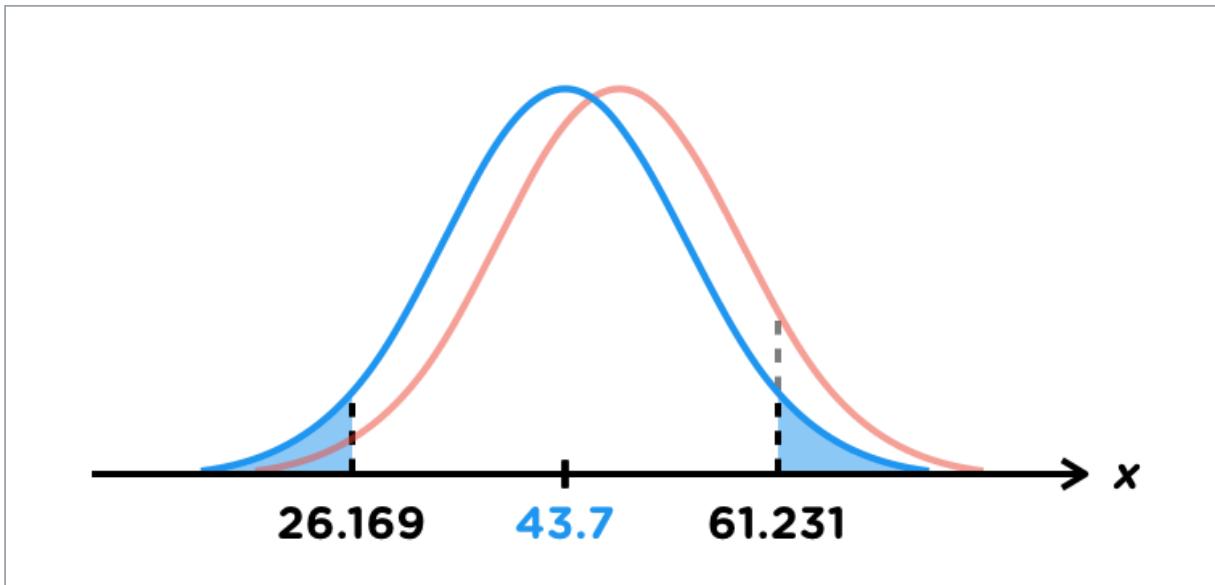
$$\frac{a - 43.7}{\sqrt{80}} = -1.96 \Rightarrow a = 26.169$$

$$\frac{b - 43.7}{\sqrt{80}} = 1.96 \Rightarrow b = 61.231$$

In summary, the critical region for the test is

$$[x \leq 26.169] \cup [x \geq 61.231]$$

The power is the probability of rejecting H_0 , given an explicit value of the parameter (i.e. μ) that contradicts H_0 (i.e. $\mu = 43.7$). We may visualize both the test size and power with the following graph:



Therefore, the answer is

$$\begin{aligned} \Pr([X \leq 26.169] \cup [X \geq 61.231] \mid \mu = 48) &= \Pr\left(\left[Z \leq \frac{26.169 - 48}{\sqrt{80}}\right] \cup \left[Z \geq \frac{61.231 - 48}{\sqrt{80}}\right]\right) \\ &= \Pr([Z \leq -2.44] \cup [Z \geq 1.48]) \\ &= \Pr(Z \leq -2.44) + \Pr(Z \geq 1.48) \\ &= [1 - \Phi(2.44)] + 1 - \Phi(1.48) \\ &= 1 - 0.9927 + 1 - 0.9306 \\ &= \mathbf{0.0767} \end{aligned}$$

Coach's Remarks

Intuitively, this low power makes sense. If $\mu = 48$, then H_0 should be rejected, since $\mu \neq 43.7$. However, there is not a large difference between $\mu = 43.7$ and $\mu = 48$. As a result, the test is unlikely to detect the distinction, and thus unlikely to correctly reject H_0 .

Example 2.3.2.2

X is normally distributed with mean μ and variance 22. The following set of hypotheses are considered:

$$H_0 : \mu = 5 \quad H_1 : \mu = 10$$

With a single observation and at the 5% significance level, calculate the power of this test.

Solution

Based on H_0 and H_1 , realize this is a right-tailed test. This is because the μ hypothesized by H_1 is larger than the one by H_0 .

First, determine the critical value in the unit of X . For $\alpha = 0.05$, we previously saw that $c = 1.645$ is the critical value in the unit of the standardized normal variable. Therefore,

$$\frac{b - 5}{\sqrt{22}} = 1.645 \quad \Rightarrow \quad b = 12.716$$

Hence, the critical region is $x \geq 12.716$.

It is implied that the power is calculated based on H_1 . This produces the answer of

$$\begin{aligned}
\Pr(X \geq 12.716 \mid \mu = 10) &= \Pr\left(Z \geq \frac{12.716 - 10}{\sqrt{22}}\right) \\
&= \Pr(Z \geq 0.58) \\
&= 1 - \Phi(0.58) \\
&= 1 - 0.719 \\
&= \mathbf{0.281}
\end{aligned}$$

■

Example 2.3.2.3

A random sample of size n is taken from a gamma distribution with shape parameter $\alpha = 3$ and scale parameter θ . We wish to test the hypotheses:

$$H_0 : \theta = 4 \quad H_1 : \theta > 4$$

The critical region is $\bar{x} \geq 14.5$, where \bar{x} is the sample mean.

Calculate the minimum sample size required such that the probability of a Type I error is at most 4% using normal approximation.

Solution

For this random sample,

$$\mathbb{E}[\bar{X}] = \mathbb{E}[X] = \alpha\theta = 12$$

$$\text{Var}[\bar{X}] = \frac{\text{Var}[X]}{n} = \frac{\alpha\theta^2}{n} = \frac{48}{n}$$

given H_0 is true.

Recall that the probability of a Type I error is the significance level. As a result, using normal approximation produces

$$\begin{aligned}\Pr(\bar{X} \geq 14.5 \mid H_0 \text{ is true}) &= \Pr\left(Z \geq \frac{14.5 - 12}{\sqrt{48/n}}\right) \\ &= 1 - \Phi\left(\frac{2.5}{\sqrt{48/n}}\right) \\ &\leq 0.04\end{aligned}$$

$$\begin{aligned}\Phi\left(\frac{2.5}{\sqrt{48/n}}\right) &\geq 0.96 \\ \frac{2.5}{\sqrt{48/n}} &\geq 1.75 \\ \sqrt{n} &\geq 1.75 \frac{\sqrt{48}}{2.5} \\ n &\geq 4.8497^2 \\ &= 23.52\end{aligned}$$

Therefore, the minimum sample size required is **24** because it is the smallest integer that satisfies $n \geq 23.52$.



2.3.3 Tests for Means

There are scenarios where a default test statistic and critical region are assumed, and thus, an exam problem need not specify them. This and the next two subsections cover these scenarios. We start with hypothesis tests that examine a distribution's mean. There are three broad categories to discuss:

1. One sample
2. Two samples
3. Two samples with paired observations

One Sample

For the case of one random sample of size n , we can divide it further into two sub-cases: known variance and unknown variance.

KNOWN VARIANCE

Let the random sample be drawn from **any** distribution with mean μ and known variance σ^2 . When n is large, then by the Central Limit Theorem, \bar{X} is approximately normally distributed. As a reminder,

$$\mathbb{E}[\bar{X}] = \mu, \quad \text{Var}[\bar{X}] = \frac{\sigma^2}{n}$$

for a random sample. Consequently,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has an approximate standard normal distribution. Therefore, to test hypotheses involving μ , we use the test statistic

$$t. s. = \frac{\bar{x} - h}{\sigma / \sqrt{n}} \quad (2.3.3.1)$$

where h is the hypothesized value of μ according to H_0 . The test decisions for this scenario can be summarized as:

	Left-Tailed	Two-Tailed	Right-Tailed
Critical Region	$t. s. \leq -z_{1-\alpha}$	$ t. s. \geq z_{1-\frac{\alpha}{2}}$	$t. s. \geq z_{1-\alpha}$

where α is the significance level, and z_q is the $100q^{\text{th}}$ percentile of the standard normal distribution, which can be calculated using the Excel formula:

`NORM.S.INV(q)`

UNKNOWN VARIANCE

When the variance σ^2 is not known, the test statistic of Equation 2.3.3.1 cannot be calculated. So, an adjustment is needed to obtain a different, yet similar, distribution.

Let the random sample be drawn from a **normal** distribution with mean μ and variance σ^2 . Recall that S^2 denotes the unbiased sample variance (Equation 2.2.1.2). Then,

$$\frac{\bar{X} - \mu}{S / \sqrt{n}}$$

follows a t -distribution with $n - 1$ degrees of freedom. This is not the standard normal Z because the unbiased sample standard deviation replaces the true standard deviation, σ . The degrees of freedom is derived from the denominator of S^2 . Moreover, n does not need to be large since we are not invoking the Central Limit Theorem.

The **t -distribution** is very similar to the standard normal distribution, as it has a bell-shaped density and is symmetric about 0. Its single parameter is called degrees of freedom; as it approaches ∞ , the t -distribution converges to the standard normal distribution.

The $100q^{\text{th}}$ percentile of a t -distribution with df degrees of freedom can be calculated using the Excel formula:

`T.INV(q, df)`

Coach's Remarks

Alternatively we can determine the percentiles using the t -distribution table provided in the exam table, which is shown below:

Tail areas (two-sided) for t-distributions

df	0.20	0.10	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
45	1.301	1.679	2.014	2.412	2.690
50	1.299	1.676	2.009	2.403	2.678
55	1.297	1.673	2.004	2.396	2.668
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.290	1.660	1.984	2.364	2.626
120	1.289	1.658	1.980	2.358	2.617
400	1.284	1.649	1.966	2.336	2.588
Inf	1.282	1.645	1.960	2.326	2.576

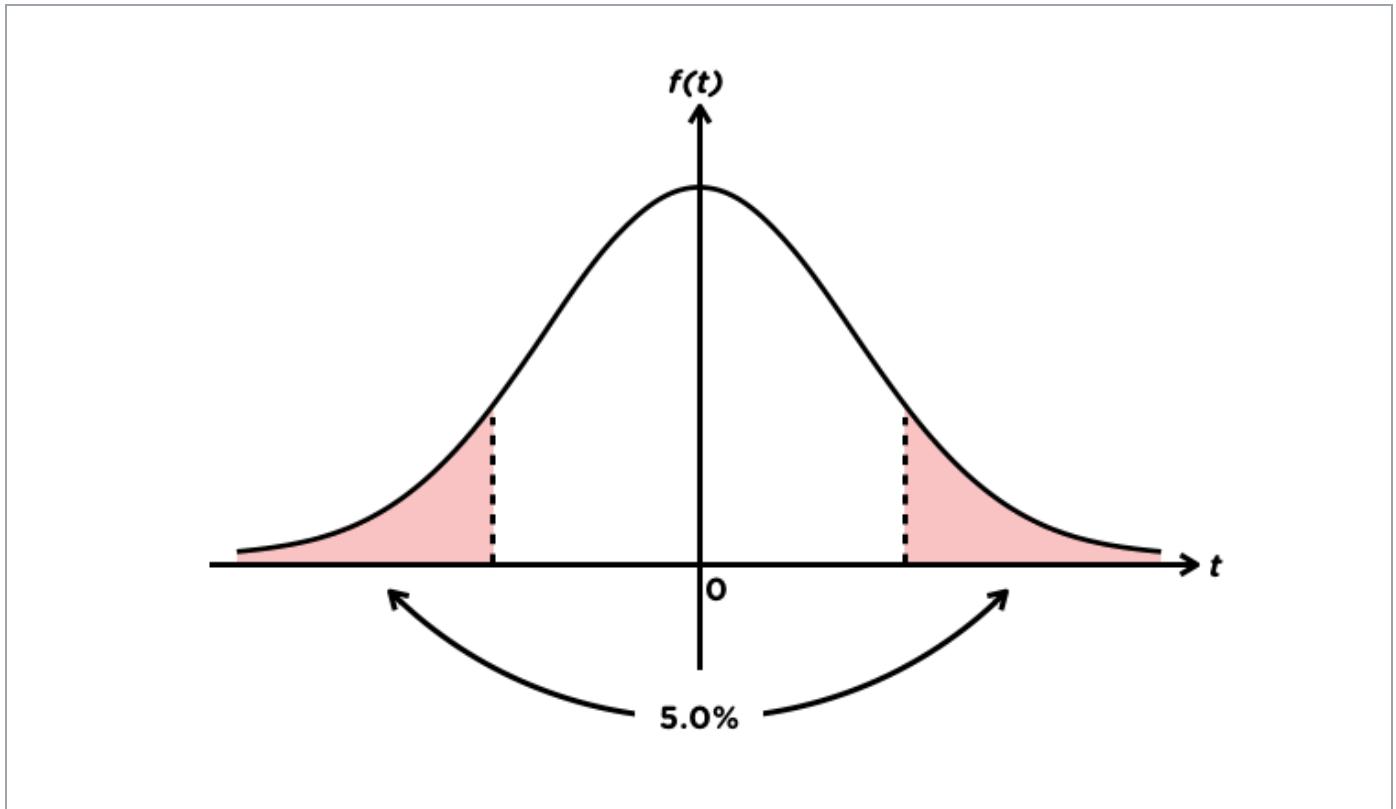
We use $t_{2(1-q), df}$ to denote the $100q^{\text{th}}$ percentile of a t -distribution with df degrees of freedom. Note that $2(1 - q)$ represents the probability in **both tails** of the distribution. Therefore, the

table indicates that 2.086 is the 97.5th percentile of a t -distribution with 20 degrees of freedom, i.e. $t_{0.05, 20} = 2.086$.

This matches the result from the Excel formula:

$$\text{T.INV}(0.975, 20) = 2.086$$

A graphical representation is shown below.



Thus, when a random sample is drawn from a normal distribution with unknown σ^2 , we use the test statistic

$$t.s. = \frac{\bar{x} - h}{s/\sqrt{n}} \quad (2.3.3.2)$$

where h is the hypothesized μ as per H_0 . Remember that this test statistic comes from a t -distribution with $n - 1$ degrees of freedom. To clarify, s^2 is the version of Equation 2.2.1.2 that is calculated from the sample, i.e.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (2.3.3.3)$$

Since both standard normal and t have densities that are symmetric about 0, the test decisions are the same as before, except t percentiles are used instead of standard normal percentiles. Specifically, the critical region forms are the same, and the critical values are the following t percentiles:

	Left-Tailed	Two-Tailed	Right-Tailed
Critical Value	$-t_{2\alpha, n-1}$	$t_{\alpha, n-1}$	$t_{2\alpha, n-1}$

Coach's Remarks

Pay attention to the subscripts we use for the standard normal and t percentiles:

- $z_q \rightarrow 100q^{\text{th}}$ percentile of the standard normal distribution
- $t_{2(1-q), \text{df}} \rightarrow 100q^{\text{th}}$ percentile of a t -distribution with df degrees of freedom

Although the subscripts have two different expressions of q , both of these percentiles correspond to the **same** cumulative probability q . The subscript in z contains the cumulative probability, whereas the subscript in t contains the probability in both tails. This is done intentionally to match how the probabilities are stated in their respective distribution tables.

As a result, note that

$$\begin{aligned} q = 1 - \frac{\alpha}{2} &\Leftrightarrow 2(1 - q) = \alpha && \text{(two-tailed)} \\ q = 1 - \alpha &\Leftrightarrow 2(1 - q) = 2\alpha && \text{(one-tailed)} \end{aligned}$$

Example 2.3.3.1

A study states that the height of North Linden residents has a mean of 1.67 and a variance of 0.5. Believing that the mean is in fact higher, scientists conduct a hypothesis test by taking a random sample of 50 residents and measuring their heights. The sample mean of the 50 heights is 1.86.

Determine the test result at the 3.5% significance level.

Solution

For μ representing the mean height of North Linden residents, the hypotheses are

$$H_0 : \mu = 1.67 \quad H_1 : \mu > 1.67$$

We are given

$$n = 50, \quad \bar{x} = 1.86, \quad \sigma^2 = 0.5$$

Notice that the problem gives the value of σ^2 , suggesting that a standard normal percentile should be used for this right-tailed test. At $\alpha = 0.035$, the critical value is $z_{0.965}$, which equals 1.81 from the normal distribution table.

Using Equation 2.3.3.1, the test statistic is

$$\frac{\bar{x} - h}{\sigma / \sqrt{n}} = \frac{1.86 - 1.67}{\sqrt{0.5} / \sqrt{50}} = 1.9$$

Since $1.9 > 1.81$, we **reject H_0 at the 3.5% significance level.**



Example 2.3.3.2

An automated candy dispenser is used in a factory line. The weight of dispensed candy per bag is normally distributed with an assumed mean of 46oz. Through a hypothesis test, a technician performs a routine check to see if the mean dispensed weight differs from 46oz by drawing a random sample of six bags. The following are the sample weights:

42 46 43 43 44 46

Determine which statements are true.

- I. The hypothesis test is a two-tailed test.
- II. The hypothesis test is a t test.
- III. The p -value for this test is less than 5%.

Solution

Statement I is true. For μ representing the mean dispensed weight, the hypotheses are

$$H_0 : \mu = 46 \quad H_1 : \mu \neq 46$$

Thus, both tails are included in the critical region.

Statement II is true. A t test is any hypothesis test where the test statistic comes from a t -distribution. In this problem, a random sample from a normal distribution with unknown variance is taken. Since we are testing a mean parameter, all the requirements are met such that

$$\frac{\bar{x} - 46}{s/\sqrt{6}}$$

from Equation 2.3.3.2 is the test statistic that comes from a t -distribution with $6 - 1 = 5$ degrees of freedom.

Statement III is true. First, compute \bar{x} and s from the observations.

$$\bar{x} = \frac{42 + 46 + 43 + 43 + 44 + 46}{6} = 44$$

$$s^2 = \frac{(42 - 44)^2 + (46 - 44)^2 + \dots + (46 - 44)^2}{6 - 1} = 2.8$$

This produces the test statistic of

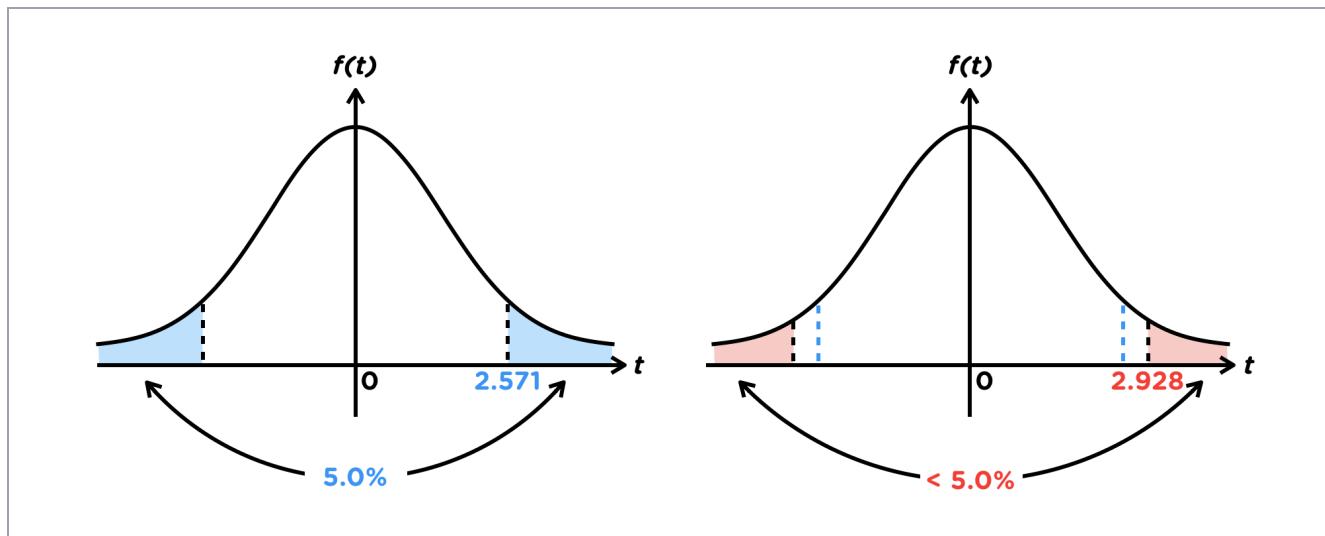
$$\frac{44 - 46}{\sqrt{2.8}/\sqrt{6}} = -2.928$$

Let T denote a random variable that follows a t -distribution with 5 degrees of freedom. Hence, the p -value for this two-tailed test is

$$\begin{aligned}\Pr(|T| \geq |t.s.|) &= \Pr(|T| \geq |-2.928|) \\ &= \Pr(|T| \geq 2.928)\end{aligned}$$

Recall that this probability expression is the probability in both tails of the t -distribution, i.e. $\Pr(|T| \geq t) = \Pr([T \leq -t] \cup [T \geq t])$ for any positive constant t . Thus, note that the t -distribution table gives $\Pr(|T| \geq 2.571) = 0.05$. As a result,

$$\Pr(|T| \geq 2.928) < \Pr(|T| \geq 2.571) = 0.05$$



Therefore, all of the statements are true. ■

Coach's Remarks

An alternative to proving Statement III is to pretend the 5% refers to a significance level. Doing so makes the statement equivalent to saying that H_0 is rejected at

$\alpha = 0.05$. With a test statistic of -2.928 and critical value of $t_{0.05, 5} = 2.571$, we would reject H_0 since $|-2.928| > 2.571$.

Coach's Remarks

Alternatively, perform the test using Excel:

1. Enter the sample data into range A1:A6
2. Calculate the sample mean and standard deviation:
 - $\bar{x} = \text{AVERAGE}(A1:A6) = 44$
 - $s = \text{STDEV.S}(A1:A6) = 1.6733$
3. Calculate the test statistic: $(44 - 46) / (1.6733 / \sqrt{6}) = -2.9277$
4. Calculate the p -value using either of the following:
 - $2 * \text{T.DIST}(-2.9277, 5, 1) = 0.0327$
 - $\text{T.DIST.2T}(2.9277, 5) = 0.0327$

The p -value is 3.27%, which is less than 5%.

Before continuing, it is helpful to note the following generalization:

In testing a mean parameter, the test statistic has the form

$$\frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}}$$

which comes from either the standard normal distribution or a t -distribution. In the one-sample case, we are using \bar{X} as an estimator of μ , which leads to \bar{x} as the test statistic's estimate.

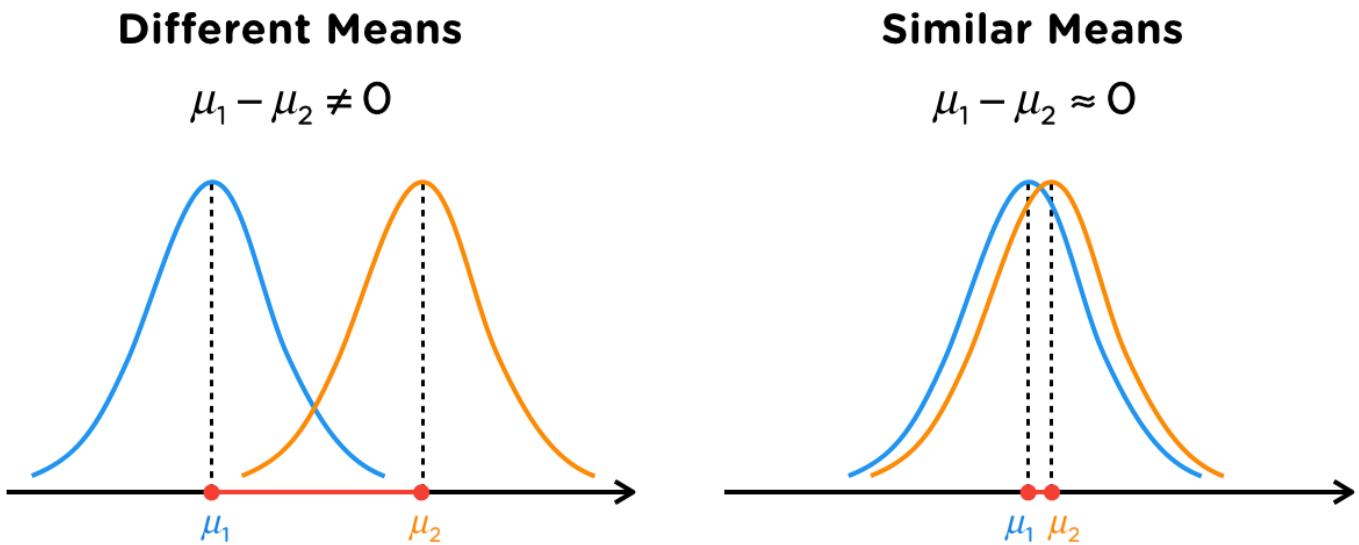
Standard error is the term for the standard deviation of an estimator. However, the test statistic's standard error is a placeholder for two possibilities:

1. The **true** standard error. This implies the known variance scenario, and hence, the standard normal distribution is used for testing.

2. An **estimated** standard error. This implies the unknown variance scenario, where the unknown variance is replaced by an appropriate unbiased sample variance, and hence, a t -distribution is used for testing. The t -distribution's degrees of freedom is equal to the denominator of the appropriate unbiased sample variance.

Two Samples

To detect whether there is a difference between two groups, we may take a sample from each group and examine whether they have similar or different means. Note that "equal means" is equivalent to "the difference in means equals 0". Consequently, we are more interested in the difference in means rather than their individual values. The following diagram illustrates how the difference in means can capture whether two samples are drawn from distinct or identical distributions.



Assume there are two random samples with sizes n_1 and n_2 , respectively. Furthermore, assume independence between the two samples. Let μ_k , σ_k^2 , and \bar{X}_k denote the mean, variance, and sample mean corresponding to Sample k .

Similar to the one-sample case, there are two sub-cases to discuss: known variances and unknown variances.

KNOWN VARIANCES

For large values of n_1 and n_2 , the Central Limit Theorem states that \bar{X}_1 and \bar{X}_2 are each approximately normally distributed.

Recall that we are interested in investigating $\mu_1 - \mu_2$. To apply the generalization for a mean parameter, we should view $\mu_1 - \mu_2$ itself as a mean, and thus, use $\bar{X}_1 - \bar{X}_2$ as its estimator. Note that

$$E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$$

$$\begin{aligned} \text{Var}[\bar{X}_1 - \bar{X}_2] &= \text{Var}[\bar{X}_1] + (-1)^2 \text{Var}[\bar{X}_2] + 2(-1)\text{Cov}[\bar{X}_1, \bar{X}_2] \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} + 0 \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

where $\bar{X}_1 - \bar{X}_2$ is also approximately normally distributed by the Central Limit Theorem. As expected from the Central Limit Theorem, this holds for **any** two distributions from which the two random samples are taken.

Therefore, to test hypotheses involving $\mu_1 - \mu_2$, we use the test statistic

$$\begin{aligned} t.s. &= \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}} \\ &= \frac{\bar{x}_1 - \bar{x}_2 - h}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \end{aligned} \tag{2.3.3.4}$$

which comes from the standard normal distribution. We alluded to the fact that $h = 0$ if we wish to test whether the two samples have similar versus different means, but the test also works for other values of h . The test decisions are exactly the same as in the one-sample case. As a reminder, we assume that the values of σ_1^2 and σ_2^2 are known in this scenario.

UNKNOWN VARIANCES

Given unknown σ_1^2 and σ_2^2 , it should be no surprise that we wish to use the t -distribution. However, additional assumptions are required. Specifically, assume

- Sample 1 and Sample 2 are each taken from a **normal** distribution, and

- $\sigma_1^2 = \sigma_2^2$.

The big change is the assumption of equal variances between the two samples; we will denote the variances simply as σ^2 since they are equal. As a result,

$$\text{Var}[\bar{X}_1 - \bar{X}_2] = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Next, define s_p^2 as the **pooled sample variance**, where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (2.3.3.5)$$

where s_k^2 is the calculated unbiased sample variance from Sample k . The pooled sample variance is an unbiased sample variance for σ^2 , so it replaces the unknown σ^2 in this test. As a result, the test statistic

$$\begin{aligned} t. s. &= \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}} \\ &= \frac{\bar{x}_1 - \bar{x}_2 - h}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \end{aligned} \quad (2.3.3.6)$$

comes from a t -distribution with $n_1 + n_2 - 2$ degrees of freedom. Recall that the degrees of freedom equals the denominator of the appropriate unbiased sample variance used (i.e. Equation 2.3.3.5 for this test). The test decisions are exactly the same as in the one-sample case, except the t percentiles are based on $n_1 + n_2 - 2$ degrees of freedom rather than $n - 1$.

Example 2.3.3.3

You are given the following data observed from two random samples:

Sample 1	6	8	9	9
Sample 2	6	7	8	9

- Each sample is drawn from a normal distribution with a common variance.
- The two samples are independent.
- You want to test the hypothesis:
 - H_0 : The Sample 1 and Sample 2 means are equal
 - H_1 : The Sample 1 mean is greater than the Sample 2 mean

Determine the test result at $\alpha = 0.1$.

Solution

For μ_k representing the mean of the normal distribution for Sample k , the hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_1 : \mu_1 - \mu_2 > 0$$

The variances of the two normal distributions are unknown, but we are told that they are common. Therefore, we should conduct a t test.

First, compute the sample means \bar{x}_1 and \bar{x}_2 .

$$\bar{x}_1 = \frac{6 + 8 + 9 + 9}{4} = 8, \quad \bar{x}_2 = \frac{6 + 7 + 8 + 9}{4} = 7.5$$

Next, calculate the pooled sample variance (Equation 2.3.3.5). Realize that $(n_k - 1)s_k^2$ is another way to express the sum of squared deviations from \bar{x}_k , i.e.

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

As a result, it is quicker to solve for the pooled sample variance as

$$s_p^2 = \frac{\left[(6-8)^2 + \dots + (9-8)^2\right] + \left[(6-7.5)^2 + \dots + (9-7.5)^2\right]}{4+4-2} = \frac{11}{6}$$

Using Equation 2.3.3.6, the test statistic is

$$\frac{\bar{x}_1 - \bar{x}_2 - h}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{8 - 7.5 - 0}{\sqrt{\frac{11}{6}} \cdot \sqrt{\frac{1}{4} + \frac{1}{4}}} = 0.522$$

For this right-tailed test, the critical value is $t_{2\alpha, n_1+n_2-2} = t_{0.2, 6} = 1.440$. Since $0.522 < 1.440$, we **fail to reject H_0** at $\alpha = 0.1$, which suggests there is evidence that the two means are equal.



Coach's Remarks

Alternatively, perform the test using Excel:

1. Enter the sample data into the ranges A1:A4 and B1:B4.
2. Calculate the p -value the Excel function T.TEST(A1:A4, B1:B4, 1, 2) = 0.3101, where:
 - A1:A4 and B1:B4 contain the two sample datasets.
 - 1 specifies a one-tailed test.
 - 2 specifies a two-sample test assuming equal variance.

Since the p -value = 0.3101 is greater than the significance level $\alpha = 0.10$, we fail to reject the null hypothesis.

Paired Observations

Some experiments are designed such that the two random samples are not independent from each other. One example is when Sample 2 consists of the same objects from Sample 1 after undergoing

some special treatment. This allows a researcher to assess whether or not the treatment is effective. The key is that the observations between the samples are **paired** in some way, thus creating a dependence between the samples.

Given the pairing of observations, both samples will have the same size, i.e. $n_1 = n_2$; we will denote the sample sizes simply as n_* since they are equal. Let i be the index for the observation pairs, where $i = 1, 2, \dots, n_*$.

Let D_i denote the difference in the i^{th} pair of X 's. In addition, let

$$\mu_D = \mathbb{E}[D_i], \quad \sigma_D^2 = \text{Var}[D_i]$$

For more information on σ_D^2 , see the appendix at the end of the section.

Although we are investigating $\mu_1 - \mu_2$, the paired structure leads to a one-sample framework because $\mu_D = \mu_1 - \mu_2$. Specifically, the sample mean of the differences, \bar{D} , is used as an estimator of $\mu_1 - \mu_2$, having a mean and variance of

$$\mathbb{E}[\bar{D}] = \mu_D = \mu_1 - \mu_2, \quad \text{Var}[\bar{D}] = \frac{\sigma_D^2}{n_*}$$

So, if σ_D^2 is known and the Central Limit Theorem applies (large n_* , etc), then we have the test statistic

$$t. s. = \frac{\bar{d} - h}{\sigma_D / \sqrt{n_*}} \tag{2.3.3.7}$$

which comes from the standard normal distribution. But, if σ_D^2 is unknown and the D_i 's are normally distributed, then we have the test statistic

$$t. s. = \frac{\bar{d} - h}{s_D / \sqrt{n_*}} \tag{2.3.3.8}$$

which comes from a t -distribution with $n_* - 1$ degrees of freedom. It should be clear that Equations 2.3.3.7 and 2.3.3.8 are just a specific application of Equations 2.3.3.1 and 2.3.3.2.

Coach's Remarks

On a technical note, simply having each sample drawn from a normal distribution does **not** guarantee normally distributed D_i 's. One way to obtain normally distributed D_i 's is when each pair follows a bivariate normal distribution. However, past exams have not shown this attention to detail, so you may treat "each sample is drawn from a normal distribution" as a reasonable proxy to mean that the D_i 's are normally distributed.

Let's reuse the data from Example 2.3.3.3 in the next example.

Example 2.3.3.4

You are given the following data observed from two random samples with paired observations:

Pair	1	2	3	4
Sample 1	6	8	9	9
Sample 2	6	7	8	9

- Each sample is drawn from a normal distribution.
- The two samples are not independent.
- You want to test the hypothesis:
 - H_0 : The Sample 1 and Sample 2 means are equal
 - H_1 : The Sample 1 mean is greater than the Sample 2 mean

Determine the inequality that describes the p -value of this test based on the given t -distribution table.

Solution

For μ_k representing the mean of the normal distribution for Sample k , the hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_1 : \mu_1 - \mu_2 > 0$$

Note that:

- The observations are paired.
- Each sample is drawn from a normal distribution; for this exam, it sufficiently indicates that the differences between each pair are normally distributed as well.
- The variance of the differences is not known.

Therefore, we should conduct a t test. First, compute the sample mean of the differences, \bar{d} , and the unbiased sample variance of the differences, s_D^2 .

$$\bar{d} = \frac{\overbrace{(6 - 6)}^0 + \overbrace{(8 - 7)}^1 + \overbrace{(9 - 8)}^1 + \overbrace{(9 - 9)}^0}{4} = 0.5$$

$$s_D^2 = \frac{(0 - 0.5)^2 + (1 - 0.5)^2 + (1 - 0.5)^2 + (0 - 0.5)^2}{4 - 1} = \frac{1}{3}$$

Therefore, the test statistic is

$$\frac{\bar{d} - h}{s_D / \sqrt{n_*}} = \frac{0.5 - 0}{\sqrt{\frac{1}{3}} / \sqrt{4}} = 1.732$$

and it comes from a t -distribution with $4 - 1 = 3$ degrees of freedom.

Let T denote a random variable that follows a t -distribution with 3 degrees of freedom. Hence, the p -value for this right-tailed test is

$$\Pr(T \geq t.s.) = \Pr(T \geq 1.732)$$

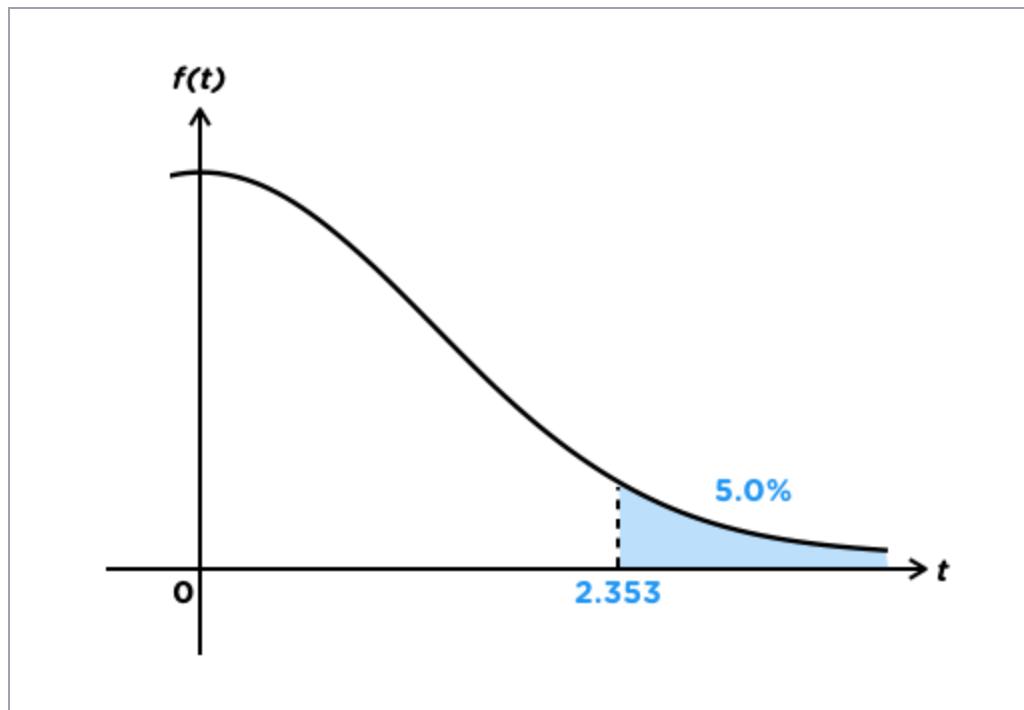
Remember that the five probabilities listed in the t -distribution table refer to the probability in both tails. In other words,

$$\Pr(|T| \geq 1.638) = 0.2 \quad \Rightarrow \quad \Pr(T \geq 1.638) = \frac{0.2}{2} = 0.1$$
$$\Pr(|T| \geq 2.353) = 0.1 \quad \Rightarrow \quad \Pr(T \geq 2.353) = \frac{0.1}{2} = 0.05$$

In conclusion,

$$\Pr(T \geq 2.353) < \Pr(T \geq 1.732) < \Pr(T \geq 1.638)$$

$$\Rightarrow 0.05 < p\text{-value} < 0.1$$



Coach's Remarks

Note that the inequality implies that, at $\alpha = 0.1$, we would reject H_0 for this test. In Example 2.3.3.3, recall that we failed to reject H_0 instead. The main driver for this change is the standard error component of the test statistic. It is smaller in this example, thus increasing the test statistic. At $\alpha = 0.1$, the increase was significant enough for the test statistic to be considered extreme.

The following table summarizes the cases and their test statistics in this subsection.

Number of Samples	Variance	Test Statistic
One	Known	$\frac{\bar{x} - h}{\sigma / \sqrt{n}}$
One	Unknown	$\frac{\bar{x} - h}{s / \sqrt{n}}$
Two	Known	$\frac{\bar{x}_1 - \bar{x}_2 - h}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
Two	Unknown	$\frac{\bar{x}_1 - \bar{x}_2 - h}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
Two, Paired	Known	$\frac{\bar{d} - h}{\sigma_D / \sqrt{n_*}}$
Two, Paired	Unknown	$\frac{\bar{d} - h}{s_D / \sqrt{n_*}}$

2.3.4 Tests for Proportions

The parameter q of a Bernoulli distribution is often of interest since it represents the proportion of a population that is considered a "success". Example 2.3.1.4 demonstrates how the sum of the random sample is a test statistic that comes from a binomial distribution. However, binomial is impractical for moderate to large sample sizes, given its discrete nature.

Recall that q is the mean of a Bernoulli distribution. Therefore, many concepts on testing means from the previous subsection do apply, but there are important caveats. The Bernoulli variance is $q(1 - q)$, which counts as an unknown variance. However, the random sample is drawn from a Bernoulli distribution rather than a normal distribution, thus disallowing the use of a t -distribution.

Instead, we depend on the Central Limit Theorem and use the standard normal distribution for testing. We also address how the variance should be handled.

One Sample

When n is large, \bar{X} is approximately normally distributed by the Central Limit Theorem, where in this case,

$$\mathbb{E}[\bar{X}] = q, \quad \text{Var}[\bar{X}] = \frac{q(1 - q)}{n}$$

While the estimate is \bar{x} , it is common to denote it as \hat{q} instead. Since we are sampling from a Bernoulli distribution, realize that

$$\hat{q} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\# \text{ of successes from } n \text{ trials}}{n}$$

In short, we use the test statistic

$$t.s. = \frac{\hat{q} - h}{\sqrt{\frac{h(1-h)}{n}}} \tag{2.3.4.1}$$

which comes from a standard normal distribution.

Two Samples

Now consider investigating $q_1 - q_2$, the difference in the success proportions of two independent samples. For large values of n_1 and n_2 , $\bar{X}_1 - \bar{X}_2$ is approximately normally distributed by the Central Limit Theorem, where in this case,

$$\mathbb{E}[\bar{X}_1 - \bar{X}_2] = q_1 - q_2, \quad \text{Var}[\bar{X}_1 - \bar{X}_2] = \frac{q_1(1-q_1)}{n_1} + \frac{q_2(1-q_2)}{n_2}$$

In the one-sample case, we handle the variance by letting q equal the hypothesized value h . However, we need values for both q_1 and q_2 in this setup; h does not help since it is a value of their difference. To resolve this, we use the test statistic

$$t.s. = \frac{\hat{q}_1 - \hat{q}_2 - h}{\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}} \quad (2.3.4.2)$$

which comes from a standard normal distribution.

Example 2.3.4.1

A block of business is said to have 20% of its insureds categorized as high risk. A random sample of 35 insureds was taken. You are given:

- The sample revealed five high risk insureds.
- You want to test the hypothesis:
 - $H_0 : q = 0.2$
 - $H_1 : q \neq 0.2$

Calculate the change in the p -value by doubling the sample size and the number of observed high risk insureds.

Solution

Start by calculating \hat{q} . Since there are 5 successes from a sample size of 35,

$$\hat{q} = \frac{5}{35} = \frac{1}{7}$$

Next, calculate the test statistic with Equation 2.3.4.1.

$$\frac{\hat{q} - h}{\sqrt{\frac{h(1-h)}{n}}} = \frac{\frac{1}{7} - 0.2}{\sqrt{\frac{0.2(1-0.2)}{35}}} = -0.85$$

As this is a two-tailed test, we obtain a p -value of

$$\begin{aligned}\Pr(|Z| \geq |-0.85|) &= 2 \cdot \Pr(Z \geq |-0.85|) \\ &= 2 \cdot \Pr(Z \geq 0.85) \\ &= 2[1 - \Phi(0.85)] \\ &= 2(1 - 0.8023) \\ &= 0.3954\end{aligned}$$

By doubling the sample size and the number of observed high risk insureds, the value of \hat{q} remains unchanged, i.e. $\frac{2(5)}{2(35)} = \frac{1}{7}$. However, the test statistic becomes

$$\frac{\hat{q} - h}{\sqrt{\frac{h(1-h)}{n}}} = \frac{\frac{1}{7} - 0.2}{\sqrt{\frac{0.2(1-0.2)}{70}}} = -1.20$$

Following similar steps from before, the new p -value is

$$\begin{aligned}\Pr(|Z| \geq |-1.20|) &= 2[1 - \Phi(1.20)] \\ &= 2(1 - 0.8849) \\ &= 0.2302\end{aligned}$$

Therefore, the change in the p -value is $0.2302 - 0.3954 = -0.1652$. Importantly, this illustrates that if all else remains equal, a larger sample size leads to a smaller p -value, and hence, a greater tendency to reject H_0 . This should make intuitive sense. Having more evidence should result in a test that is more sensitive to discrepancies between the data and the null hypothesis.



Coach's Remarks

We could also calculate the probabilities using the Excel formula:

`NORM.S.DIST()`

- $\Phi(0.85) = \text{NORM.S.DIST}(0.85, \text{ TRUE}) = 0.8023$
- $\Phi(1.20) = \text{NORM.S.DIST}(1.20, \text{ TRUE}) = 0.8849$

2.3.5 Tests for Variances

We may also conduct hypothesis tests for variances. Unfortunately, we cannot use most of the specifics that relate to testing means or proportions.

One Sample

For one random sample of size n drawn from a normal distribution with variance σ^2 ,

$$\frac{(n-1)S^2}{\sigma^2}$$

has a chi-square distribution with $n - 1$ degrees of freedom. As a result, a hypothesis test for σ^2 has the test statistic

$$t.s. = \frac{(n-1)s^2}{h} \quad (2.3.5.1)$$

which comes from a chi-square distribution with $n - 1$ degrees of freedom. It should be no surprise that h is the hypothesized σ^2 as per H_0 . In addition, recall that $(n - 1)s^2$ is another way to express the sum of squared deviations from \bar{x} .

A chi-square random variable is constructed as follows. Let Z_1, Z_2, \dots, Z_ν be independent standard normal random variables. Then, the random variable

$$\sum_{i=1}^{\nu} Z_i^2$$

follows a **chi-square distribution** with ν degrees of freedom as its only parameter. This means $\frac{(n-1)S^2}{\sigma^2}$ has the same distribution as the sum of $n - 1$ independent squared standard normal random variables. In contrast to the standard normal distribution and t -distribution, the chi-square distribution is only valid on non-negative numbers. Furthermore, it is a right-skewed distribution.

We use $\chi_{q, df}^2$ to denote the $100q^{\text{th}}$ percentile of a chi-square distribution with df degrees of freedom, which can be calculated using the Excel formula:

CHISQ.INV(q, df)

Coach's Remarks

The final page of the exam table has the chi-square distribution table, which is shown below:

		Lower-tail areas for Chi-square distributions							
		0.005	0.010	0.025	0.050	0.950	0.975	0.990	0.995
df									
1		0.00	0.00	0.00	0.00	3.84	5.02	6.63	7.88
2		0.01	0.02	0.05	0.10	5.99	7.38	9.21	10.60
3		0.07	0.11	0.22	0.35	7.81	9.35	11.34	12.84
4		0.21	0.30	0.48	0.71	9.49	11.14	13.28	14.86
5		0.41	0.55	0.83	1.15	11.07	12.83	15.09	16.75
6		0.68	0.87	1.24	1.64	12.59	14.45	16.81	18.55
7		0.99	1.24	1.69	2.17	14.07	16.01	18.48	20.28
8		1.34	1.65	2.18	2.73	15.51	17.53	20.09	21.95
9		1.73	2.09	2.70	3.33	16.92	19.02	21.67	23.59
10		2.16	2.56	3.25	3.94	18.31	20.48	23.21	25.19
11		2.60	3.05	3.82	4.57	19.68	21.92	24.72	26.76
12		3.07	3.57	4.40	5.23	21.03	23.34	26.22	28.30
13		3.57	4.11	5.01	5.89	22.36	24.74	27.69	29.82
14		4.07	4.66	5.63	6.57	23.68	26.12	29.14	31.32
15		4.60	5.23	6.26	7.26	25.00	27.49	30.58	32.80
16		5.14	5.81	6.91	7.96	26.30	28.85	32.00	34.27
17		5.70	6.41	7.56	8.67	27.59	30.19	33.41	35.72
18		6.26	7.01	8.23	9.39	28.87	31.53	34.81	37.16
19		6.84	7.63	8.91	10.12	30.14	32.85	36.19	38.58
20		7.43	8.26	9.59	10.85	31.41	34.17	37.57	40.00
21		8.03	8.90	10.28	11.59	32.67	35.48	38.93	41.40
22		8.64	9.54	10.98	12.34	33.92	36.78	40.29	42.80
23		9.26	10.20	11.69	13.09	35.17	38.08	41.64	44.18
24		9.89	10.86	12.40	13.85	36.42	39.36	42.98	45.56
25		10.52	11.52	13.12	14.61	37.65	40.65	44.31	46.93
26		11.16	12.20	13.84	15.38	38.89	41.92	45.64	48.29
27		11.81	12.88	14.57	16.15	40.11	43.19	46.96	49.64
28		12.46	13.56	15.31	16.93	41.34	44.46	48.28	50.99
29		13.12	14.26	16.05	17.71	42.56	45.72	49.59	52.34
30		13.79	14.95	16.79	18.49	43.77	46.98	50.89	53.67
31		14.46	15.66	17.54	19.28	44.99	48.23	52.19	55.00
32		15.13	16.36	18.29	20.07	46.19	49.48	53.49	56.33
33		15.82	17.07	19.05	20.87	47.40	50.73	54.78	57.65
34		16.50	17.79	19.81	21.66	48.60	51.97	56.06	58.96
35		17.19	18.51	20.57	22.47	49.80	53.20	57.34	60.27
36		17.89	19.23	21.34	23.27	51.00	54.44	58.62	61.58
37		18.59	19.96	22.11	24.07	52.19	55.67	59.89	62.88
38		19.29	20.69	22.88	24.88	53.38	56.90	61.16	64.18
39		20.00	21.43	23.65	25.70	54.57	58.12	62.43	65.48
40		20.71	22.16	24.43	26.51	55.76	59.34	63.69	66.77
41		21.42	22.91	25.21	27.22	56.94	60.56	64.85	68.05

For example, the table indicates that 31.41 is the 95th percentile of a chi-square distribution with 20 degrees of freedom, i.e. $\chi_{0.95, 20}^2 = 31.41$. This matches the result from the Excel formula:

$$\text{CHISQ.INV}(0.95, 20) = 31.41$$

Notice that, **unlike** the *t*-distribution table, the probabilities listed in this table are cumulative probabilities.

Depending on the type of test, the critical regions are:

- For left-tailed tests, reject H_0 when

$$t.s. \leq \chi_{\alpha, n-1}^2$$

- For two-tailed tests, reject H_0 when

$$\left[t.s. \leq \chi_{\frac{\alpha}{2}, n-1}^2 \right] \cup \left[t.s. \geq \chi_{1-\frac{\alpha}{2}, n-1}^2 \right]$$

- For right-tailed tests, reject H_0 when

$$t.s. \geq \chi_{1-\alpha, n-1}^2$$

Since the chi-square distribution is **not** symmetrical, these inequalities do not simplify further.

Example 2.3.5.1

SAT test scores are reported to follow a normal distribution with variance 125. A researcher suspects that the assumed variance is different, and takes a random sample of 20 test scores. The sample variance of the 20 scores is 62.

Test whether the researcher is justified when the probability of a Type I error is 5%.

Solution

For σ^2 representing the variance of SAT test scores, the hypotheses are

$$H_0 : \sigma^2 = 125 \quad H_1 : \sigma^2 \neq 125$$

This is a two-tailed test. Recall that the probability of a Type I error equals the significance level. For $\alpha = 0.05$, the critical values are

$$\chi_{\frac{\alpha}{2}, n-1}^2 = \chi_{0.025, 19}^2 = 8.91$$

$$\chi_{1-\frac{\alpha}{2}, n-1}^2 = \chi_{0.975, 19}^2 = 32.85$$

The test statistic is

$$\frac{(n-1)s^2}{\sigma^2} = \frac{19 \cdot 62}{125} = 9.424$$

Since $8.91 < 9.424 < 32.85$, we **fail to reject H_0 at the 5% significance level**, so the researcher is not justified.



Two Samples

For two independent random samples of respective sizes n_1 and n_2 , each drawn from a normal distribution, having respective variances σ_1^2 and σ_2^2 , the expression

$$\frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2}$$

follows an F -distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom.

An F random variable is constructed as follows. Let:

- $C_1 \sim \text{Chi-Square} (\nu_1)$
- $C_2 \sim \text{Chi-Square} (\nu_2)$
- C_1 and C_2 be independent

Then, the random variable

$$\frac{C_1/\nu_1}{C_2/\nu_2}$$

follows an **F -distribution** with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom for its two parameters. As chi-square distributions are valid on non-negative numbers, this guarantees the same for F -distributions. We encourage you to prove on your own how $(S_1^2/S_2^2) (\sigma_2^2/\sigma_1^2)$ is F distributed using this definition. Moreover, we can infer from the definition that

$$Y \sim F(\nu_1, \nu_2) \Rightarrow Y^{-1} \sim F(\nu_2, \nu_1)$$

The $100q^{\text{th}}$ percentile of an F -distribution with ndf numerator degrees of freedom and ddf denominator degrees of freedom can be calculated using the Excel formula:

`F.INV(q, ndf, ddf)`

Coach's Remarks

The exam table dedicates three pages to the F -distribution table. Here is a portion of the table:

Selected Upper-tail areas for F-distributions

		Numerator df	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20
Denominator df	Upper-tail																
1	0.20		9.472	12	13.064	13.644	14.008	14.258	14.439	14.577	14.685	14.772	14.844	14.904	14.998	15.07	15.171
1	0.10		39.86	49.5	53.59	55.83	57.24	58.2	58.91	59.44	59.86	60.19	60.47	60.71	61.07	61.35	61.74
1	0.05		161.4	199.5	215.7	224.6	230.2	234	236.8	238.9	240.5	241.9	243	243.9	245.4	246.5	248
1	0.02		1013	1249	1351	1406	1441	1464	1482	1495	1505	1514	1521	1526	1535	1542	1552
1	0.01		4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6083	6106	6143	6170	6209
2	0.20		3.556	4	4.156	4.236	4.284	4.317	4.34	4.358	4.371	4.382	4.391	4.399	4.41	4.419	4.432
2	0.10		8.526	9	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392	9.401	9.408	9.42	9.429	9.441
2	0.05		18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4	19.4	19.41	19.42	19.43	19.45
2	0.02		48.51	49	49.17	49.25	49.3	49.33	49.36	49.37	49.39	49.4	49.41	49.42	49.43	49.44	49.45
2	0.01		98.5	99	99.17	99.25	99.3	99.33	99.36	99.37	99.39	99.4	99.41	99.42	99.43	99.44	99.45
3	0.20		2.682	2.886	2.936	2.956	2.965	2.971	2.974	2.976	2.978	2.979	2.98	2.981	2.982	2.982	2.983
3	0.10		5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.24	5.23	5.222	5.216	5.205	5.196	5.184
3	0.05		10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.763	8.745	8.715	8.692	8.66
3	0.02		20.62	18.86	18.11	17.69	17.43	17.25	17.11	17.01	16.93	16.86	16.81	16.76	16.69	16.63	16.55
3	0.01		34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.13	27.05	26.92	26.83	26.69
4	0.20		2.351	2.472	2.485	2.483	2.478	2.473	2.469	2.465	2.462	2.46	2.457	2.455	2.452	2.449	2.445
4	0.10		4.545	4.325	4.191	4.107	4.051	4.01	3.979	3.955	3.936	3.92	3.907	3.896	3.878	3.864	3.844
4	0.05		7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.936	5.912	5.873	5.844	5.803
4	0.02		14.04	12.142	11.344	10.899	10.616	10.419	10.274	10.162	10.074	10.003	9.944	9.894	9.815	9.755	9.67
4	0.01		21.2	18	16.69	15.98	15.52	15.21	14.98	14.8	14.66	14.55	14.45	14.37	14.25	14.15	14.02

Notice that each subtable corresponds to a single denominator degrees of freedom. This means the best way to read the table is to first identify the denominator degrees of freedom.

We use $F_{1-q, \text{ndf}, \text{ddf}}$ to denote the $100q^{\text{th}}$ percentile of an F distribution with ndf numerator degrees of freedom and ddf denominator degrees of freedom. Therefore, the table indicates that 19.37 is the 95th percentile of an F -distribution with $\text{ndf} = 8$ and $\text{ddf} = 2$, i.e.

$F_{0.05, 8, 2} = 19.37$. This matches the result from the Excel formula:

$$\text{F.INV}(0.95, 8, 2) = 19.37$$

Unlike other distribution tables on the exam sheet, this is the only one that expresses probabilities in terms of **survival probabilities**. For that reason, it aligns more directly with the Excel formula:

$$\text{F.INV.RT}(0.05, 8, 2) = 19.37$$

Here, RT stands for "right tail", which means the formula, like the table, is focused on the upper tail of the distribution.

Based on the given probabilities, the table directly supplies the 80th, 90th, 95th, 98th, and 99th percentiles. However, it is also possible to obtain the 1st, 2nd, 5th, 10th, and 20th percentiles.

For $Y \sim F(\nu_1, \nu_2)$,

$$\Pr(Y > F_{1-q, \nu_1, \nu_2}) = 1 - q$$

$$\Rightarrow \Pr\left[Y^{-1} < (F_{1-q, \nu_1, \nu_2})^{-1}\right] = 1 - q$$

Since we know that $Y^{-1} \sim F(\nu_2, \nu_1)$, this illustrates that the $100(1 - q)^{\text{th}}$ percentile of Y^{-1} equals the reciprocal of the $100q^{\text{th}}$ percentile of Y . Keep in mind that the values of ndf and ddf are exchanged between Y and Y^{-1} . In general,

$$F_{q, \nu_2, \nu_1} = \frac{1}{F_{1-q, \nu_1, \nu_2}} \quad (2.3.5.2)$$

Determine the 2nd percentile of an F -distribution with 2 numerator degrees of freedom and 4 denominator degrees of freedom.

We want to take the reciprocal of the 98th percentile of an F -distribution with ndf = 4 and ddf = 2, i.e. compute $(F_{0.02, 4, 2})^{-1}$. The table states that $F_{0.02, 4, 2} = 49.25$. Thus, the answer is

$$F_{0.98, 2, 4} = \frac{1}{49.25} = \mathbf{0.020}$$

This, again, matches the result from the Excel formula:

$$\text{F.INV}(0.02, 4, 2) = 0.020$$

Remember that we denote F percentiles with the survival probability written in the subscript, e.g. the 2nd percentile produces a survival probability of 0.98. This is done intentionally to match how the probabilities are stated in the table.

To test the variances, we investigate their ratio rather than their difference. In other words,

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = h$$

By letting $h = 1$, we can examine whether the variances are the same. In any case, we use the test statistic

$$t.s. = \frac{s_1^2}{s_2^2} \cdot \frac{1}{h} \quad (2.3.5.3)$$

which comes from an F -distribution with $\text{ndf} = n_1 - 1$ and $\text{ddf} = n_2 - 1$.

RIGHT-TAILED TESTS

In this case, the alternative hypothesis is

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > h$$

Therefore, we reject H_0 when

$$t.s. \geq F_{\alpha, n_1-1, n_2-1}$$

LEFT-TAILED TESTS

In this case, the alternative hypothesis is

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} < h$$

Therefore, we reject H_0 when

$$t.s. \leq F_{1-\alpha, n_1-1, n_2-1}$$

There are two possible approaches. The first is to solve for the critical value using Equation 2.3.5.2.

$$F_{1-\alpha, n_1-1, n_2-1} = \frac{1}{F_{\alpha, n_2-1, n_1-1}}$$

The second is to restate the critical region by taking the reciprocal on both sides of the inequality. Remember to flip the inequality sign as well.

$$\frac{1}{t.s.} \geq \frac{1}{F_{1-\alpha, n_1-1, n_2-1}} = F_{\alpha, n_2-1, n_1-1}$$

Either way, we need the percentile F_{α, n_2-1, n_1-1} from the F -distribution table. But what makes the second approach attractive is in realizing how we can rewrite H_1 as

$$\frac{\sigma_1^2}{\sigma_2^2} < h \quad \Rightarrow \quad \frac{\sigma_2^2}{\sigma_1^2} > \frac{1}{h} = h_*$$

and that

$$\frac{1}{t.s.} = \frac{1}{\frac{s_1^2}{s_2^2} \cdot \frac{1}{h}} = \frac{s_2^2}{s_1^2} \cdot \frac{1}{h_*}$$

In other words, this left-tailed test is equivalent to a right-tailed test where Sample 2 acts as Sample 1, and vice versa. In summary, we may perform a left-tailed test by following these steps:

1. Write the null hypothesis as

$$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = h_*$$

2. Perform a right-tailed test based on this H_0 .

TWO-TAILED TESTS

In this case, the alternative hypothesis is

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq h$$

Therefore, we reject H_0 when

$$\left[t. s. \leq \left(F_{\frac{\alpha}{2}, n_2-1, n_1-1} \right)^{-1} \right] \cup \left[t. s. \geq F_{\frac{\alpha}{2}, n_1-1, n_2-1} \right]$$

The left-tail critical value is in fact $F_{1-\frac{\alpha}{2}, n_1-1, n_2-1}$; we have applied Equation 2.3.5.2 so that it can be solved using the table.

Example 2.3.5.2

Two independent random samples of sizes 6 and 4 are taken from two separate normal distributions with variances σ_1^2 and σ_2^2 , respectively.

- For the first sample, $\sum_{i=1}^6 (x_i - \bar{x})^2 = 274$, where \bar{x} is its sample mean.
- For the second sample, $\sum_{i=1}^4 (x_i - \bar{x})^2 = 307$, where \bar{x} is its sample mean.
- You want to test the hypothesis:
 - $H_0 : \sigma_1^2 = 3\sigma_2^2$
 - $H_1 : \sigma_1^2 < 3\sigma_2^2$

Determine the inequality that describes the p -value of this test based on the given F -distribution table.

Solution

Start by writing the hypotheses as

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 3 \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 3$$

Then, compute the unbiased sample variances of the two samples.

$$s_1^2 = \frac{274}{6-1} = 54.8, \quad s_2^2 = \frac{307}{4-1} = 102.333$$

Therefore, the test statistic is

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{h} = \frac{54.8}{102.333} \cdot \frac{1}{3} = 0.1785$$

Let Y denote a random variable that follows an F -distribution with $\text{ndf} = 5$ and $\text{ddf} = 3$. Hence, the p -value for this left-tailed test is

$$\Pr(Y \leq t.s.) = \Pr(Y \leq 0.1785)$$

Remember that the five probabilities listed in the F -distribution table refer to survival probabilities. This means it is easier to calculate the p -value as a survival probability by taking reciprocals and changing the direction of the inequality.

$$\Pr(Y \leq 0.1785) = \Pr(Y^{-1} \geq 0.1785^{-1}) = \Pr(Y^{-1} \geq 5.602)$$

Note that Y^{-1} is F distributed with $\text{ndf} = 3$ and $\text{ddf} = 5$. From the F -distribution table, we see that

$$\Pr(Y^{-1} \geq 5.409) = 0.05, \quad \Pr(Y^{-1} \geq 8.670) = 0.02$$

In conclusion,

$$\Pr(Y^{-1} \geq 8.670) < \Pr(Y^{-1} \geq 5.602) < \Pr(Y^{-1} \geq 5.409)$$

$$\Rightarrow 0.02 < p\text{-value} < 0.05$$



Alternative Solution

Since this is a left-tailed test, we can approach it as a right-tailed test if we let Sample 2 take the role of Sample 1, and vice versa. Begin by writing the hypotheses as

$$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = \frac{1}{3} \quad H_1 : \frac{\sigma_2^2}{\sigma_1^2} > \frac{1}{3}$$

After solving for $s_2^2 = 102.333$ and $s_1^2 = 54.8$, calculate the alternate test statistic as

$$\frac{102.333}{54.8} \cdot \frac{1}{1/3} = 5.602$$

It is no coincidence that this test statistic is the reciprocal of the test statistic in the previous solution.

Let Y denote a random variable that follows an F -distribution with $\text{ndf} = 3$ and $\text{ddf} = 5$. Realize that Y in this solution is not the same as Y in the previous solution; they are reciprocals of each other. The p -value of interest is

$$\Pr(Y \geq 5.602)$$

as it is now based on a right-tailed test. At this point, the comparison of probabilities is exactly the same as in the previous solution. Thus,

$$\Pr(Y \geq 8.670) < \Pr(Y \geq 5.602) < \Pr(Y \geq 5.409)$$

$$\Rightarrow 0.02 < p\text{-value} < 0.05$$



2.3.6 Strategies and Takeaways

🕒 5m

It is challenging to learn and remember all of the hypothesis test details covered so far. Many other topics will depend on and/or tie in with the core concepts from this subsection, so it is vital to know them well now.

Here are some suggestions to help you better familiarize yourself with these concepts, including problem-solving strategies that are specific to this subsection:

1. Make sure you **understand** the core ideas behind hypothesis testing, i.e. Sections 2.3.1 and 2.3.2. In particular, there should be no confusion between critical value versus test statistic, as well as between α versus p -value. Similarly, know what makes a test two-tailed versus one-tailed, as well as left-tailed versus right-tailed. Blind memorization will likely make things far more difficult.
2. There should be no need to memorize any of the critical region inequalities. They should flow intuitively from your comprehension of what is a critical region. In light of that, sketching a distribution's density should help tremendously in determining the critical value(s).

This also means you should not blindly memorize our notation for percentiles. Take time to understand what they represent, and why we have written them in that way. This will help you be less dependent on the numerous symbols which can be very hard to memorize flawlessly.

3. For the four distribution tables provided on the exam table, remember that the probabilities listed in each table are:

Distribution Table	Probability Type
Standard normal	Cumulative
t	Both tails
Chi-square	Cumulative
F	Survival

4. Problems from this subsection can be categorized into two groups: those that mention a critical region, and those that do not.

For the first group, you will need to recognize the applicable distribution on your own based on the test statistic. See Examples 2.3.1.4 and 2.3.1.5.

For the second group, it will fall under the tests discussed in Sections 2.3.3 to 2.3.5.

5. When memorizing the test statistic formulas in Sections 2.3.3 and 2.3.4, pay attention to their similar form and appreciate how they are connected.

2.3 Summary

Notation

Symbol	Concept
H_0	Null hypothesis
H_1	Alternative hypothesis
$t. s.$	Test statistic
α	Significance level, Size, Probability of Type I error
z_q	100 q^{th} percentile of the standard normal distribution
df	Degrees of freedom
$t_{2(1-q), \text{df}}$	100 q^{th} percentile of a t -distribution
$\chi^2_{q, \text{df}}$	100 q^{th} percentile of a chi-square distribution
ndf	Numerator degrees of freedom
ddf	Denominator degrees of freedom
$F_{1-q, \text{ndf}, \text{ddf}}$	100 q^{th} percentile of an F -distribution

Terminology

- Test statistic: A value calculated from data that assumes H_0 is true.
- Critical region: The range of $t. s.$ values where we reject H_0 .
- Critical value: A value that borders the critical region, separating it from the rest of the possible $t. s.$ values.
- Two-tailed test: A test that includes both tails of a distribution in its critical region.
- Right-tailed test: A test that only includes the right tail of a distribution in its critical region.
- Left-tailed test: A test that only includes the left tail of a distribution in its critical region.
- Significance level: The probability of rejecting H_0 , assuming it is true.
- p -value: The probability of observing $t. s.$ or a more extreme value, assuming H_0 is true.
- Type I error: The incorrect decision of rejecting H_0 when it is true.
- Type II error: The incorrect decision of not rejecting H_0 when it is false.
- Power: The probability of rejecting H_0 , assuming it is false.

Key Ideas

- The meaning of "more extreme" depends on whether the test is two-tailed, right-tailed, or left-tailed.
- Reject H_0 when $t. s.$ is more extreme than the critical value(s), or equivalently, when $p\text{-value} \leq \alpha$.
- Critical values are percentiles calculated based on α .

Tests for Means

ONE SAMPLE

- $H_0 : \mu = h$
- If σ^2 is known and the Central Limit Theorem applies, then

$$t. s. = \frac{\bar{x} - h}{\sigma / \sqrt{n}}$$

comes from the standard normal distribution. The critical regions are:

Test Type	Critical Region
Left-tailed	$t. s. \leq -z_{1-\alpha}$
Two-tailed	$ t. s. \geq z_{1-\frac{\alpha}{2}}$
Right-tailed	$t. s. \geq z_{1-\alpha}$

- If σ^2 is unknown and the random sample is drawn from a normal distribution, then

$$t. s. = \frac{\bar{x} - h}{s / \sqrt{n}}$$

comes from a t -distribution with $n - 1$ degrees of freedom. The critical regions are:

Test Type	Critical Region
Left-tailed	$t. s. \leq -t_{2\alpha, n-1}$

Test Type	Critical Region
Two-tailed	$ t.s. \geq t_{\alpha, n-1}$
Right-tailed	$t.s. \geq t_{2\alpha, n-1}$

TWO SAMPLES

- $H_0 : \mu_1 - \mu_2 = h$
- Both samples are independent.
- If σ_1^2 and σ_2^2 are known and the Central Limit Theorem applies, then

$$t.s. = \frac{\bar{x}_1 - \bar{x}_2 - h}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

comes from the standard normal distribution. The critical regions are the same as the one-sample case.

- If σ_1^2 and σ_2^2 are unknown, each random sample is drawn from a normal distribution, and $\sigma_1^2 = \sigma_2^2$, then

$$t.s. = \frac{\bar{x}_1 - \bar{x}_2 - h}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

comes from a t -distribution with $n_1 + n_2 - 2$ degrees of freedom. The critical regions are the same as the one-sample case, with updated degrees of freedom.

- The pooled sample variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

PAIRED OBSERVATIONS

- $H_0 : \mu_1 - \mu_2 = h$
- Both samples are not independent; the observations form pairs between the samples.

- This is identical to the one-sample case with the following substitutions:

- $\bar{x} \rightarrow \bar{d}$
- $\sigma^2 \rightarrow \sigma_D^2$
- $n \rightarrow n_*$
- $s^2 \rightarrow s_D^2$

Tests for Proportions

ONE SAMPLE

- $H_0 : q = h$
- $\hat{q} = \frac{\# \text{ of successes from } n \text{ trials}}{n}$
- If the Central Limit Theorem applies, then

$$t.s. = \frac{\hat{q} - h}{\sqrt{\frac{h(1-h)}{n}}}$$

comes from the standard normal distribution. The critical regions are the same as the one-sample case for testing means with known variance.

TWO SAMPLES

- $H_0 : q_1 - q_2 = h$
- Both samples are independent.
- If the Central Limit Theorem applies, then

$$t.s. = \frac{\hat{q}_1 - \hat{q}_2 - h}{\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}}$$

comes from the standard normal distribution. The critical regions are the same as the one-sample case for testing means with known variance.

Tests for Variances

ONE SAMPLE

- $H_0 : \sigma^2 = h$
- If the random sample is drawn from a normal distribution, then

$$t.s. = \frac{(n-1)s^2}{h}$$

comes from a chi-square distribution with $n - 1$ degrees of freedom. The critical regions are:

Test Type	Critical Region
Left-tailed	$t.s. \leq \chi_{\alpha, n-1}^2$
Two-tailed	$\left[t.s. \leq \chi_{\frac{\alpha}{2}, n-1}^2 \right] \cup \left[t.s. \geq \chi_{1-\frac{\alpha}{2}, n-1}^2 \right]$
Right-tailed	$t.s. \geq \chi_{1-\alpha, n-1}^2$

- Chi-square distributions are not symmetric, so the two-tailed test requires checking both critical values.

TWO SAMPLES

- $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = h$
- Both samples are independent.
- If each random sample is drawn from a normal distribution, then

$$t.s. = \frac{s_1^2}{s_2^2} \cdot \frac{1}{h}$$

comes from an F -distribution with $\text{ndf} = n_1 - 1$ and $\text{ddf} = n_2 - 1$. The critical regions are:

Test Type	Critical Region
Left-tailed	$t.s. \leq F_{1-\alpha, n_1-1, n_2-1}$
Two-tailed	$[t.s. \leq (F_{\frac{\alpha}{2}, n_2-1, n_1-1})^{-1}] \cup [t.s. \geq F_{\frac{\alpha}{2}, n_1-1, n_2-1}]$
Right-tailed	$t.s. \geq F_{\alpha, n_1-1, n_2-1}$

- F distributions are not symmetric, so the two-tailed test requires checking both critical values.
- A left-tailed test can be performed by writing H_0 in terms of $\frac{\sigma_2^2}{\sigma_1^2}$ instead and doing a right-tailed test.
- $F_{q, \nu_2, \nu_1} = (F_{1-q, \nu_1, \nu_2})^{-1}$

Appendix

🕒 5m

Variance of Difference for Paired Observations

Let $X_{i,k}$ denote the i^{th} observation from Sample k , for $i = 1, \dots, n_*$, and $k = 1, 2$. This means

$$D_i = X_{i,1} - X_{i,2}$$

In addition, let

- σ_k^2 be the variance for Sample k
- ρ be the correlation of $X_{i,1}$ and $X_{i,2}$

Then,

$$\begin{aligned}\sigma_D^2 &= \text{Var}[D_i] \\ &= \text{Var}[X_{i,1} - X_{i,2}] \\ &= \text{Var}[X_{i,1}] + (-1)^2 \text{Var}[X_{i,2}] + 2(-1)\text{Cov}[X_{i,1}, X_{i,2}] \\ &= \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\end{aligned}$$

If the two samples are independent, then $\rho = 0$, and in turn

$$\sigma_D^2 = \sigma_1^2 + \sigma_2^2$$

But if the two samples are dependent due to the observations being paired, then it may be reasonable that $\rho > 0$ (i.e. $X_{i,1}$ and $X_{i,2}$ tend to increase or decrease together), resulting in

$$\sigma_D^2 < \sigma_1^2 + \sigma_2^2$$

Therefore, the variance of D_i could be smaller when the observations are paired. This is beneficial in that \bar{D} would be an estimator with a smaller variance.

2.4.0 Overview

 5m

Previously, we learned of hypothesis tests and their associated critical regions. For example, a two-tailed test such as

$$H_0 : \mu = h \quad H_1 : \mu \neq h$$

involves a critical region that covers both tails of a distribution. Although this makes sense intuitively, is there a concrete justification for this critical region made of symmetric tails? Can we show that, all else being equal, a different critical region will result in a worse test performance?

In general, these questions are very challenging to answer, particularly with the immense variety of possible tests. It also may not be obvious how one should define or measure test performance.

This subsection addresses similar questions for only a handful of cases. Specifically, we will discuss the Neyman-Pearson theorem and the concept of uniformly most powerful (UMP) tests.

2.4.1 Neyman-Pearson Theorem

Simple and Composite Hypotheses

A **simple hypothesis** is a hypothesis that fully specifies the distribution(s) associated with the sample.

A **composite hypothesis** is a hypothesis that does not fully specify the distribution(s) associated with the sample; it is composed of many simple hypotheses.

To clarify, consider these examples.

A random sample is drawn from a normal distribution with mean μ and known variance σ^2 . A hypothesis test has $H_0 : \mu = 5$.

Is H_0 a simple or composite hypothesis?

Without H_0 , the only unspecified component of the normal distribution is the value of its mean. Since H_0 specifies a mean of 5, it fully specifies the normal distribution. Thus, **H_0 is a simple hypothesis**.

A random sample is drawn from a normal distribution with mean μ and known variance σ^2 . A hypothesis test has $H_0 : \mu \leq 5$.

Is H_0 a simple or composite hypothesis?

Here, H_0 does not specify an exact value for the mean. Since it does not fully specify the normal distribution, **H_0 is a composite hypothesis**.

A random sample is drawn from a normal distribution with mean μ and **unknown** variance σ^2 . A hypothesis test has $H_0 : \mu = 5$.

Is H_0 a simple or composite hypothesis?

Even though H_0 specifies an exact value for the mean, the normal distribution is still not fully specified due to the unknown σ^2 . Thus, **H_0 is a composite hypothesis**.

Most Powerful Test

By limiting our scope such that

- both H_0 and H_1 are simple hypotheses, and
- tests have the same significance level α ,

we define a test as having the best critical region of size α if it has the largest power compared to all other tests. Such a test is called the **most powerful test** of size α .

A single observation, x , drawn from a normal distribution with mean μ and variance 1 is used for testing the hypothesis

$$H_0 : \mu = 2 \quad H_1 : \mu = 3$$

Two tests are considered:

- Test A has a critical region of the form $x \leq c$
- Test B has a critical region of the form $x \geq c$

Show that Test A is not the most powerful test of size 0.05.

We simply need to show that the Test A power is less than the Test B power. To compute the powers, we need the exact critical regions.

Solve for Test A's critical value as

$$\begin{aligned}
 \Pr(X \leq c \mid \mu = 2) &= 0.05 \\
 \Pr\left(Z \leq \frac{c-2}{1}\right) &= 0.05 \\
 c-2 &= \Phi^{-1}(0.05) \\
 c-2 &= -\Phi^{-1}(0.95) \\
 c &= 2 - 1.645 \\
 &= 0.355
 \end{aligned}$$

Thus, the power of Test A is

$$\begin{aligned}
 \Pr(X \leq 0.355 \mid \mu = 3) &= \Pr\left(Z \leq \frac{0.355-3}{1}\right) \\
 &= \Phi(-2.645) \\
 &= 1 - \Phi(2.645) \\
 &= 0.0040
 \end{aligned}$$

Solve for Test B's critical value as

$$\begin{aligned}
 \Pr(X \geq c \mid \mu = 2) &= 0.05 \\
 \Pr\left(Z \geq \frac{c-2}{1}\right) &= 0.05 \\
 1 - \Phi(c-2) &= 0.05 \\
 c-2 &= \Phi^{-1}(0.95) \\
 c &= 2 + 1.645 \\
 &= 3.645
 \end{aligned}$$

Thus, the power of Test B is

$$\begin{aligned}
 \Pr(X \geq 3.645 \mid \mu = 3) &= \Pr\left(Z \geq \frac{3.645-3}{1}\right) \\
 &= 1 - \Phi(0.645) \\
 &= 0.2578
 \end{aligned}$$

Since $0.0040 < 0.2578$, we have shown that **Test A is not the most powerful test of size 0.05**; Test B is far more powerful than Test A at $\alpha = 0.05$.

Notice that the critical region forms already hint at this result. If the mean is higher (i.e. H_1 is true), then x would tend to be larger than anticipated. So, it makes sense to reject H_0 if x is large rather than small, i.e. Test B makes more sense than Test A. In turn, Test B should have a much higher chance of correctly rejecting H_0 relative to Test A.

Neyman-Pearson Theorem

The **Neyman-Pearson theorem** or **lemma** provides a way to identify the test that is most powerful. Consider the following set of simple hypotheses involving a generic parameter θ and two constants h_0 and h_1 :

$$H_0 : \theta = h_0 \quad H_1 : \theta = h_1$$

Additionally, let:

- $L(\theta)$ be the likelihood function based on the sample
- k be a positive constant

A test of size α is most powerful (i.e. has the best critical region) if and only if the following are satisfied:

Condition	Criterion
When the sample is within the critical region	$\frac{L(h_0)}{L(h_1)} \leq k$
When the sample is outside the critical region	$\frac{L(h_0)}{L(h_1)} \geq k$

Note that $L(\theta)$ is in terms of the observed values x_1, \dots, x_n . Likewise, the ratio of likelihood functions is also in terms of the observed values. As a result, it is typical to begin with the inequality

$$\frac{L(h_0)}{L(h_1)} \leq k \tag{2.4.1.1}$$

and manipulate it until a meaningful statistic emerges, such as \bar{x} . This helps to determine the best critical region.

With the manipulated inequality specifying the critical region, we can also identify the test statistic and critical value. This implies that the value of k is likely unimportant, since manipulating the inequality will cause k and other constants to be absorbed into the critical value.

There is an intuitive perspective to why Equation 2.4.1.1 corresponds with the best critical region. If $L(h_1)$ is significantly greater than $L(h_0)$, meaning the likelihood is much higher if H_1 is true, then the data seems to favor H_1 more than H_0 . This leads to a smaller $L(h_0) \div L(h_1)$ ratio. When the ratio is lower than or equal to some threshold k , we decide that the data is significantly more likely under H_1 than H_0 , hence H_0 is rejected.

Example 2.4.1.1

A certain brand of LED lightbulbs have exponentially distributed lifetimes with PDF

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0$$

A random sample of 15 lightbulbs is used to test the hypothesis:

- $H_0 : \theta = 4$
- $H_1 : \theta = 3$

Determine the form of the critical region for the most powerful test.

Solution

First, determine the likelihood function.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{15} f(x_i) \\ &= \frac{1}{\theta^{15}} e^{-\left(\sum_{i=1}^{15} x_i\right)/\theta} \end{aligned}$$

Therefore,

$$\begin{aligned}
 \frac{L(4)}{L(3)} &= \frac{\frac{1}{4^{15}} e^{-\left(\sum_{i=1}^{15} x_i\right)/4}}{\frac{1}{3^{15}} e^{-\left(\sum_{i=1}^{15} x_i\right)/3}} \\
 &= \left(\frac{\frac{1}{4}}{\frac{1}{3}}\right)^{15} e^{-\frac{\sum_{i=1}^{15} x_i}{4} - \left(-\frac{\sum_{i=1}^{15} x_i}{3}\right)} \\
 &= \left(\frac{3}{4}\right)^{15} e^{\frac{-\left(3 \cdot \sum_{i=1}^{15} x_i\right) + \left(4 \cdot \sum_{i=1}^{15} x_i\right)}{12}} \\
 &= \left(\frac{3}{4}\right)^{15} e^{\left(\sum_{i=1}^{15} x_i\right)/12}
 \end{aligned}$$

Applying the theorem,

$$\begin{aligned}
 \left(\frac{3}{4}\right)^{15} e^{\left(\sum_{i=1}^{15} x_i\right)/12} &\leq k \\
 \Rightarrow e^{\left(\sum_{i=1}^{15} x_i\right)/12} &\leq k \left(\frac{4}{3}\right)^{15} \\
 \Rightarrow \frac{\sum_{i=1}^{15} x_i}{12} &\leq \ln \left[k \left(\frac{4}{3}\right)^{15} \right] \\
 \Rightarrow \sum_{i=1}^{15} x_i &\leq 12 \ln \left[k \left(\frac{4}{3}\right)^{15} \right]
 \end{aligned}$$

Manipulating the inequality resulted in a summation test statistic emerging on the left side of inequality. Thus, the right side of the inequality is the critical value in the unit of total lifetime. Since the critical value is determined by the distribution specified under H_0 , its current form is inconsequential. The best critical region has the form

$$\sum_{i=1}^{15} x_i \leq c$$

This makes sense; if the sum of the observations is sufficiently small, then a smaller exponential mean seems more reasonable than a larger one, thus favoring H_1 over H_0 .

Coach's Remarks

To compute a critical value c , we need a significance level. In addition, note that

$$\sum_{i=1}^{15} X_i \sim \text{Gamma}(15, \theta)$$

With how challenging it is to solve by hand, this is not suitable to be tested on the exam. Nevertheless, with a 5% significance level,

$$\Pr\left(\sum_{i=1}^{15} X_i \leq c \mid \theta = 4\right) = 0.05$$

results in $c = 36.985$. Consequently, the power is

$$\Pr\left(\sum_{i=1}^{15} X_i \leq 36.985 \mid \theta = 3\right) = 0.2585$$

Therefore, the Neyman-Pearson theorem guarantees that all other tests of size 0.05 will not exceed a power of 0.2585.

While the Neyman-Pearson theorem was introduced in a context with one generic parameter θ , it is not a requirement for the theorem. The key is that H_0 and H_1 are simple hypotheses; they can even be statements describing more than just parameters. To generalize, " $L(h_0)$ " represents the likelihood function given by H_0 , and " $L(h_1)$ " represents the likelihood function given by H_1 .

Example 2.4.1.2

You consider the following hypotheses about a random variable, X :

- H_0 : X has a Pareto distribution with parameters $\alpha = 2$ and $\theta = 2$.
- H_1 : X has an inverse Pareto distribution with parameters $\tau = 1$ and $\theta = 2$.

To evaluate this hypothesis, you use a single observation and desire the best critical region to have a power of 0.75.

Calculate the probability of a Type I error.

Solution

Let x denote the single observation for this test. Then, use the exam table to determine the likelihood function given by H_0 .

$$\frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} = \frac{2(2)^2}{(x+2)^3}$$

Similarly, the likelihood function given by H_1 is

$$\frac{\tau\theta x^{\tau-1}}{(x+\theta)^{\tau+1}} = \frac{2}{(x+2)^2}$$

The ratio of likelihood functions is

$$\frac{\frac{2(2)^2}{(x+2)^3}}{\frac{2}{(x+2)^2}} = \frac{4}{x+2}$$

Applying the theorem,

$$\frac{4}{x+2} \leq k \quad \Rightarrow \quad \frac{x+2}{4} \geq \frac{1}{k} \quad \Rightarrow \quad x \geq \frac{4}{k} - 2$$

As a result, the best critical region has the form $x \geq c$. Given a power of 0.75, determine the critical value c . Refer to the exam table for the appropriate inverse Pareto probability expression.

$$\begin{aligned}\Pr(X \geq c \mid H_1 \text{ is true}) &= 0.75 \\ 1 - \left(\frac{c}{c+2}\right)^1 &= 0.75 \\ \frac{c}{c+2} &= 0.25 \\ c &= 0.25c + 0.5 \\ c &= \frac{0.5}{1 - 0.25} \\ &= \frac{2}{3}\end{aligned}$$

With the best critical region of $x \geq \frac{2}{3}$, calculate the probability of a Type I error, i.e. the significance level. Use the exam table to obtain the appropriate Pareto probability expression.

$$\begin{aligned}\Pr\left(X \geq \frac{2}{3} \mid H_0 \text{ is true}\right) &= \left(\frac{2}{\frac{2}{3} + 2}\right)^2 \\ &= \mathbf{0.5625}\end{aligned}$$



So far, we have seen expressions for the ratio of likelihood functions that are rather simple. In other words, it was straightforward how Equation 2.4.1.1 led to $\sum_{i=1}^{15} x_i \leq c$ in Example 2.4.1.1 and to $x \geq c$ in Example 2.4.1.2 as the best critical regions. However, there are situations where manipulating the Neyman-Pearson inequality is not as straightforward. It is important to remember that the inequality gives an **ordering** of the likelihood ratio values, where smaller values correspond to rejecting H_0 . We demonstrate this in the next two examples.

Example 2.4.1.3

You are given the following information for a hypothesis test on parameter θ of a

discrete distribution:

- $H_0 : \theta = 1$
- $H_1 : \theta = 2$
- Based on the hypotheses, the probability mass functions are:

	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$
$p(x \theta = 1)$	0.01	0.05	0.50	0.43	0.01
$p(x \theta = 2)$	0.02	0.24	0.25	0.25	0.24

- A single observation is used to determine the test result.

Determine the best critical region for tests of size 0.06.

Solution

Now that the likelihood is expressed with constants instead of a variable x , it is impossible to manipulate the Neyman-Pearson inequality just like before. However, we still need the ratio of the likelihoods. Let Λ denote the ratio. When $x = 1$, then

$$\Lambda = \frac{p(1 | \theta = 1)}{p(1 | \theta = 2)} = 0.5$$

Because there are only five possibilities for x , there are only five possible values for Λ :

	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$
Λ	0.5	0.208	2	1.72	0.042

Since we want to reject H_0 for smaller values of Λ , it makes sense to sort Λ in ascending order.

	$x = 5$	$x = 2$	$x = 1$	$x = 4$	$x = 3$
Λ	0.042	0.208	0.5	1.72	2

Consequently, the best critical region will include the values of x starting from the left of this table. Since $p(5 | \theta = 1) + p(2 | \theta = 1)$ equals the desired significance level of 0.06,

the best critical region here is

$$x = 2, 5$$

The table of probability mass functions agrees instinctively with this result. For example, imagine if the observed value is 5. Then, we would be suspicious as to whether H_0 is true because it is a very unlikely occurrence (with a 1% chance). We would be more willing to consider H_1 as true instead; observing a 5 then is far more likely (with a 24% chance). This similarly holds for an observed value of 2.

However, note that $x = 1$ is not part of the best critical region, even though it is also an unlikely occurrence under H_0 (with a 1% chance). This is because observing a 1 under H_1 is not very likely either (with a 2% chance). In comparing H_0 against H_1 , the values of 2 and 5 are "more extreme" than a value of 1. All of this information is captured in the ratio of the likelihoods; keep in mind that Examples 2.4.1.1 and 2.4.1.2 operate on the same reasoning.



Alternative Solution

Considering the values of $p(x \mid \theta = 1)$, there are only two possible critical regions with size 0.06:

- $x = 1, 2$
- $x = 2, 5$

The best critical region has the highest power. The power for the first critical region is

$$\Pr(X = 1 \text{ or } 2 \mid \theta = 2) = 0.02 + 0.24 = 0.26$$

whereas the power for the second critical region is

$$\Pr(X = 2 \text{ or } 5 \mid \theta = 2) = 0.24 + 0.24 = 0.48$$

Therefore, the best critical region is the second one, i.e.

$$x = 2, 5$$



Coach's Remarks

With discrete distributions, there are unattainable values for α . For example, $\alpha = 0.03$ cannot be attained from $p(x | \theta = 1)$ in this problem because no combination of x values can produce that probability. We believe that exam problems will avoid this situation.

Example 2.4.1.4

You consider the following hypotheses about a random variable, X :

- $H_0 : X$ is uniformly distributed on $[2, 7]$.
- $H_1 : X$ is uniformly distributed on $[0, 5]$.

To evaluate this hypothesis, you use a single observation.

Calculate the power of the most powerful test of size 0.05.

Solution

Let x denote the single observation for this test, and Λ denote the ratio of likelihood functions.

Notice that the distributions specified in H_0 and H_1 are valid on different intervals. So, calculating Λ requires using the likelihood value that corresponds to the proper range.

Under H_0 , the uniform PDF (i.e. likelihood) is

$$f(x) = \begin{cases} \frac{1}{5}, & 2 \leq x \leq 7 \\ 0, & \text{otherwise} \end{cases}$$

Under H_1 , the uniform PDF (i.e. likelihood) is

$$f(x) = \begin{cases} \frac{1}{5}, & 0 \leq x \leq 5 \\ 0, & \text{otherwise} \end{cases}$$

Therefore,

$$\begin{aligned} \Lambda &= \begin{cases} \frac{0}{1/5}, & 0 \leq x < 2 \\ \frac{1/5}{1/5}, & 2 \leq x \leq 5 \\ \frac{1/5}{0}, & 5 < x \leq 7 \end{cases} \\ &= \begin{cases} 0, & 0 \leq x < 2 \\ 1, & 2 \leq x \leq 5 \\ \infty, & 5 < x \leq 7 \end{cases} \end{aligned}$$

While dividing a non-zero number by 0 is technically undefined, we can treat it as dividing by a very small positive number instead, thus producing a result that goes to ∞ .

Since Λ is not expressed with x , we will not manipulate the Neyman-Pearson inequality. Instead, observe that Λ is non-decreasing in x , so smaller x values are associated with smaller Λ values. Therefore, the best critical region has the form $x \leq c$. Rejecting H_0 for small values of x should make sense; for example, if x is anything less than 2, then it is impossible for H_0 to be true.

With a test size of 0.05, the critical value c is

$$\begin{aligned} \Pr(X \leq c \mid H_0 \text{ is true}) &= 0.05 \\ \frac{c-2}{5} &= 0.05 \\ c &= 0.05(5) + 2 \\ &= 2.25 \end{aligned}$$

With the best critical region of $x \leq 2.25$, the power is

$$\begin{aligned}\Pr(X \leq 2.25 \mid H_1 \text{ is true}) &= \frac{2.25 - 0}{5} \\ &= \mathbf{0.45}\end{aligned}$$



2.4.2 Uniformly Most Powerful Tests

Previously, we have seen many hypothesis tests where H_0 and H_1 were not both simple hypotheses. While this means the Neyman-Pearson theorem does not apply directly to those cases, it is still useful in identifying ***uniformly most powerful (UMP) tests***.

The usual scenario is when H_0 is simple and H_1 is composite. Here, a UMP test of size α has a critical region which is the best critical region for testing H_0 against each simple hypothesis in H_1 . Said differently, if the best critical regions for all possibilities described in H_1 are all **the same**, then that critical region corresponds to a UMP test.

A random sample of size n is drawn from a Poisson distribution with mean λ to investigate the hypotheses

$$H_0 : \lambda = 2 \quad H_1 : \lambda > 2$$

Determine whether a UMP test exists for this setup.

Note that H_0 is simple and H_1 is composite. Let h_1 be a constant that satisfies a λ value in H_1 , i.e. h_1 is some constant greater than 2. Next, treat $\lambda = h_1$ as an alternative hypothesis and find the best critical region. Start by finding the likelihood function, and then the ratio of the likelihoods.

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^n \frac{e^{-\lambda} \cdot \lambda^{x_i}}{x_i!} \\ &= \frac{e^{-\lambda n} \cdot \lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \end{aligned}$$

$$\begin{aligned} \frac{L(2)}{L(h_1)} &= \frac{e^{-2n} \cdot 2^{\sum_{i=1}^n x_i} \div \prod_{i=1}^n x_i!}{e^{-h_1 \cdot n} \cdot h_1^{\sum_{i=1}^n x_i} \div \prod_{i=1}^n x_i!} \\ &= e^{(h_1-2)n} \left(\frac{2}{h_1} \right)^{\sum_{i=1}^n x_i} \end{aligned}$$

Applying the Neyman-Pearson theorem,

$$\begin{aligned}
& e^{(h_1-2)n} \left(\frac{2}{h_1} \right)^{\sum_{i=1}^n x_i} \leq k \\
& \Rightarrow \left(\frac{2}{h_1} \right)^{\sum_{i=1}^n x_i} \leq \frac{k}{e^{(h_1-2)n}} \\
& \Rightarrow \left(\sum_{i=1}^n x_i \right) \ln \left(\frac{2}{h_1} \right) \leq \ln \left[\frac{k}{e^{(h_1-2)n}} \right] \\
& \Rightarrow \sum_{i=1}^n x_i \geq \ln \left[\frac{k}{e^{(h_1-2)n}} \right] \div \ln \left(\frac{2}{h_1} \right)
\end{aligned}$$

Note that $\ln \left(\frac{2}{h_1} \right)$ is negative because $h_1 > 2 \Rightarrow \frac{2}{h_1} < 1$, which leads to flipping the direction of the inequality in the last line. Therefore, the best critical region has the form $\sum_{i=1}^n x_i \geq c$. Keep in mind this result stemmed from treating $\lambda = h_1$ as an alternative hypothesis.

The key is to realize the following: for this setup, the result of $\sum_{i=1}^n x_i \geq c$ as the best critical region holds true for **any** h_1 larger than 2; the exact value of h_1 does not matter here. This means $\lambda = h_1$ can represent any simple hypothesis in $H_1 : \lambda > 2$, and each one leads to the same best critical region of $\sum_{i=1}^n x_i \geq c$. Therefore, a **UMP test does exist** with $\sum_{i=1}^n x_i \geq c$ as the critical region.

A random sample of size n is drawn from a Poisson distribution with mean λ to investigate the hypotheses

$$H_0 : \lambda = 2 \quad H_1 : \lambda \neq 2$$

Determine whether a UMP test exists for this setup.

Again, H_0 is simple and H_1 is composite. Let h_1 be a constant that satisfies a λ value in H_1 . Next, treat $\lambda = h_1$ as an alternative hypothesis and find the best critical region. The steps are the same as the previous example until the manipulation of the Neyman-Pearson inequality at

$$\left(\sum_{i=1}^n x_i \right) \ln \left(\frac{2}{h_1} \right) \leq \ln \left[\frac{k}{e^{(h_1-2)n}} \right]$$

At this stage, we must consider the impact of h_1 being either greater than or less than 2 since both scenarios are possible in H_1 .

If $h_1 > 2$, then the result is the same as the previous example. As a reminder, the best critical region would have the form $\sum_{i=1}^n x_i \geq c$.

If $h_1 < 2$, then $\ln\left(\frac{2}{h_1}\right)$ is positive. The direction of the inequality does **not** change, and the best critical region would have the form $\sum_{i=1}^n x_i \leq c$ instead.

Because there are simple hypotheses in $H_1 : \lambda \neq 2$ that produce different best critical regions, a **UMP test does not exist here**.

Example 2.4.2.1

A random sample of size 10 is drawn from a normal distribution with mean 0 and variance σ^2 to test the hypotheses

$$H_0 : \sigma^2 = 5 \quad H_1 : \sigma^2 < 5$$

Determine the critical region for the uniformly most powerful test of size 0.025.

Solution

Let h_1 be a constant that satisfies a σ^2 value in H_1 , i.e. h_1 is some constant between 0 and 5. The likelihood function and the ratio of the likelihoods are as follows:

$$\begin{aligned} L(\sigma^2) &= \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - 0)^2}{2\sigma^2}\right] \\ &= \frac{1}{(2\pi\sigma^2)^{10/2}} \exp\left[-\frac{\sum_{i=1}^{10} x_i^2}{2\sigma^2}\right] \end{aligned}$$

$$\begin{aligned}
\frac{L(5)}{L(h_1)} &= \frac{\frac{1}{(2\pi \cdot 5)^5} \exp \left[-\frac{\sum_{i=1}^{10} x_i^2}{2(5)} \right]}{\frac{1}{(2\pi \cdot h_1)^5} \exp \left[-\frac{\sum_{i=1}^{10} x_i^2}{2h_1} \right]} \\
&= \left(\frac{h_1}{5} \right)^5 \exp \left[\frac{\sum_{i=1}^{10} x_i^2}{2h_1} - \frac{\sum_{i=1}^{10} x_i^2}{2(5)} \right] \\
&= \left(\frac{h_1}{5} \right)^5 \exp \left[\frac{(5-h_1) \sum_{i=1}^{10} x_i^2}{10h_1} \right]
\end{aligned}$$

Applying the Neyman-Pearson theorem,

$$\begin{aligned}
\left(\frac{h_1}{5} \right)^5 \exp \left[\frac{(5-h_1) \sum_{i=1}^{10} x_i^2}{10h_1} \right] &\leq k \\
\Rightarrow \exp \left[\frac{(5-h_1) \sum_{i=1}^{10} x_i^2}{10h_1} \right] &\leq k \left(\frac{5}{h_1} \right)^5 \\
\Rightarrow \frac{(5-h_1) \sum_{i=1}^{10} x_i^2}{10h_1} &\leq \ln \left[k \left(\frac{5}{h_1} \right)^5 \right] \\
\Rightarrow \sum_{i=1}^{10} x_i^2 &\leq \frac{10h_1}{5-h_1} \cdot \ln \left[k \left(\frac{5}{h_1} \right)^5 \right]
\end{aligned}$$

Note that $5 - h_1$ is positive because $h_1 < 5$, so the Neyman-Pearson inequality stays in the less-than direction. Since we are told that the test is UMP, the critical region has the form $\sum_{i=1}^{10} x_i^2 \leq c$. We encourage you to spend some time convincing yourself that the test is indeed UMP.

Given $\alpha = 0.025$, we want to solve for c . Use the fact that

$$\sum_{i=1}^{10} Z_i^2 = \sum_{i=1}^{10} \left(\frac{X_i - 0}{\sigma} \right)^2 = \frac{\sum_{i=1}^{10} X_i^2}{\sigma^2}$$

follows a chi-square distribution with 10 degrees of freedom. Thus,

$$\Pr\left(\sum_{i=1}^{10} X_i^2 \leq c \mid \sigma^2 = 5\right) = 0.025$$

$$\Pr\left(\frac{\sum_{i=1}^{10} X_i^2}{5} \leq \frac{c}{5}\right) = 0.025$$

$$\Rightarrow \frac{c}{5} = 3.25$$

$$c = 16.25$$

Therefore, the critical region for the UMP test of size 0.025 is

$$\sum_{i=1}^{10} x_i^2 \leq 16.25$$



If H_0 is also composite, there are specific setups where we can define a UMP test. In testing a generic parameter θ , consider the hypotheses

$$H_0 : \theta \leq h \quad H_1 : \theta > h$$

Define constants a and b as possible values of θ such that $a < b$, and

$$\Lambda = \frac{L(a)}{L(b)}$$

As seen in previous examples, Λ is a function of the observed values that form a statistic. For this test to be UMP, $L(\theta)$ must have a **monotone likelihood ratio** in a statistic y . This means that Λ is a monotone function of y .

A monotone **decreasing** function of y cannot increase in y , so small values of the function are associated with large y values.

A monotone **increasing** function of y cannot decrease in y , so small values of the function are associated with small y values.

Recall that smaller Λ values should lead to rejecting H_0 . Hence, depending on the type of monotone likelihood ratio, this test is UMP for size α with the following critical region forms:

Monotone Likelihood Ratio Type	Critical Region	α
Decreasing	$y \geq c$	$\Pr(Y \geq c \theta = h)$
Increasing	$y \leq c$	$\Pr(Y \leq c \theta = h)$

Since Λ must be a monotone function of y , manipulating the inequality

$$\frac{L(h)}{L(h_1)} \leq k$$

will produce one of the critical region forms mentioned in the table, where h_1 is any constant greater than h .

A test with hypotheses

$$H_0 : \theta \geq h \quad H_1 : \theta < h$$

can also be UMP. It only differs from the previous setup in one area: a and b are possible values of θ such that $a > b$. This also means that h_1 would instead be any constant less than h .

A random sample of size n is drawn from an exponential distribution with mean θ to investigate the hypotheses

$$H_0 : \theta \leq 7 \quad H_1 : \theta > 7$$

Determine whether a UMP test exists for this setup.

Note that both H_0 and H_1 are composite. This test is UMP if there exists a monotone likelihood ratio. First, solve for the likelihood function.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \frac{1}{\theta} e^{-x_i / \theta} \\ &= \frac{1}{\theta^n} e^{-(\sum_{i=1}^n x_i) / \theta} \end{aligned}$$

Let constants a and b be possible values of θ , such that $a < b$. Then,

$$\begin{aligned} \Lambda &= \frac{L(a)}{L(b)} \\ &= \frac{\frac{1}{a^n} e^{-(\sum_{i=1}^n x_i) / a}}{\frac{1}{b^n} e^{-(\sum_{i=1}^n x_i) / b}} \\ &= \left(\frac{b}{a} \right)^n e^{(a-b)(\sum_{i=1}^n x_i) / ab} \end{aligned}$$

Λ is a function of the statistic $\sum_{i=1}^n x_i$. Since $a < b$, note that $a - b$ is negative. As a function of an arbitrary variable t , realize that e^{-t} is strictly decreasing. Therefore, as $\sum_{i=1}^n x_i$ increases, Λ must decrease. So, there exists a monotone decreasing likelihood ratio in the statistic $\sum_{i=1}^n x_i$.

As a result, a **UMP test does exist** with $\sum_{i=1}^n x_i \geq c$ as the critical region. As practice, we encourage you to derive for yourself this critical region form by manipulating the inequality

$$\frac{L(7)}{L(h_1)} \leq k$$

where h_1 is any constant greater than 7.

2.4 Summary

🕒 5m

A simple hypothesis fully specifies the distribution(s) associated with the sample.

A composite hypothesis does not fully specify the distribution(s) associated with the sample.

Most Powerful Test

When H_0 and H_1 are both simple, the most powerful test of size α has the largest power (i.e. the best critical region) among all tests with the same significance level α .

Neyman-Pearson Theorem

- If $L(h_0)$ represents the likelihood given by a simple H_0 , and $L(h_1)$ represents the likelihood given by a simple H_1 , then embedded in

$$\frac{L(h_0)}{L(h_1)} \leq k$$

is the best critical region.

- The key is that smaller values of the ratio of the likelihoods support rejecting H_0 .

Uniformly Most Powerful (UMP) Tests

- For a simple H_0 and composite H_1 , a test is UMP when the best critical region is the same for testing H_0 against each simple hypothesis in H_1 .
- The test with composite hypotheses

$$H_0 : \theta \leq h \quad H_1 : \theta > h$$

is UMP if there is a monotone likelihood ratio in a statistic y , i.e. $\frac{L(a)}{L(b)}$ is a monotone function of y , where $a < b$. Specifically:

- For a monotone decreasing likelihood ratio, the critical region has the form $y \geq c$.
- For a monotone increasing likelihood ratio, the critical region has the form $y \leq c$.
- Similarly, the test with composite hypotheses

$$H_0 : \theta \geq h \quad H_1 : \theta < h$$

could be UMP with the only difference being $a > b$.

2.5.0 Overview

5m

Most of the hypothesis tests we encountered so far investigate a parameter's value. But as shown in Example 2.4.1.2, we can use hypothesis tests to compare two distributions. Since parameters are defining characteristics of distributions, a broader view of these tests is that they investigate distributions.

In this subsection, we will study tests that emphasize checking the quality of a proposed distribution in fitting the data. The four tests we will consider are:

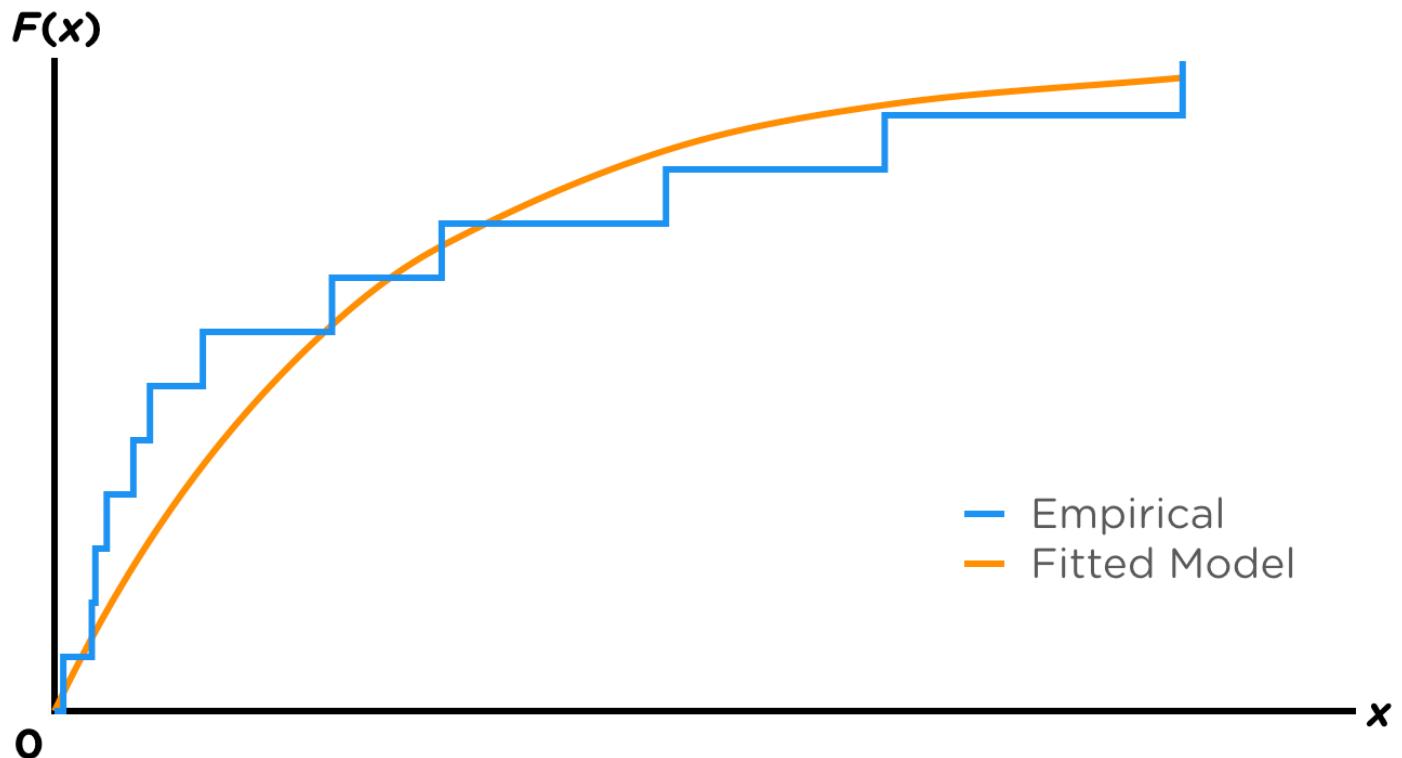
- The Kolmogorov-Smirnov test
- The chi-square goodness-of-fit test
- The chi-square test of independence
- The likelihood ratio test

2.5.1 Kolmogorov-Smirnov Test

A proposed distribution should accurately reflect the data across many aspects. One aspect is the similarity in cumulative probabilities. The **Kolmogorov-Smirnov test** examines this feature. We may state the hypotheses as follows:

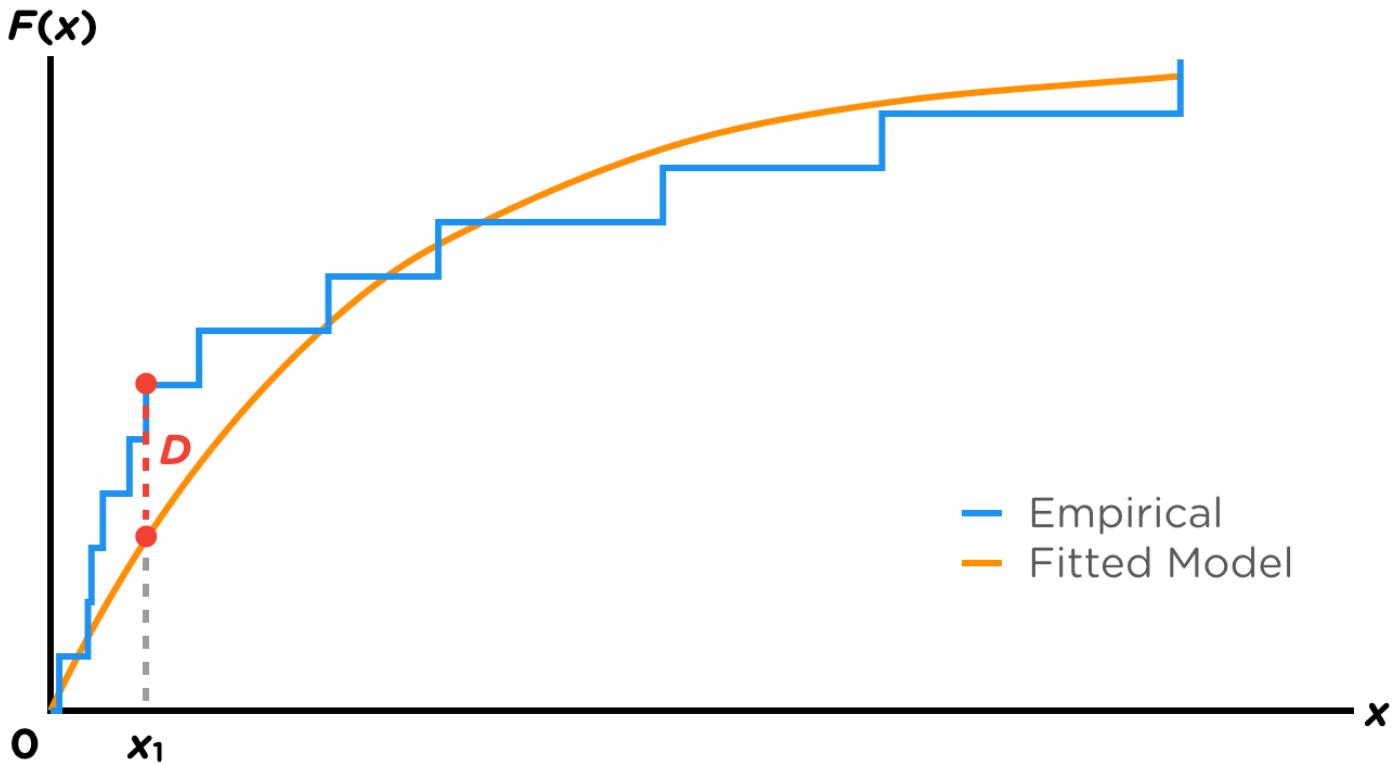
- H_0 : The proposed distribution adequately fits the data.
- H_1 : The proposed distribution does not adequately fit the data.

The following graph compares an empirical distribution function and a proposed/fitted distribution's CDF. The **empirical distribution function** calculates the sample cumulative probabilities by counting the observed values.



Looking at the graph, one may deduce that the proposed distribution does not fit the data very well for smaller values of x , but fits the data relatively well for larger values of x . These conclusions are subjective; the Kolmogorov-Smirnov test aims to reach an objective decision.

The test uses the largest difference between the empirical distribution function and the fitted CDF to evaluate the fit. The red dotted line in the graph below represents the largest difference between the two functions.



Let $\hat{F}(x)$ be the empirical distribution function and $F^*(x)$ be the CDF under the proposed distribution. The empirical distribution function is defined as:

$$\hat{F}(x) = \frac{\text{\# of observations} \leq x}{n} \quad (2.5.1.1)$$

where n is the sample size. Next, for $x_{(j)}$ denoting the j^{th} unique observed value in ascending order, let

$$D_j = \max\left(\left|\hat{F}(x_{(j)}) - F^*(x_{(j)})\right|, \left|\hat{F}(x_{(j-1)}) - F^*(x_{(j)})\right|\right) \quad (2.5.1.2)$$

Note that $\hat{F}(x_{(0)}) = 0$. Then, the Kolmogorov-Smirnov test statistic is

$$D = \max_j D_j \quad (2.5.1.3)$$

In other words, Equation 2.5.1.2 finds whether $F^*(x)$ differs more with $\hat{F}(x)$ at $x_{(j)}$ or immediately before $x_{(j)}$, taking the larger of the two. Then, Equation 2.5.1.3 compares those larger differences to find the largest difference overall. At first glance, these equations can be intimidating. Rather than memorizing them, we encourage you to understand the methodology and practice applying it. The examples are a good place to start.

Conceptually, D is the maximum absolute difference between $\hat{F}(x)$ and $F^*(x)$ for any given set of data. Therefore, D captures the worst disparity between the cumulative distributions; the larger D is, the more we suspect that the proposed distribution is a poor fit for the sample data.

This is a right-tailed test, and the critical values are given by the following table:

Significance Level, α	Critical Value
0.10	$\frac{1.22}{\sqrt{n}}$
0.05	$\frac{1.36}{\sqrt{n}}$
0.01	$\frac{1.63}{\sqrt{n}}$

There is no need to memorize the critical values. If they are needed, they will be given in the problem itself.

Example 2.5.1.1

You are given the following observations:

12 16 35 68 120

The Kolmogorov-Smirnov test is used to test the hypothesis that the data came from an exponential distribution with $\theta = 60$.

Calculate the Kolmogorov-Smirnov test statistic.

Solution

First, construct the fitted CDF and the empirical distribution function.

We are testing how well the exponential distribution fits the data. Thus, the proposed CDF is

$$F^*(x) = 1 - e^{-x/60}$$

Then, use Equation 2.5.1.1 to construct the empirical distribution function for the 5 data points.

$$\hat{F}(x) = \frac{\# \text{ of observations} \leq x}{5}$$

Next, calculate $F^*(x_{(j)})$, $\hat{F}(x_{(j)})$, and D_j for each observed value. For example, the fitted CDF at $x_{(1)} = 12$ is

$$F^*(12) = 1 - e^{-12/60} = 0.1813$$

and the empirical distribution function at $x_{(1)} = 12$ is

$$\hat{F}(12) = \frac{1}{5}$$

because there is only one observation that is at most 12. Then, calculate D_1 using Equation 2.5.1.2.

$$\left| \hat{F}(12) - F^*(12) \right| = |0.20 - 0.1813| = 0.0187$$

$$\left| \hat{F}(x_{(0)}) - F^*(12) \right| = |0.00 - 0.1813| = 0.1813$$

$$D_1 = \max(0.0187, 0.1813) = 0.1813$$

The corresponding values for all five data points are tabulated below.

j	$x_{(j)}$	$F^*(x_{(j)})$	$\hat{F}(x_{(j-1)})$	$\hat{F}(x_{(j)})$	D_j
1	12	0.1813	0.00	0.20	0.1813
2	16	0.2341	0.20	0.40	0.1659
3	35	0.4420	0.40	0.60	0.1580
4	68	0.6780	0.60	0.80	0.1220
5	120	0.8647	0.80	1.00	0.1353

Finally, calculate the test statistic. It is the largest D_j , so the Kolmogorov-Smirnov test statistic is $D = 0.1813$.

Coach's Remarks

Notice that $\hat{F}(x)$ increases uniformly from one observed value to the next. This occurs when all the observed values are unique. This is a good way to check whether the empirical distribution function columns are accurate.

Example 2.5.1.2

You fit an exponential distribution with $\theta = 100$ to the following data:

36 97 111 111 111 140 174 200

The critical values for this test are given in the table below:

Significance Level, α	Critical Value
0.10	$\frac{1.22}{\sqrt{n}}$
0.05	$\frac{1.36}{\sqrt{n}}$
0.01	$\frac{1.63}{\sqrt{n}}$

Determine the result of the Kolmogorov-Smirnov test.

Solution

Construct the fitted CDF and the empirical distribution function.

$$F^*(x) = 1 - e^{-x/100}$$

$$\hat{F}(x) = \frac{\# \text{ of observations} \leq x}{8}$$

Next, calculate $F^*(x_{(j)})$, $\hat{F}(x_{(j)})$, and D_j for each unique observed value. Notice that some values are repeated. For example, the empirical distribution function at $x_{(3)} = 111$ is

$$\begin{aligned}\hat{F}(111) &= \frac{\# \text{ of observations} \leq 111}{8} \\ &= \frac{5}{8} \\ &= 0.625\end{aligned}$$

Thus, calculate D_3 as follows:

$$\left| \hat{F}(111) - F^*(111) \right| = |0.625 - 0.6704| = 0.0454$$

$$\left| \hat{F}(97) - F^*(111) \right| = |0.250 - 0.6704| = 0.4204$$

$$D_3 = \max(0.0454, 0.4204) = 0.4204$$

The necessary values are tabulated below. Note that the Kolmogorov-Smirnov test statistic is 0.4959.

j	$x_{(j)}$	$F^*(x_{(j)})$	$\hat{F}(x_{(j-1)})$	$\hat{F}(x_{(j)})$	D_j
1	36	0.3023	0.000	0.125	0.3023
2	97	0.6209	0.125	0.250	0.4959
3	111	0.6704	0.250	0.625	0.4204
4	140	0.7534	0.625	0.750	0.1284
5	174	0.8245	0.750	0.875	0.0745
6	200	0.8647	0.875	1.000	0.1353

The critical values for this test are

α	Critical Value
0.10	$\frac{1.22}{\sqrt{8}} = 0.4313$
0.05	$\frac{1.36}{\sqrt{8}} = 0.4808$
0.01	$\frac{1.63}{\sqrt{8}} = 0.5763$

Since $0.4808 < 0.4959 < 0.5763$, we **reject H_0 at the 5% significance level, but not at the 1% significance level**. There is evidence that the exponential distribution with $\theta = 100$

does not fit the data well.



Coach's Remarks

Because the data has repeated values, $\hat{F}(x)$ does not increase uniformly. Even so, there is still a discernible pattern between the empirical distribution function columns that makes it easy to populate them.

Furthermore, instead of calculating the critical values, it is more efficient to calculate $D\sqrt{n}$ and compare it to the numerators of the critical values. In this example, $D\sqrt{8} = 1.4027$. Note that $1.36 < D\sqrt{8} < 1.63$, which brings us to the same conclusion more quickly.

Truncation and Censoring

If the data is incomplete, adjustments are needed. If the data is left-truncated at d , then

$$F^*(x) = \frac{F(x) - F(d)}{1 - F(d)}, \quad x > d \quad (2.5.1.4)$$

where $F(\cdot)$ is the CDF of the proposed distribution for the "ground-up" amount. Said differently, if $F(x)$ is the CDF of X , then $F^*(x)$ must be the CDF of $(X \mid X > d)$ to match the truncated data.

With censored data, $F^*(x)$ is unaffected. For data right-censored at m , the value of $\hat{F}(m)$ is undefined rather than 1. Moreover, n is the total number of observed values, including censored ones.

The following examples illustrate these concepts.

Example 2.5.1.3

You are given the following observations:

25 47 50 68 120

The Kolmogorov-Smirnov test is used to examine whether the data comes from a Pareto distribution with $\alpha = 2$ and $\theta = 300$. The data is left-truncated at 20.

The critical values for this test are given in the table below:

Significance Level	Critical Value
0.10	$\frac{1.22}{\sqrt{n}}$
0.05	$\frac{1.36}{\sqrt{n}}$
0.01	$\frac{1.63}{\sqrt{n}}$

Determine the result of the Kolmogorov-Smirnov test.

Solution

Since the data is left-truncated, an additional step is needed to determine $F^*(x)$. The proposed distribution is Pareto, so

$$F(x) = 1 - \left(\frac{300}{x+300} \right)^2$$

Consequently,

$$\begin{aligned} F^*(x) &= \frac{F(x) - F(20)}{1 - F(20)} \\ &= \frac{\left[1 - \left(\frac{300}{x+300} \right)^2 \right] - \left[1 - \left(\frac{300}{20+300} \right)^2 \right]}{1 - \left[1 - \left(\frac{300}{20+300} \right)^2 \right]} \\ &= \frac{\left(\frac{300}{20+300} \right)^2 - \left(\frac{300}{x+300} \right)^2}{\left(\frac{300}{20+300} \right)^2} \\ &= 1 - \left(\frac{320}{x+300} \right)^2 \end{aligned}$$

The empirical distribution function is

$$\hat{F}(x) = \frac{\# \text{ of observations} \leq x}{5}$$

Next, calculate $F^*(x_{(j)})$, $\hat{F}(x_{(j)})$, and D_j for each observed value. Obtain a test statistic of 0.5805.

j	$x_{(j)}$	$F^*(x_{(j)})$	$\hat{F}(x_{(j-1)})$	$\hat{F}(x_{(j)})$	D_j
1	25	0.0305	0.00	0.20	0.1695
2	47	0.1496	0.20	0.40	0.2504
3	50	0.1641	0.40	0.60	0.4359
4	68	0.2439	0.60	0.80	0.5561
5	120	0.4195	0.80	1.00	0.5805

To determine the test result efficiently, calculate $D\sqrt{n}$ and compare it to the numerators of the critical values.

$$0.5805\sqrt{5} = 1.2980$$

Since $1.22 < 1.2980 < 1.36$, we **reject H_0 at the 10% significance level, but not the 5% significance level.**



Example 2.5.1.4

You are given the following observations:

25 47 50 68 89 100 100 100

The Kolmogorov-Smirnov test is used to examine whether the data comes from a Weibull distribution with $\theta = 50$ and $\tau = 3$. The data is right-censored at 100.

The critical values for this test are given in the table below:

Significance Level, α	Critical Value
0.10	$\frac{1.22}{\sqrt{n}}$
0.05	$\frac{1.36}{\sqrt{n}}$
0.01	$\frac{1.63}{\sqrt{n}}$

Determine the test result at the 10% significance level.

Solution

Construct the fitted CDF and the empirical distribution function.

$$F^*(x) = 1 - e^{-(x/50)^3}$$

$$\hat{F}(x) = \frac{\# \text{ of observations} \leq x}{8}, \quad x < 100$$

While the censored values of 100 are included in $n = 8$, the empirical distribution function is not defined at $x = 100$.

Next, calculate $F^*(x_{(j)})$, $\hat{F}(x_{(j)})$, and D_j for each unique observed value, including 100. Obtain a test statistic of 0.5442.

j	$x_{(j)}$	$F^*(x_{(j)})$	$\hat{F}(x_{(j-1)})$	$\hat{F}(x_{(j)})$	D_j
1	25	0.1175	0.000	0.125	0.1175
2	47	0.5642	0.125	0.250	0.4392
3	50	0.6321	0.250	0.375	0.3821
4	68	0.9192	0.375	0.500	0.5442
5	89	0.9964	0.500	0.625	0.4964
6	100	0.9997	0.625	-	0.3747

Since $0.5442\sqrt{8} = 1.54 > 1.22$, we **reject H_0 at the 10% significance level.**



2.5.2 Chi-Square Goodness-of-Fit Test

The previous test checks the quality of a distribution's fit to the data by finding the largest disparity in the cumulative distributions. Now, consider an approach that does the same by measuring disparity across the distribution's domain instead.

The **chi-square goodness-of-fit test** starts by splitting the domain of the proposed distribution into k mutually exclusive intervals. This means each of the n observed values is found in only one of the k intervals. In addition, let

- q_j be the probability of being in interval j under the proposed distribution, and
- n_j be the number of observed values that fall in interval j ,

where $j = 1, \dots, k$. Then, the hypotheses are

- $H_0 : q_j$ is the true probability of being in interval j , for all $j = 1, \dots, k$
- $H_1 : \text{At least one } q_j \text{ is not the true probability of being in interval } j, \text{ for } j = 1, \dots, k$

In other words, we reject the idea that the proposed distribution fits the data well if, in at least one interval, the proposed probability poorly matches the corresponding observed proportion from the data, i.e. $q_j \neq \frac{n_j}{n}$.

The test statistic is

$$t.s. = \sum_{j=1}^k \frac{(n_j - nq_j)^2}{nq_j} \quad (2.5.2.1)$$

Coach's Remarks

The interpretation of nq_j is the **expected** number of observations in interval j , according to the proposed distribution. Remember that n_j is the **actual** observation count from the data. Hence, the test statistic captures the disparity between the data and distribution by including each interval's squared difference in the actual and expected counts.

Before discussing the distribution that the test statistic comes from, let's first practice calculating the test statistic.

Example 2.5.2.1

Company XYZ's one-year loss experience for 150 policies is as follows:

Interval	Number of Policies
0 - 7,500	79
7,500 - 17,500	40
17,500 - 32,500	16
32,500 - 67,500	11
67,500 - ∞	4

You test the null hypothesis that the loss size per policy follows an exponential distribution with $\theta = 12,750$.

Calculate the chi-square goodness-of-fit test statistic.

Solution

Start by calculating the probability for each interval. For instance, the probabilities for the first and second intervals are:

$$\begin{aligned} q_1 &= F(7,500) \\ &= 1 - e^{-7,500 / 12,750} \\ &= 0.4447 \end{aligned}$$

$$\begin{aligned} q_2 &= F(17,500) - F(7,500) \\ &= \left(1 - e^{-17,500 / 12,750}\right) - \left(1 - e^{-7,500 / 12,750}\right) \\ &= 0.3018 \end{aligned}$$

The policy counts and the probabilities are tabulated below:

	Interval		
1	0 - 7,500	79	0.4447

<i>j</i>	Interval	<i>n_j</i>	<i>q_j</i>
2	7,500 - 17,500	40	0.3018
3	17,500 - 32,500	16	0.1753
4	32,500 - 67,500	11	0.0731
5	67,500 - ∞	4	0.0050

There are $n = 150$ policies in total. Finally, calculate the test statistic.

$$\sum_{j=1}^5 \frac{(n_j - 150q_j)^2}{150q_j} = \frac{(79 - 150 \cdot 0.4447)^2}{150 \cdot 0.4447} + \dots + \frac{(4 - 150 \cdot 0.0050)^2}{150 \cdot 0.0050} \\ = \mathbf{20.9096}$$



Coach's Remarks

To verify the q_j calculations, sum them to check whether it equals 1. Here,
 $\sum_{j=1}^5 q_j = 0.4447 + \dots + 0.0050 = 1$.

Example 2.5.2.2

An experiment makes 24 rolls of a six-sided die. The outcomes are:

5	4	6	4	5	2	3	4
3	6	6	3	1	4	2	1
3	3	2	3	4	3	2	3

You want to test the null hypothesis that the die is fair using the chi-square goodness-of-fit test.

Calculate the chi-square goodness-of-fit test statistic.

Solution

A fair die means the proposed distribution is a discrete distribution with equal probability for each of the six outcomes. It is implied that the domain is split such that each possible value constitutes an "interval".

The die roll counts and the probabilities are tabulated below:

j	Die Roll	n_j	q_j
1	1	2	$\frac{1}{6}$
2	2	4	$\frac{1}{6}$
3	3	8	$\frac{1}{6}$
4	4	5	$\frac{1}{6}$
5	5	2	$\frac{1}{6}$
6	6	3	$\frac{1}{6}$

With $n = 24$, the test statistic is

$$\begin{aligned} \sum_{j=1}^6 \frac{(n_j - 24q_j)^2}{24q_j} &= \frac{(2 - 24 \cdot \frac{1}{6})^2}{24 \cdot \frac{1}{6}} + \dots + \frac{(3 - 24 \cdot \frac{1}{6})^2}{24 \cdot \frac{1}{6}} \\ &= 6.5 \end{aligned}$$



Let r denote the number of free parameters in the proposed distribution. A **free parameter** is a parameter whose value is not specified; it is typically estimated by maximum likelihood for this test.

For example, if the proposed distribution is Poisson with λ to be estimated by maximum likelihood, then $r = 1$. However, if the proposed distribution is Poisson with a predetermined value of λ (i.e. **not** estimated), then $r = 0$.

The test statistic (Equation 2.5.2.1) comes from an approximate chi-square distribution with $k - 1 - r$ degrees of freedom. The distribution is approximate because it is asymptotically chi-square as $n \rightarrow \infty$. Therefore, this test is right-tailed with a critical value of $\chi^2_{1-\alpha, k-1-r}$ at the α significance level; we reject H_0 when

$$t.s. \geq \chi^2_{1-\alpha, k-1-r}$$

Coach's Remarks

Degrees of freedom describes the number of unconstrained data points used in calculating a statistic. The calculation of Equation 2.5.2.1 involves k data points: n_1, \dots, n_k . However, not all k of them are free to take on any value. They must follow the constraint of

$$\sum_{j=1}^k n_j = n$$

Furthermore, when free parameters are estimated from the data, an additional constraint is imposed for each one. So, a total of $1 + r$ constraints will force $1 + r$ values on the n_j 's once the other $k - 1 - r$ values are specified. In other words, only $k - 1 - r$ values of the n_j 's are unconstrained in calculating the test statistic, i.e. there are $k - 1 - r$ degrees of freedom.

Example 2.5.2.3

Company XYZ's one-year loss experience for 300 policyholders is given as follows:

Number of Claims per Policy	Number of Policies
0	124
1	84
2	62
3	28
4	2

A Poisson distribution is fitted to the data. The mean of the Poisson distribution is estimated using maximum likelihood.

Determine the result of chi-square goodness-of-fit test based on the given chi-square table.

Solution

This goodness-of-fit test splits the entire domain into mutually exclusive intervals. Note that Poisson's domain includes all non-negative integers. Consequently, there are six intervals; the last interval covers 5 or more claims per policy with 0 of the 300 policies in it.

To calculate the probability for each interval, we must first estimate the Poisson's mean by MLE. For complete data, the MLE estimate of λ is the sample mean.

$$\hat{\lambda} = \bar{x} = \frac{0(124) + 1(84) + 2(62) + 3(28) + 4(2)}{300} = 1$$

Proceed to calculate the probabilities using a Poisson with mean 1. The policy counts and the probabilities are tabulated below:

j	Number of Claims	n_j	q_j
1	0	124	0.3679
2	1	84	0.3679
3	2	62	0.1839
4	3	28	0.0613
5	4	2	0.0153
6	5+	0	0.0037

Next, calculate the test statistic.

$$\begin{aligned} \sum_{j=1}^6 \frac{(n_j - 300q_j)^2}{300q_j} &= \frac{(124 - 300 \cdot 0.3679)^2}{300 \cdot 0.3679} + \dots + \frac{(0 - 300 \cdot 0.0037)^2}{300 \cdot 0.0037} \\ &= 16.4080 \end{aligned}$$

We have $k = 6$ and $r = 1$ (i.e. one free parameter was estimated), so the degrees of freedom is $6 - 1 - 1 = 4$. From the chi-square table, $\chi^2_{0.995, 4} = 14.86$. Since $16.4080 > 14.86$, we **reject H_0 at the 0.5% significance level**, meaning the Poisson distribution does not fit the data well.

In the next example, we revisit Example 2.5.2.1 with an added requirement: the expected number of losses in each interval must satisfy a minimum value.

Example 2.5.2.4

Company XYZ's one-year loss experience for 150 policies is as follows:

Interval	Number of Policies
0 - 7,500	79
7,500 - 17,500	40
17,500 - 32,500	16
32,500 - 67,500	11
67,500 - ∞	4

You test the null hypothesis that the loss size per policy follows an exponential distribution with $\theta = 12,750$.

Any interval in which the expected number of losses is less than five is combined with the previous interval.

Determine the result of chi-square goodness-of-fit test based on the given chi-square table.

Solution

Recall the following table from the solution of Example 2.5.2.1:

j	Interval	n_j	q_j
1	0 - 7,500	79	0.4447
2	7,500 - 17,500	40	0.3018
3	17,500 - 32,500	16	0.1753
4	32,500 - 67,500	11	0.0731
5	67,500 - ∞	4	0.0050

The expected number of losses in interval j refers to nq_j . Note that only $150q_5$ is less than five.

j	$150q_j$
1	66.7040
2	45.2769
3	26.2953
4	10.9706
5	0.7532

Thus, we fold the 5th interval into the 4th interval. This ensures that the expected number of losses in every interval is at least five.

j	Interval	n_j	q_j
1	0 - 7,500	79	0.4447
2	7,500 - 17,500	40	0.3018
3	17,500 - 32,500	16	0.1753
4	32,500 - ∞	15	0.0782

Next, calculate the test statistic.

$$\sum_{j=1}^4 \frac{(n_j - 150q_j)^2}{150q_j} = \frac{(79 - 150 \cdot 0.4447)^2}{150 \cdot 0.4447} + \dots + \frac{(15 - 150 \cdot 0.0782)^2}{150 \cdot 0.0782} \\ = 7.8280$$

We have $k = 4$ and $r = 0$ (i.e. θ was predetermined), so the degrees of freedom is $4 - 1 - 0 = 3$. From the chi-square table,

$$7.81 < 7.8280 < 9.35$$

$$\Rightarrow \chi^2_{0.95, 3} < 7.8280 < \chi^2_{0.975, 3}$$

Therefore, we **reject H_0 at the 5% significance level, but not at the 2.5% significance level.**

Coach's Remarks

Compared to this example, the test statistic in Example 2.5.2.1 is much larger. This demonstrates how the small $150q_5$ overwhelmed the test statistic; the fitted probability of a loss exceeding 67,500 is too low, relative to a mere 150 observations. A larger sample size is needed to properly assess an interval with that low a probability. Alternatively, combining intervals can help prevent low probabilities from skewing the test statistic.

Statistical literature recommends this so that the approximate chi-square distribution would be more reliable. But for exam purposes, follow the instruction in the problem.

2.5.3 Chi-Square Test of Independence

The **chi-square test of independence** is for data that can be organized into a contingency table. A **contingency table** records the frequency of observations described by two categorical variables. The test can examine the presence of dependence between the two variables. This is achieved using a similar procedure as the chi-square goodness-of-fit test. The hypotheses are

- H_0 : The two variables are independent.
- H_1 : The two variables are dependent.

One variable has a number of categories, while the other variable has b . Each of the n observations belongs to one of the a -by- b combinations. Let

- n_{ij} be the number of observations in Category i for the first variable and Category j for the second variable,
- $n_{i\bullet}$ be the subtotal number of observations in Category i for the first variable, across all categories of the second variable, and
- $n_{\bullet j}$ be the subtotal number of observations in Category j for the second variable, across all categories of the first variable,

for $i = 1, \dots, a$ and $j = 1, \dots, b$. Thus, a contingency table resembles

		Second Variable				Total
		Cat. 1	Cat. 2	...	Cat. b	
First Variable	Cat. 1	n_{11}	n_{12}	...	n_{1b}	$n_{1\bullet}$
	Cat. 2	n_{21}	n_{22}	...	n_{2b}	$n_{2\bullet}$
	:	:	:	..	:	:
	Cat. a	n_{a1}	n_{a2}	...	n_{ab}	$n_{a\bullet}$
Total		$n_{\bullet 1}$	$n_{\bullet 2}$...	$n_{\bullet b}$	n

The test statistic is

$$t.s. = \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \frac{(n_{ij} n - n_{i\bullet} n_{\bullet j})^2}{n_{i\bullet} n_{\bullet j}} \quad (2.5.3.1)$$

which comes from an approximate chi-square distribution with $(a - 1)(b - 1)$ degrees of freedom. Therefore, reject H_0 at the α significance level when

$$t.s. \geq \chi^2_{1-\alpha, (a-1)(b-1)}$$

Example 2.5.3.1

150 vehicles were stopped at random by the police for inspection.

		Year		Total
		< 2015	≥ 2015	
Type	Cars	40	60	100
	Motorcycles	10	40	50
Total		50	100	150

Test whether the vehicle type and year are independent based on the given chi-square table.

Solution

Note that $a = 2$ and $b = 2$ since each variable, vehicle type and year, has two categories.

Calculate the test statistic to obtain

$$\begin{aligned}
& \frac{1}{150} \sum_{i=1}^2 \sum_{j=1}^2 \frac{[n_{ij}(150) - n_{i\bullet} n_{\bullet j}]^2}{n_{i\bullet} n_{\bullet j}} \\
&= \frac{1}{150} \left\{ \frac{[40(150) - 100(50)]^2}{100(50)} + \frac{[60(150) - 100(100)]^2}{100(100)} \right. \\
&\quad \left. + \frac{[10(150) - 50(50)]^2}{50(50)} + \frac{[40(150) - 50(100)]^2}{50(100)} \right\} \\
&= 6
\end{aligned}$$

This test involves $(2 - 1)(2 - 1) = 1$ degree of freedom. From the chi-square table,

$$5.02 < 6 < 6.63$$

$$\Rightarrow \chi^2_{0.975, 1} < 6 < \chi^2_{0.99, 1}$$

In conclusion, we **reject H_0 at the 2.5% significance level, but not at the 1% level**, suggesting strong evidence that vehicle type and year are dependent.



2.5.4 Likelihood Ratio Test

The *likelihood ratio test* investigates the hypotheses

- H_0 : The data is drawn from Distribution A.
- H_1 : The data is drawn from Distribution B.

where Distribution A is a special case of Distribution B. For example:

- Exponential is a special case of gamma (with shape parameter of 1).
- A distribution with predetermined parameter values is a special case of the same distribution with free parameters.

This relation between the distributions implies that Distribution B is more complex than Distribution A, i.e. the former has more free parameters. In short, we prefer a simpler model (i.e. fewer free parameters), but will defer to a more complex model (i.e. more free parameters) if there is significant improvement.

Define

- r_0 as the number of free parameters in Distribution A,
- r_1 as the number of free parameters in Distribution B,
- L_0 as the maximized likelihood under H_0 , and
- L_1 as the maximized likelihood under H_1 .

As the name suggests, a *maximized likelihood* is simply $L(\cdot)$ evaluated at its MLE estimate(s). It follows that a *maximized log-likelihood* is the natural log of a maximized likelihood, or equivalently, $l(\cdot)$ evaluated at its MLE estimate(s).

Let l_0 and l_1 be the maximized log-likelihoods under their corresponding hypotheses, i.e. $l_0 = \ln(L_0)$ and $l_1 = \ln(L_1)$. Then, the likelihood ratio test statistic is

$$t.s. = -2 \ln \left(\frac{L_0}{L_1} \right) \quad (2.5.4.1)$$

$$\begin{aligned} &= -2 [\ln(L_0) - \ln(L_1)] \\ &= 2(l_1 - l_0) \end{aligned} \quad (2.5.4.2)$$

which comes from an approximate chi-square distribution with $r_1 - r_0$ degrees of freedom. Being a right-tailed test, we reject H_0 at the α significance level when

$$t.s. \geq \chi^2_{1-\alpha, r_1-r_0}$$

When calculating a numerical value for the test statistic, Equation 2.5.4.2 tends to be preferred over Equation 2.5.4.1. However, when obtaining a non-numerical test statistic (see Example 2.5.4.3), the formula of choice is typically up to personal preference.

Coach's Remarks

In addition to the aforementioned requirement on Distributions A and B, the approximate chi-square distribution also depends on the following:

- Large n (i.e. asymptotically chi-square)
- Typical regularity conditions hold
- Under both hypotheses, the maximum likelihood estimators are consistent solutions to the score equations

Example 2.5.4.1

You are given the following loss amounts:

12 16 35 68 120

Two models are fitted to the data:

- Exponential distribution with one parameter θ_E .
- Weibull distribution with two parameters τ and θ_W .

The maximum likelihood estimate of θ_E is 50.2. The maximum likelihood estimates of τ and θ_W are 1.28 and 54.43, respectively.

Calculate the likelihood ratio test statistic.

Solution

Note that the exponential distribution is a special case of the Weibull distribution where $\tau = 1$; it has one fewer free parameter. Thus, the exponential distribution is the model under H_0 .

Compute the test statistic using Equation 2.5.4.2. To obtain l_0 , begin by finding the log-likelihood under H_0 .

$$\begin{aligned}
 L(\theta) &= \prod_{i=1}^5 \frac{1}{\theta} e^{-x_i/\theta} \\
 &= \frac{1}{\theta^5} e^{-\left(\sum_{i=1}^5 x_i\right)/\theta} \\
 l(\theta) &= -5 \ln \theta - \frac{\sum_{i=1}^5 x_i}{\theta}
 \end{aligned}$$

$$\begin{aligned}
 l_0 &= l(\hat{\theta}_E) \\
 &= l(50.2) \\
 &= -5 \ln 50.2 - \frac{12 + 16 + 35 + 68 + 120}{50.2} \\
 &= -24.5801
 \end{aligned}$$

Next, find the log-likelihood under H_1 to solve for l_1 .

$$\begin{aligned}
 L(\theta, \tau) &= \prod_{i=1}^5 \frac{\tau x_i^{\tau-1} e^{-(x_i/\theta)^\tau}}{\theta^\tau} \\
 &= \frac{\tau^5 \left(\prod_{i=1}^5 x_i\right)^{\tau-1} e^{-\left(\sum_{i=1}^5 x_i^\tau\right)/\theta^\tau}}{\theta^{5\tau}} \\
 l(\theta, \tau) &= 5 \ln \tau + (\tau - 1) \sum_{i=1}^5 (\ln x_i) - \frac{\sum_{i=1}^5 x_i^\tau}{\theta^\tau} - 5\tau \ln \theta
 \end{aligned}$$

$$\begin{aligned}
 l_1 &= l(\hat{\theta}_W, \hat{\tau}) \\
 &= l(54.43, 1.28) \\
 &= 5 \ln 1.28 + 0.28 (\ln 12 + \dots + \ln 120) - \frac{12^{1.28} + \dots + 120^{1.28}}{54.43^{1.28}} - 5(1.28) \ln 54.43 \\
 &= -24.3582
 \end{aligned}$$

Thus, the test statistic is

$$2 [-24.3582 - (-24.5801)] = \mathbf{0.4437}$$

Coach's Remarks

Recall that we may drop multiplicative constants from $L(\cdot)$ when solving for MLE estimates. However, the likelihood ratio test statistic requires evaluating the actual $L(\cdot)$ or $l(\cdot)$. Make sure the MLE estimates are plugged into the right expressions, not those with dropped constants.

Example 2.5.4.2

In fitting a distribution to 100 observations, you consider the hypotheses:

- H_0 : The distribution is gamma with $\alpha = 4.5$ and θ
- H_1 : The distribution is gamma with α and θ

Under H_0 , the maximum likelihood estimate of θ is 57.0. Under H_1 , the maximum likelihood estimates of α and θ are 3.4 and 74.6, respectively.

You are also given:

- $\sum_{i=1}^{100} x_i = 25,600$
- $\sum_{i=1}^{100} \ln x_i = 539$
- $\Gamma(4.5) = 11.63$
- $\Gamma(3.4) = 2.98$

Determine the result of the likelihood ratio test at the 2.5% significance level.

Solution

Start by finding the log-likelihood function.

$$\begin{aligned}
L(\alpha, \theta) &= \prod_{i=1}^{100} \frac{x_i^{\alpha-1} e^{-x_i/\theta}}{\theta^\alpha \Gamma(\alpha)} \\
&= \frac{\left(\prod_{i=1}^{100} x_i\right)^{\alpha-1} e^{-\left(\sum_{i=1}^{100} x_i\right)/\theta}}{[\theta^\alpha \Gamma(\alpha)]^{100}} \\
l(\alpha, \theta) &= (\alpha - 1) \sum_{i=1}^{100} (\ln x_i) - \frac{\sum_{i=1}^{100} x_i}{\theta} - 100\alpha \ln \theta - 100 \ln [\Gamma(\alpha)] \\
&= (\alpha - 1)539 - \frac{25,600}{\theta} - 100\alpha \ln \theta - 100 \ln [\Gamma(\alpha)]
\end{aligned}$$

Next, calculate l_0 and l_1 .

$$\begin{aligned}
l_0 &= l(\alpha, \hat{\theta}) \\
&= l(4.5, 57.0) \\
&= 3.5 \cdot 539 - \frac{25,600}{57} - 100(4.5) \ln 57 - 100 \ln 11.63 \\
&= -627.3547
\end{aligned}$$

$$\begin{aligned}
l_1 &= l(\hat{\alpha}, \hat{\theta}) \\
&= l(3.4, 74.6) \\
&= 2.4 \cdot 539 - \frac{25,600}{74.6} - 100(3.4) \ln 74.6 - 100 \ln 2.98 \\
&= -624.8836
\end{aligned}$$

Therefore, the likelihood ratio test statistic is

$$2[-624.8836 - (-627.3547)] = 4.9421$$

Since the value of α was specified under H_0 , only θ is a free parameter, i.e. $r_0 = 1$. On the other hand, both α and θ are free parameters under H_1 , i.e. $r_1 = 2$. Hence, this test involves $2 - 1 = 1$ degree of freedom.

At the 2.5% significance level, the critical value is $\chi^2_{0.975, 1} = 5.02$. Since $4.9421 < 5.02$, we **fail to reject H_0 at the 2.5% significance level**. In this case, we prefer the simpler gamma distribution whose shape parameter is preset to 4.5.

Example 2.5.4.3

For a random sample of size n drawn from a normal distribution with mean μ and variance 25, the likelihood ratio test is used to examine the hypotheses:

- $H_0 : \mu = 6$
- $H_1 : \mu \neq 6$

Determine the critical region of the test.

- A. $\bar{x}^2 \geq c$
- B. $\sum_{i=1}^n x_i \geq c$
- C. $\sum_{i=1}^n x_i^2 \geq c$
- D. $|\bar{x} - 6| \geq c$
- E. $\sum_{i=1}^n (x_i - 6)^2 \geq c$

Solution

Let's practice calculating the likelihood ratio test statistic using Equation 2.5.4.1 instead.
Start by solving for the likelihood.

$$\begin{aligned} L(\mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi(25)}} \exp \left[-\frac{(x_i - \mu)^2}{2 \cdot 25} \right] \\ &= \frac{1}{(50\pi)^{n/2}} \exp \left[-\frac{\sum_{i=1}^n (x_i - \mu)^2}{50} \right] \end{aligned}$$

Recall that the sample mean is the MLE estimate of μ for a normal distribution with fixed variance. Then, calculate L_0 and L_1 .

$$L_0 = L(\mu) = L(6) = \frac{1}{(50\pi)^{n/2}} \exp \left[-\frac{\sum_{i=1}^n (x_i - 6)^2}{50} \right]$$

$$L_1 = L(\hat{\mu}) = L(\bar{x}) = \frac{1}{(50\pi)^{n/2}} \exp \left[-\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{50} \right]$$

Use Equation 2.5.4.1 to determine the test statistic.

$$\begin{aligned} \frac{L_0}{L_1} &= \frac{\frac{1}{(50\pi)^{n/2}} \exp \left[-\frac{\sum_{i=1}^n (x_i - 6)^2}{50} \right]}{\frac{1}{(50\pi)^{n/2}} \exp \left[-\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{50} \right]} \\ &= \exp \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2 - \sum_{i=1}^n (x_i - 6)^2}{50} \right] \\ &= \exp \left[\frac{\sum_{i=1}^n x_i^2 - 2\bar{x}(\sum_{i=1}^n x_i) + n\bar{x}^2 - \{\sum_{i=1}^n x_i^2 - 2(6)(\sum_{i=1}^n x_i) + 12n\}}{50} \right] \\ &= \exp \left[\frac{-2\bar{x}(n\bar{x}) + n\bar{x}^2 + 2(6)(n\bar{x}) - n6^2}{50} \right] \\ &= \exp \left[-\frac{n(\bar{x}^2 - 2\bar{x}6 + 6^2)}{50} \right] \\ &= \exp \left[-\frac{n(\bar{x} - 6)^2}{50} \right] \\ -2 \ln \left(\frac{L_0}{L_1} \right) &= -2 \left[-\frac{n(\bar{x} - 6)^2}{50} \right] \\ &= \frac{n}{25} (\bar{x} - 6)^2 \end{aligned}$$

If k represents the critical value in terms of the approximate chi-square distribution, then the critical region has the form

$$\frac{n}{25} (\bar{x} - 6)^2 \geq k$$

Thus, we must manipulate the inequality in order to match one of the answer choices.

$$\begin{aligned}
 & \frac{n}{25}(\bar{x} - 6)^2 \geq k \\
 \Rightarrow & (\bar{x} - 6)^2 \geq \frac{25k}{n} \\
 \Rightarrow & \left[\bar{x} - 6 \leq -\sqrt{\frac{25k}{n}} \right] \cup \left[\bar{x} - 6 \geq \sqrt{\frac{25k}{n}} \right] \\
 \Rightarrow & |\bar{x} - 6| \geq \sqrt{\frac{25k}{n}}
 \end{aligned}$$

In conclusion, the answer is (D).



Coach's Remarks

For extra credit, notice the test statistic can be written as

$$\begin{aligned}
 \frac{n}{25}(\bar{x} - 6)^2 &= \frac{(\bar{x} - 6)^2}{25/n} \\
 &= \frac{(\bar{x} - 6)^2}{(5/\sqrt{n})^2} \\
 &= \left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right)^2 \\
 &= z^2
 \end{aligned}$$

Therefore, testing a normal distribution's mean when the variance is known produces a likelihood ratio test statistic that comes from an **exact** chi-square distribution with 1 degree of freedom. In general, it is not easy to obtain exact distributions, so we rely on the asymptotic result of an approximate chi-square distribution for likelihood ratio tests.

2.5 Summary

🕒 5m

Kolmogorov-Smirnov Test

$F^*(x)$ is the CDF of the proposed distribution, and $\hat{F}(x)$ is the empirical distribution function, where

$$\hat{F}(x) = \frac{\text{\# of observations} \leq x}{n}$$

The test statistic, D , is the maximum absolute difference between $\hat{F}(x)$ and $F^*(x)$. Reject H_0 if D is greater than the critical value.

If the data is left-truncated at d ,

$$F^*(x) = \frac{F(x) - F(d)}{1 - F(d)}$$

If the data is right-censored at m , then $\hat{F}(m)$ is undefined, not 1. The censored values are included in n .

Chi-Square Goodness-of-Fit Test

- The domain of the proposed distribution is split into k mutually exclusive intervals.
- q_j is the probability of being in interval j under the proposed distribution.
- n_j is the number of observed values that fall in interval j .
- r is the number of free parameters in the proposed distribution.

$$t. s. = \sum_{j=1}^k \frac{(n_j - nq_j)^2}{nq_j}$$

For a size α test, reject H_0 if $t. s. \geq \chi^2_{1-\alpha, k-1-r}$.

Chi-Square Test of Independence

For a categorical variable with a categories and another with b categories,

- n_{ij} is the number of observations in Category i for the first variable and Category j for the second variable
- $n_{i\bullet}$ is the subtotal number of observations in Category i for the first variable, across all categories of the second variable
- $n_{\bullet j}$ is the subtotal number of observations in Category j for the second variable, across all categories of the first variable

for $i = 1, \dots, a$ and $j = 1, \dots, b$.

$$t. s. = \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \frac{(n_{ij} n - n_{i\bullet} n_{\bullet j})^2}{n_{i\bullet} n_{\bullet j}}$$

For a size α test, reject H_0 if $t. s. \geq \chi^2_{1-\alpha, (a-1)(b-1)}$.

Likelihood Ratio Test

With H_0 supporting Distribution A and H_1 supporting Distribution B, let

- Distribution A be a special case of Distribution B
- r_0 be the number of free parameters in Distribution A
- r_1 be the number of free parameters in Distribution B
- L_0 be the maximized likelihood under H_0
- L_1 be the maximized likelihood under H_1
- $l_0 = \ln(L_0)$
- $l_1 = \ln(L_1)$

$$t. s. = -2 \ln \left(\frac{L_0}{L_1} \right) = 2(l_1 - l_0)$$

For a size α test, reject H_0 if $t. s. \geq \chi^2_{1-\alpha, r_1 - r_0}$.

2.6.0 Overview

 5m

This subsection introduces the second approach to estimating a parameter: interval estimation. While point estimation provides a single-value "best guess", **interval estimation** produces a "best guess" for a parameter in the form of a range of values. We achieve this by constructing a confidence interval.

We present confidence intervals as they apply to parameters like means, proportions, and variances. These are the same types of parameters discussed in Section 2.3: Hypothesis Testing. In fact, the specific scenarios seen in Sections 2.3.3 to 2.3.5 are revisited. The connections between confidence intervals and hypothesis tests are also covered here.

We begin with some introductory concepts at the core of confidence intervals.

2.6.1 Introduction

A **confidence interval** suggests where the value of a parameter of interest lies with a certain level of confidence. Broadly speaking, we require the following to obtain a confidence interval:

- A methodology
- A suitable distribution
- A confidence level
- Data

In general, consider

$$\Pr(L \leq \theta \leq U) = k$$

This describes a random interval (L, U) that encloses a generic parameter θ . Notice that the random components above are L and U . To be clear, θ is **not** random since it is a parameter. When data is used to evaluate L and U , the resulting numerical interval is a $100k\%$ confidence interval for θ . Thus, k is known as the **confidence level**.

Pivotal Method

We construct a confidence interval using the pivotal method. It relies on a **pivotal quantity**: a function of random variables from a sample and unknown parameters of the sample's distribution. Thus, a pivotal quantity is also a random variable. Importantly, the distribution of a pivotal quantity must not depend on the aforementioned parameters.

Coach's Remarks

The definition of a pivotal quantity is similar yet distinct from the definition of a statistic (Section 2.2.1). Here is how they differ:

If a pivotal quantity consists of unknown parameters, then the pivotal quantity is not a statistic.

If a statistic's distribution depends on unknown parameters, then the statistic is not a pivotal quantity.

The following simplified example demonstrates how this method works.

Random variable X has CDF

$$F_X(x) = 1 - \left(\frac{\theta}{x + \theta} \right)^2, \quad x > 0$$

A pivotal quantity is $Y = \frac{\theta}{X}$, with CDF

$$F_Y(y) = \left(\frac{y}{y + 1} \right)^2, \quad y > 0$$

noting that its distribution does not depend on θ .

Assuming X was observed to equal 3.5, use this pivotal quantity to construct a 95% confidence interval for the parameter θ .

Start with the following probability statement involving the pivotal quantity:

$$\Pr(a \leq Y \leq b) = 0.95$$

Let a be the 2.5th percentile of Y , and b the 97.5th percentile of Y . Thus,

$$\left(\frac{a}{a + 1} \right)^2 = 0.025 \Rightarrow a = 0.1878$$

$$\left(\frac{b}{b + 1} \right)^2 = 0.975 \Rightarrow b = 78.4968$$

Now obtain $\Pr(L \leq \theta \leq U) = 0.95$ by manipulating the initial probability statement.

$$\begin{aligned}\Pr(0.1878 \leq Y \leq 78.4968) &= 0.95 \\ \Rightarrow \Pr\left(0.1878 \leq \frac{\theta}{X} \leq 78.4968\right) &= 0.95 \\ \Rightarrow \Pr(0.1878X \leq \theta \leq 78.4968X) &= 0.95\end{aligned}$$

This means the random interval (L, U) is $(0.1878X, 78.4968X)$. Having observed $x = 3.5$, the 95% confidence interval for θ is

$$(0.1878 [3.5], 78.4968 [3.5]) = (\mathbf{0.6573}, \mathbf{274.7389})$$

As a result, we produced an interval which we assert – with reasonably high confidence – includes the value of θ . The key takeaways are:

- In general, L and U are functions of random variables and constants. The random variables are evaluated at the data to compute the confidence interval.
- The constants in L and U come from the desired confidence level and the pivotal quantity's distribution.

Coach's Remarks

There is often confusion in interpreting a confidence interval. For the example above, the temptation is to conclude that there is a 95% **probability** that θ is between 0.6573 and 274.7389; this is an **incorrect** interpretation of the interval. Remember that θ has a fixed (albeit unknown) value. The value is simply within the interval, or is not. There is no element of chance or randomness.

Instead, it is accurate to say that 95% of numerous, similarly constructed intervals would enclose θ . While $(0.6573, 274.7389)$ is one of these intervals, it is practically impossible to know whether it is among the 95% that would enclose θ , or the 5% that would not. This is what it means to be 95% confident.

2.6.2 Intervals for Means

As in Section 2.3.3, the same three broad categories are addressed here:

1. One sample
2. Two samples
3. Two samples with paired observations

We similarly consider the sub-cases of known variance and unknown variance in each category.

One Sample

KNOWN VARIANCE

Let one random sample of size n be drawn from any distribution with mean μ and known variance σ^2 . When n is large, then by the Central Limit Theorem, \bar{X} is approximately normally distributed. In turn,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has an approximate standard normal distribution. Note that the expression is a pivotal quantity, since

- it is a function of X_1, \dots, X_n , and unknown μ , and
- the standard normal distribution does not depend on μ .

For this setup, the bounds of a symmetric random interval (L, U) that encloses μ with probability k are:

$$L = \bar{X} + z_{(1-k)/2} \frac{\sigma}{\sqrt{n}}$$

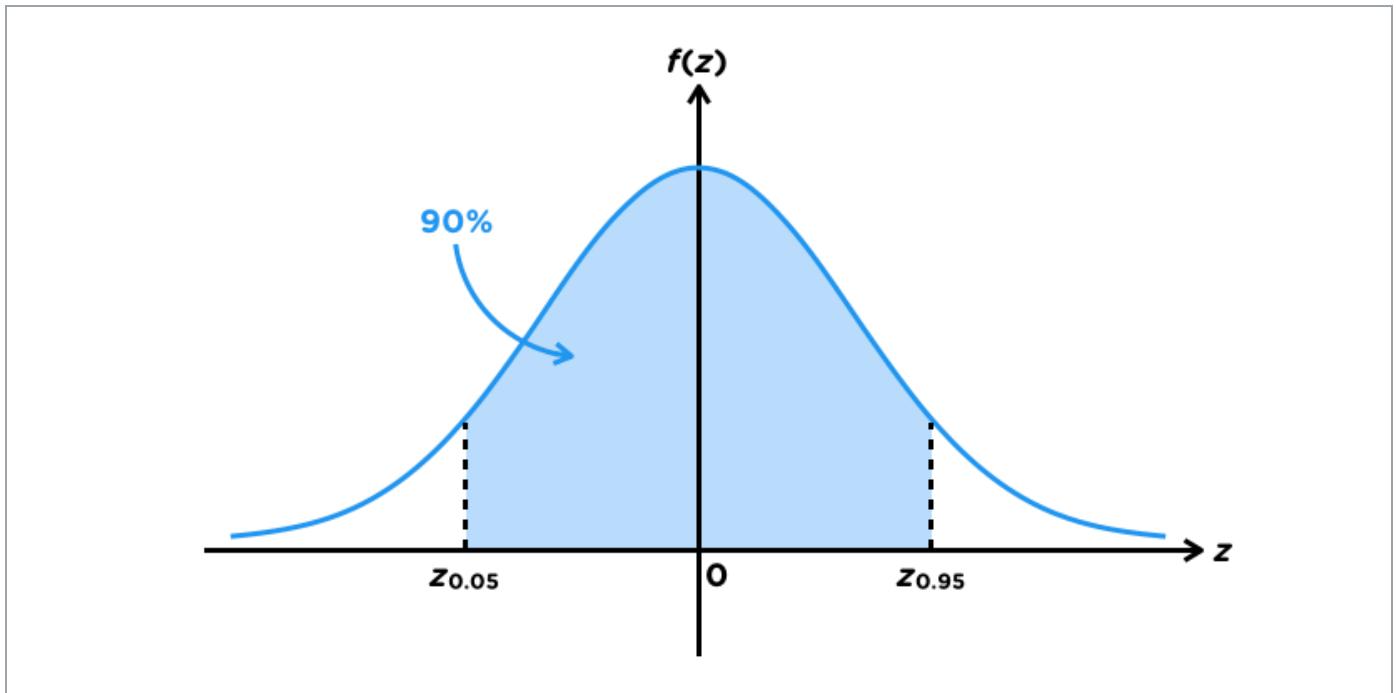
$$U = \bar{X} + z_{(1+k)/2} \frac{\sigma}{\sqrt{n}}$$

where z_q is the $100q^{\text{th}}$ percentile of the standard normal distribution.

Coach's Remarks

Realize that $z_{(1-k)/2}$ and $z_{(1+k)/2}$ produce a symmetric interval where Z is between them with probability k . The graph below illustrates the case when $k = 0.9$, hence

$$z_{(1-0.9)/2} = z_{0.05}, \quad z_{(1+0.9)/2} = z_{0.95}$$



In addition, realize that

$$z_{(1-k)/2} = -z_{(1+k)/2}$$

since the standard normal distribution is symmetric about 0. Therefore, the $100k\%$ confidence interval for μ is often given by the expression

$$\bar{x} \pm z_{(1+k)/2} \frac{\sigma}{\sqrt{n}} \quad (2.6.2.1)$$

Next, consider instead the random interval $(-\infty, U)$ that encloses μ with probability k . Then,

$$U = \bar{X} + z_k \frac{\sigma}{\sqrt{n}}$$

This leads to the $100k\%$ left-sided (or upper bound) confidence interval for μ given by

$$\left(-\infty, \bar{x} + z_k \frac{\sigma}{\sqrt{n}}\right) \quad (2.6.2.2)$$

Likewise, the random interval (L, ∞) that encloses μ with probability k has

$$\begin{aligned} L &= \bar{X} + z_{1-k} \frac{\sigma}{\sqrt{n}} \\ &= \bar{X} - z_k \frac{\sigma}{\sqrt{n}} \end{aligned}$$

which leads to the $100k\%$ right-sided (or lower bound) confidence interval for μ given by

$$\left(\bar{x} - z_k \frac{\sigma}{\sqrt{n}}, \infty\right) \quad (2.6.2.3)$$

UNKNOWN VARIANCE

If the random sample is drawn from a normal distribution instead, then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a t -distribution with $n - 1$ degrees of freedom. It is also a pivotal quantity. In other words, there are only two changes relative to the known variance scenario:

1. S replaces σ
2. t percentiles replace standard normal percentiles

With $t_{2(1-q), df}$ denoting the $100q^{\text{th}}$ percentile of a t -distribution with df degrees of freedom, we have

- the $100k\%$ confidence interval for μ given by

$$\bar{x} \pm t_{1-k, n-1} \frac{s}{\sqrt{n}} \quad (2.6.2.4)$$

- the $100k\%$ left-sided confidence interval for μ given by

$$\left(-\infty, \bar{x} + t_{2(1-k), n-1} \frac{s}{\sqrt{n}}\right) \quad (2.6.2.5)$$

- the $100k\%$ right-sided confidence interval for μ given by

$$\left(\bar{x} - t_{2(1-k), n-1} \frac{s}{\sqrt{n}}, \infty\right) \quad (2.6.2.6)$$

Example 2.6.2.1

After administering a standardized test, an exam committee obtained a random sample of 15 tests. Test scores are assumed to have mean μ and standard deviation 1.25. The sample average score of the 15 tests is 5.

Construct the 80% confidence interval for μ .

Solution

Since the variance is known, the confidence interval expression for μ is

$$\bar{x} \pm z_{(1+k)/2} \frac{\sigma}{\sqrt{n}}$$

We are given

$$n = 15, \quad \bar{x} = 5, \quad \sigma = 1.25, \quad k = 0.8$$

From the exam table, obtain $z_{(1+0.8)/2} = z_{0.9} = 1.282$.

Calculate the answer as

$$\left(5 - 1.282 \cdot \frac{1.25}{\sqrt{15}}, 5 + 1.282 \cdot \frac{1.25}{\sqrt{15}} \right) = (4.586, 5.414)$$



Example 2.6.2.2

After administering a standardized test, an exam committee obtained a random sample of 15 tests. Test scores are assumed to be normal distributed with mean μ . The sample average score and sample standard deviation of the 15 tests are 5 and 1.25, respectively.

Construct the 80% confidence interval for μ .

Solution

Since the variance is unknown, the confidence interval expression for μ is

$$\bar{x} \pm t_{1-k, n-1} \frac{s}{\sqrt{n}}$$

We are given

$$n = 15, \quad \bar{x} = 5, \quad s = 1.25, \quad k = 0.8$$

From the exam table, obtain $t_{1-0.8, 15-1} = t_{0.2, 14} = 1.345$.

Calculate the answer as

$$\left(5 - 1.345 \cdot \frac{1.25}{\sqrt{15}}, 5 + 1.345 \cdot \frac{1.25}{\sqrt{15}} \right) = (4.566, 5.434)$$

■

A recurring theme is that the test statistics in Sections 2.3.3 to 2.3.5 are realizations of pivotal quantities used in creating confidence intervals. Consequently, there is an appealing connection between hypothesis tests and confidence intervals:

For any two-tailed hypothesis test presented in Sections 2.3.3 to 2.3.5, H_0 will fail to be rejected at the α significance level if the hypothesized value h is within the $100(1 - \alpha)\%$ confidence interval for the parameter. Otherwise, H_0 will be rejected. See the appendix at the end of this section for the proof. The same holds for one-tailed tests and their corresponding one-sided confidence intervals.

Moreover, confidence intervals for means have the following generalization:

A pivotal quantity of the form

$$\frac{\text{estimator} - \text{mean parameter}}{\text{standard error}}$$

will result in the $100k\%$ confidence interval for the mean parameter having the form

$$\text{estimate} \pm (\text{percentile}) (\text{standard error})$$

where percentile is the $100\left(\frac{1+k}{2}\right)^{\text{th}}$ percentile of either the standard normal distribution or a t -distribution. In addition, the $100k\%$ left/right-sided confidence interval produces an upper/lower bound of the form

estimate $+/-$ (percentile) (standard error)

where percentile is the $100k^{\text{th}}$ percentile of either the standard normal distribution or a t -distribution.

Two Samples

KNOWN VARIANCES

Let two independent random samples of sizes n_1 and n_2 be drawn from any two distributions with means μ_1 and μ_2 , and known variances σ_1^2 and σ_2^2 . When n_1 and n_2 are large, then by the Central Limit Theorem, the pivotal quantity

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

follows an approximate standard normal distribution.

Therefore, we have

- the $100k\%$ confidence interval for $\mu_1 - \mu_2$ given by

$$\bar{x}_1 - \bar{x}_2 \pm z_{(1+k)/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (2.6.2.7)$$

- the $100k\%$ left-sided confidence interval for $\mu_1 - \mu_2$ given by

$$\left(-\infty, \bar{x}_1 - \bar{x}_2 + z_k \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \quad (2.6.2.8)$$

- the $100k\%$ right-sided confidence interval for $\mu_1 - \mu_2$ given by

$$\left(\bar{x}_1 - \bar{x}_2 - z_k \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \infty \right) \quad (2.6.2.9)$$

UNKNOWN VARIANCES

If the random samples are each drawn from a normal distribution and $\sigma_1^2 = \sigma_2^2$, then the pivotal quantity

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

follows a t -distribution with $n_1 + n_2 - 2$ degrees of freedom.

Therefore, we have

- the 100k% confidence interval for $\mu_1 - \mu_2$ given by

$$\bar{x}_1 - \bar{x}_2 \pm t_{1-k, n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (2.6.2.10)$$

- the 100k% left-sided confidence interval for $\mu_1 - \mu_2$ given by

$$\left(-\infty, \bar{x}_1 - \bar{x}_2 + t_{2(1-k), n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \quad (2.6.2.11)$$

- the 100k% right-sided confidence interval for $\mu_1 - \mu_2$ given by

$$\left(\bar{x}_1 - \bar{x}_2 - t_{2(1-k), n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \infty \right) \quad (2.6.2.12)$$

Example 2.6.2.3

Scientists study the daily average temperatures of two independent cities. For each city, they randomly sample five days. Assume the two samples are drawn from two normal distributions having the same variance.

The 90% confidence interval for City A's mean temperature minus City B's mean temperature is (-11.06, -0.94).

Calculate the pooled sample variance.

Solution

Since this problem describes the two-sample, unknown variances setup, the confidence interval has the form of Equation 2.6.2.10. Given that the interval is symmetric, its midpoint must be $\bar{x}_A - \bar{x}_B$.

$$\bar{x}_A - \bar{x}_B = \frac{-11.06 + (-0.94)}{2} = -6$$

Additionally, we are given

$$n_A = n_B = 5, \quad k = 0.9$$

From the exam table, obtain $t_{1-0.9, 5+5-2} = t_{0.1, 8} = 1.860$.

Finally, solve for s_p^2 using either the upper or lower bound of the confidence interval. We demonstrate with the upper bound.

$$\bar{x}_A - \bar{x}_B + t_{0.1, 8} \cdot s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = -0.94$$

$$-6 + 1.860 \cdot s_p \sqrt{\frac{1}{5} + \frac{1}{5}} = -0.94$$

$$s_p = \frac{6 - 0.94}{1.860 \sqrt{2/5}}$$

$$s_p^2 = (4.3014)^2$$

$$= \mathbf{18.5}$$

■

Paired Observations

As previously discussed, the differences of paired observations lead to a one-sample setup. This means the confidence intervals for $\mu_1 - \mu_2$ are identical to Equations 2.6.2.1 to 2.6.2.6 with the following substitutions:

- $\bar{x} \rightarrow \bar{d}$
- $\sigma^2 \rightarrow \sigma_D^2$
- $n \rightarrow n_*$
- $s^2 \rightarrow s_D^2$

2.6.3 Intervals for Proportions

It should be no surprise that confidence intervals for proportions are very similar to confidence intervals for means, given the details covered in Section 2.3.4. Assuming we can rely on the Central Limit Theorem, the standard normal distribution is used to create these intervals.

One Sample

If a random sample from a Bernoulli distribution has a large sample size, then by the Central Limit Theorem, the pivotal quantity

$$\frac{\hat{q} - q}{\sqrt{\frac{\hat{q}(1 - \hat{q})}{n}}}$$

follows an approximate standard normal distribution. Note that this pivotal quantity does **not** exactly match the test statistic of Equation 2.3.4.1; the standard error term uses \hat{q} here. This makes the confidence interval general forms we saw for means applicable to proportions.

Therefore, we have

- the $100k\%$ confidence interval for q given by

$$\hat{q} \pm z_{(1+k)/2} \sqrt{\frac{\hat{q}(1 - \hat{q})}{n}} \quad (2.6.3.1)$$

- the $100k\%$ left-sided confidence interval for q given by

$$\left(-\infty, \hat{q} + z_k \sqrt{\frac{\hat{q}(1 - \hat{q})}{n}} \right) \quad (2.6.3.2)$$

- the $100k\%$ right-sided confidence interval for q given by

$$\left(\hat{q} - z_k \sqrt{\frac{\hat{q}(1 - \hat{q})}{n}}, \infty \right) \quad (2.6.3.3)$$

Two Samples

If two independent random samples, each from a Bernoulli distribution, have large sample sizes, then by the Central Limit Theorem, the pivotal quantity

$$\frac{\hat{q}_1 - \hat{q}_2 - (q_1 - q_2)}{\sqrt{\frac{\hat{q}_1(1 - \hat{q}_1)}{n_1} + \frac{\hat{q}_2(1 - \hat{q}_2)}{n_2}}}$$

follows an approximate standard normal distribution.

Therefore, we have

- the $100k\%$ confidence interval for $q_1 - q_2$ given by

$$\hat{q}_1 - \hat{q}_2 \pm z_{(1+k)/2} \sqrt{\frac{\hat{q}_1(1 - \hat{q}_1)}{n_1} + \frac{\hat{q}_2(1 - \hat{q}_2)}{n_2}} \quad (2.6.3.4)$$

- the $100k\%$ left-sided confidence interval for $q_1 - q_2$ given by

$$\left(-\infty, \hat{q}_1 - \hat{q}_2 + z_k \sqrt{\frac{\hat{q}_1(1 - \hat{q}_1)}{n_1} + \frac{\hat{q}_2(1 - \hat{q}_2)}{n_2}} \right) \quad (2.6.3.5)$$

- the $100k\%$ right-sided confidence interval for $q_1 - q_2$ given by

$$\left(\hat{q}_1 - \hat{q}_2 - z_k \sqrt{\frac{\hat{q}_1(1 - \hat{q}_1)}{n_1} + \frac{\hat{q}_2(1 - \hat{q}_2)}{n_2}}, \infty \right) \quad (2.6.3.6)$$

Example 2.6.3.1

Researchers tossed two coins each 100 times. You are given:

- H_0 : The probability of tossing a heads with Coin 1 is more than the probability for Coin 2 by h
- H_1 : The probability of tossing a heads with Coin 1 is not more than the probability for Coin 2 by h
- 30 heads were recorded for Coin 1.
- 26 heads were recorded for Coin 2.

Determine the largest value of h where H_0 is not rejected at the 5% significance level.

Solution

Note that h refers to the difference in probabilities. Thus, this is a two-tailed test.

$$H_0 : q_1 - q_2 = h \quad H_1 : q_1 - q_2 \neq h$$

Failing to reject H_0 at $\alpha = 0.05$ means that the hypothesized value h is within the $100(1 - 0.05) = 95\%$ confidence interval. So, the largest h that satisfies not rejecting H_0 is the upper bound of the 95% confidence interval for $q_1 - q_2$.

Note that

$$n_1 = n_2 = 100, \quad \hat{q}_1 = \frac{30}{100} = 0.3, \quad \hat{q}_2 = \frac{26}{100} = 0.26, \quad k = 0.95$$

From the exam table, obtain $z_{(1+0.95)/2} = z_{0.975} = 1.96$.

Therefore, the answer is

$$\begin{aligned} \hat{q}_1 - \hat{q}_2 + z_{0.975} \sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \frac{\hat{q}_2(1-\hat{q}_2)}{n_2}} &= 0.3 - 0.26 + 1.96 \sqrt{\frac{0.3 \cdot 0.7}{100} + \frac{0.26 \cdot 0.74}{100}} \\ &= \mathbf{0.16} \end{aligned}$$



2.6.4 Intervals for Variances

In this subsection, the pivotal quantities do not have the form

$$\frac{\text{estimator} - \text{mean parameter}}{\text{standard error}}$$

Thus, confidence intervals for variances are very different from those in previous subsections.

One Sample

For one random sample of size n drawn from a normal distribution with variance σ^2 , the pivotal quantity

$$\frac{(n-1)S^2}{\sigma^2}$$

follows a chi-square distribution with $n - 1$ degrees of freedom.

With $\chi_{q, \text{df}}^2$ denoting the $100q^{\text{th}}$ percentile of a chi-square distribution with df degrees of freedom, we have

- the $100k\%$ confidence interval for σ^2 given by

$$\left(\frac{(n-1)s^2}{\chi_{(1+k)/2, n-1}^2}, \frac{(n-1)s^2}{\chi_{(1-k)/2, n-1}^2} \right) \quad (2.6.4.1)$$

- the $100k\%$ left-sided confidence interval for σ^2 given by

$$\left(0, \frac{(n-1)s^2}{\chi_{1-k, n-1}^2} \right) \quad (2.6.4.2)$$

- the $100k\%$ right-sided confidence interval for σ^2 given by

$$\left(\frac{(n-1)s^2}{\chi_{k,n-1}^2}, \infty \right) \quad (2.6.4.3)$$

The chi-square distribution being right-skewed contributes to Equation 2.6.4.1 not having any symmetry. Unlike the confidence intervals for means and proportions, the midpoint of the bounds in Equation 2.6.4.1 is not useful.

Two Samples

For two independent random samples of sizes n_1 and n_2 , each drawn from a normal distribution, with variances σ_1^2 and σ_2^2 , the pivotal quantity

$$\frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2}$$

follows an F -distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom.

With $F_{1-q, \text{ndf}, \text{ddf}}$ denoting the $100q^{\text{th}}$ percentile of an F -distribution with ndf numerator degrees of freedom and ddf denominator degrees of freedom, we have

- the $100k\%$ confidence interval for σ_1^2/σ_2^2 given by

$$\left(\frac{s_1^2}{s_2^2} \cdot (F_{(1-k)/2, n_1-1, n_2-1})^{-1}, \frac{s_1^2}{s_2^2} \cdot F_{(1-k)/2, n_2-1, n_1-1} \right) \quad (2.6.4.4)$$

- the $100k\%$ left-sided confidence interval for σ_1^2/σ_2^2 given by

$$\left(0, \frac{s_1^2}{s_2^2} \cdot F_{1-k, n_2-1, n_1-1} \right) \quad (2.6.4.5)$$

- the $100k\%$ right-sided confidence interval for σ_1^2/σ_2^2 given by

$$\left(\frac{s_1^2}{s_2^2} \cdot (F_{1-k, n_1-1, n_2-1})^{-1}, \infty \right) \quad (2.6.4.6)$$

To remember these six confidence interval formulas more efficiently, consider using the following: with the test statistic formula, i.e. Equations 2.3.5.1 or 2.3.5.3, replace h with the critical value(s), where $\alpha = 1 - k$.

To illustrate, suppose that we want to find the 95% right-sided confidence interval for σ_1^2 / σ_2^2 , where $n_1 = 9$ and $n_2 = 7$. Note that the corresponding right-tailed test at the $\alpha = 1 - 0.95 = 0.05$ significance level has a critical value of

$$F_{0.05, 9-1, 7-1} = 4.147$$

Therefore, a bound of the desired confidence interval using Equation 2.3.5.3 is

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{h} \Rightarrow \frac{s_1^2}{s_2^2} \cdot \frac{1}{4.147}$$

We know this is the lower bound because we are constructing a right-sided confidence interval.

Example 2.6.4.1

An astronomer studies how frequently a star is visible in a year. Assume that the number of appearances in a year is normally distributed. She takes a random sample of 8 years from a database and obtains the following data:

4 5 5 4 3 2 3 5

Construct the 95% confidence interval for the variance of the number of appearances in a year.

Solution

Recall that $(n - 1)s^2$ is the numerator of s^2 . As a result,

$$\bar{x} = \frac{4 + 5 + \dots + 5}{8} = 3.875$$

$$(n - 1)s^2 = (4 - 3.875)^2 + (5 - 3.875)^2 + \dots + (5 - 3.875)^2 = 8.875$$

From the exam table, obtain

$$\chi^2_{(1+0.95)/2, 8-1} = \chi^2_{0.975, 7} = 16.01$$

$$\chi^2_{(1-0.95)/2, 8-1} = \chi^2_{0.025, 7} = 1.69$$

This produces the answer of

$$\left(\frac{8.875}{16.01}, \frac{8.875}{1.69} \right) = (0.554, 5.251)$$



2.6 Summary

Notation

Symbol	Concept
k	Confidence level
z_q	100 q^{th} percentile of the standard normal distribution
$t_{2(1-q), \text{df}}$	100 q^{th} percentile of a t -distribution
$\chi^2_{q, \text{df}}$	100 q^{th} percentile of a chi-square distribution
$F_{1-q, \text{ndf}, \text{ddf}}$	100 q^{th} percentile of an F -distribution

Key Ideas

- A confidence interval estimates a parameter with a range of values, suggesting the parameter's value lies within the interval based on a certain level of confidence.
- Under the same circumstance, H_0 will fail to be rejected at the α significance level if the hypothesized value h is within the 100 $(1 - \alpha)\%$ confidence interval for the parameter. Otherwise, H_0 will be rejected.

Intervals for Means

ONE SAMPLE

In the known σ^2 scenario,

- the 100 $k\%$ confidence interval for μ is

$$\bar{x} \pm z_{(1+k)/2} \frac{\sigma}{\sqrt{n}}$$

- the 100 $k\%$ left-sided confidence interval for μ is

$$\left(-\infty, \bar{x} + z_k \frac{\sigma}{\sqrt{n}}\right)$$

- the $100k\%$ right-sided confidence interval for μ is

$$\left(\bar{x} - z_k \frac{\sigma}{\sqrt{n}}, \infty\right)$$

In the unknown σ^2 scenario,

- the $100k\%$ confidence interval for μ is

$$\bar{x} \pm t_{1-k, n-1} \frac{s}{\sqrt{n}}$$

- the $100k\%$ left-sided confidence interval for μ is

$$\left(-\infty, \bar{x} + t_{2(1-k), n-1} \frac{s}{\sqrt{n}}\right)$$

- the $100k\%$ right-sided confidence interval for μ is

$$\left(\bar{x} - t_{2(1-k), n-1} \frac{s}{\sqrt{n}}, \infty\right)$$

TWO SAMPLES

In the known σ_1^2 and σ_2^2 scenario,

- the $100k\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 \pm z_{(1+k)/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- the $100k\%$ left-sided confidence interval for $\mu_1 - \mu_2$ is

$$\left(-\infty, \bar{x}_1 - \bar{x}_2 + z_k \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

- the 100k% right-sided confidence interval for $\mu_1 - \mu_2$ is

$$\left(\bar{x}_1 - \bar{x}_2 - z_k \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \infty \right)$$

In the unknown σ_1^2 and σ_2^2 scenario,

- the 100k% confidence interval for $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 \pm t_{1-k, n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- the 100k% left-sided confidence interval for $\mu_1 - \mu_2$ is

$$\left(-\infty, \bar{x}_1 - \bar{x}_2 + t_{2(1-k), n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

- the 100k% right-sided confidence interval for $\mu_1 - \mu_2$ is

$$\left(\bar{x}_1 - \bar{x}_2 - t_{2(1-k), n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \infty \right)$$

PAIRED OBSERVATIONS

This is identical to the one-sample case with the following substitutions:

- $\bar{x} \rightarrow \bar{d}$

- $\sigma^2 \rightarrow \sigma_D^2$
- $n \rightarrow n_*$
- $s^2 \rightarrow s_D^2$

Intervals for Proportions

ONE SAMPLE

The $100k\%$ confidence interval for q is

$$\hat{q} \pm z_{(1+k)/2} \sqrt{\frac{\hat{q}(1-\hat{q})}{n}}$$

The $100k\%$ left-sided confidence interval for q is

$$\left(-\infty, \hat{q} + z_k \sqrt{\frac{\hat{q}(1-\hat{q})}{n}} \right)$$

The $100k\%$ right-sided confidence interval for q is

$$\left(\hat{q} - z_k \sqrt{\frac{\hat{q}(1-\hat{q})}{n}}, \infty \right)$$

TWO SAMPLES

The $100k\%$ confidence interval for $q_1 - q_2$ is

$$\hat{q}_1 - \hat{q}_2 \pm z_{(1+k)/2} \sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}$$

The $100k\%$ left-sided confidence interval for $q_1 - q_2$ is

$$\left(-\infty, \hat{q}_1 - \hat{q}_2 + z_k \sqrt{\frac{\hat{q}_1 (1 - \hat{q}_1)}{n_1} + \frac{\hat{q}_2 (1 - \hat{q}_2)}{n_2}} \right)$$

The $100k\%$ right-sided confidence interval for $q_1 - q_2$ is

$$\left(\hat{q}_1 - \hat{q}_2 - z_k \sqrt{\frac{\hat{q}_1 (1 - \hat{q}_1)}{n_1} + \frac{\hat{q}_2 (1 - \hat{q}_2)}{n_2}}, \infty \right)$$

Intervals for Variances

ONE SAMPLE

The $100k\%$ confidence interval for σ^2 is

$$\left(\frac{(n-1)s^2}{\chi_{(1+k)/2, n-1}^2}, \frac{(n-1)s^2}{\chi_{(1-k)/2, n-1}^2} \right)$$

The $100k\%$ left-sided confidence interval for σ^2 is

$$\left(0, \frac{(n-1)s^2}{\chi_{1-k, n-1}^2} \right)$$

The $100k\%$ right-sided confidence interval for σ^2 is

$$\left(\frac{(n-1)s^2}{\chi_{k, n-1}^2}, \infty \right)$$

TWO SAMPLES

The $100k\%$ confidence interval for σ_1^2 / σ_2^2 is

$$\left(\frac{s_1^2}{s_2^2} \cdot (F_{(1-k)/2, n_1-1, n_2-1})^{-1}, \frac{s_1^2}{s_2^2} \cdot F_{(1-k)/2, n_2-1, n_1-1} \right)$$

The $100k\%$ left-sided confidence interval for σ_1^2 / σ_2^2 is

$$\left(0, \frac{s_1^2}{s_2^2} \cdot F_{1-k, n_2-1, n_1-1} \right)$$

The $100k\%$ right-sided confidence interval for σ_1^2 / σ_2^2 is

$$\left(\frac{s_1^2}{s_2^2} \cdot (F_{1-k, n_1-1, n_2-1})^{-1}, \infty \right)$$

Appendix

🕒 5m

Confidence Interval and Hypothesis Test Connection

This illustrates – in the one-sample, known variance setup – how the confidence interval and the two-tailed hypothesis test for a mean are related. Other setups and types of parameters follow the same logic.

Recall that the critical region is

$$\left| \frac{\bar{x} - h}{\sigma/\sqrt{n}} \right| \geq z_{1-\frac{\alpha}{2}}$$

Therefore,

$$\begin{aligned} \Pr\left(\left| \frac{\bar{X} - h}{\sigma/\sqrt{n}} \right| \geq z_{1-\frac{\alpha}{2}}\right) &= \alpha \\ \Rightarrow \Pr\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - h}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right) &= 1 - \alpha \end{aligned}$$

As a result, the inequality

$$-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{x} - h}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}$$

leads to not rejecting H_0 . On the other hand, the last probability statement is what we should manipulate to determine the confidence interval bounds. First, let $1 - \alpha = k$. This leads to

$$z_{1-\frac{\alpha}{2}} = z_{(2-\alpha)/2} = z_{(1+k)/2}$$

Thus, rearranging the inequality in the probability statement produces

$$\begin{aligned}
 -z_{(1+k)/2} &\leq \frac{\bar{X} - h}{\sigma/\sqrt{n}} \leq z_{(1+k)/2} \\
 -z_{(1+k)/2} \frac{\sigma}{\sqrt{n}} &\leq \bar{X} - h \leq z_{(1+k)/2} \frac{\sigma}{\sqrt{n}} \\
 -z_{(1+k)/2} \frac{\sigma}{\sqrt{n}} &\leq h - \bar{X} \leq z_{(1+k)/2} \frac{\sigma}{\sqrt{n}} \\
 \bar{X} - z_{(1+k)/2} \frac{\sigma}{\sqrt{n}} &\leq h \leq \bar{X} + z_{(1+k)/2} \frac{\sigma}{\sqrt{n}}
 \end{aligned}$$

This shows that the $100(1 - \alpha)\%$ confidence interval for the mean is given by

$$\bar{x} \pm z_{(1+k)/2} \frac{\sigma}{\sqrt{n}}$$

and, when h is within the bounds of the interval, it corresponds to the test statistic being outside the critical region at the α significance level.

2.7.0 Overview

 5m

Order statistics were briefly introduced in Section 1.4.3. We now formally develop the subject from first principles. We also address several special cases and demonstrate how the concepts tie back to other topics we already discussed.

2.7.1 First Principles

The ***first order statistic*** is the smallest observation of the sample X_1, \dots, X_n , i.e.

$$X_{(1)} = \min(X_1, \dots, X_n) \quad (2.7.1.1)$$

The ***nth order statistic*** is the largest observation of the sample X_1, \dots, X_n , i.e.

$$X_{(n)} = \max(X_1, \dots, X_n) \quad (2.7.1.2)$$

In general, the ***kth order statistic*** is the k^{th} observation in ascending order, denoted by $X_{(k)}$, for $k = 1, \dots, n$.

Realize that these are mere statistics of the sample; Example 2.2.4.5 covers this idea. Therefore, we are interested in how they are distributed.

Assume X_1, \dots, X_n is a random sample. Then, we can derive the probability function of $X_{(k)}$ using the distribution from which the sample is taken, i.e.

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)! (n-k)!} \cdot [F_X(x)]^{k-1} \cdot f_X(x) \cdot [S_X(x)]^{n-k} \quad (2.7.1.3)$$

This formula is applicable to both discrete and continuous distributions. Intuitively, the probability function can be viewed as: for the k^{th} smallest of n observations, we need $k-1$ observations to be at most x , one observation at x , and $n-k$ observations to be greater than x .

On the other hand, it is often simpler to calculate probabilities by reasoning things through rather than using this probability function. For example, consider the probability that the smallest observation is greater than x . This event only occurs when all n sample observations are greater than x , so

$$\Pr(X_{(1)} > x) = [S_X(x)]^n \quad (2.7.1.4)$$

Now consider the probability that the largest observation is at most x . This event only occurs when all n sample observations are at most x , so

$$\Pr(X_{(n)} \leq x) = [F_X(x)]^n \quad (2.7.1.5)$$

These probabilities are inherently tied to a more general approach for the k^{th} order statistic that uses the binomial distribution. Example 2.7.1.2 introduces the general approach.

Example 2.7.1.1

A random sample of size 4 is drawn from an inverse exponential distribution with $\theta = 0.1$.

Calculate the 30th percentile of the largest of the sample.

Solution

We want to solve for $\pi_{0.3}$, where

$$\Pr(X_{(4)} \leq \pi_{0.3}) = 0.3$$

Thus, determine the CDF of $X_{(4)}$. Use Equation 2.7.1.5, and look up the exam table for the CDF of an inverse exponential distribution.

$$\begin{aligned}\Pr(X_{(4)} \leq x) &= [F_X(x)]^4 \\ &= [e^{-0.1/x}]^4 \\ &= e^{-0.4/x}\end{aligned}$$

Then, solve for $\pi_{0.3}$.

$$\begin{aligned}\Pr(X_{(4)} \leq \pi_{0.3}) &= 0.3 \\ e^{-0.4/\pi_{0.3}} &= 0.3 \\ -\frac{0.4}{\pi_{0.3}} &= \ln 0.3 \\ \pi_{0.3} &= -\frac{0.4}{\ln 0.3} \\ &= \mathbf{0.332}\end{aligned}$$



Example 2.7.1.2

A random sample of size 4 is drawn from an inverse exponential distribution with $\theta = 0.1$.

Calculate the probability that the second smallest of the sample is less than 0.25.

Solution

The goal is to calculate

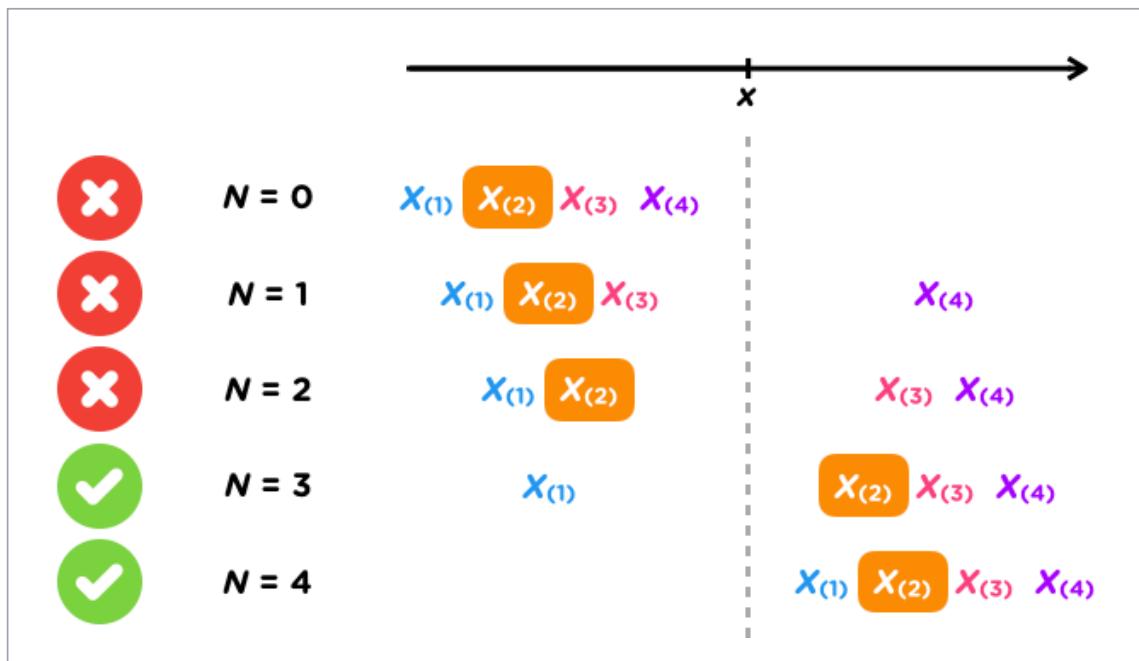
$$\Pr(X_{(2)} < 0.25)$$

When dealing with probabilities of an order statistic other than the first or n^{th} , use the binomial approach. First, note that Equations 2.7.1.4 and 2.7.1.5 give the survival function of $X_{(1)}$ and the CDF of $X_{(n)}$. This suggests that it is easier to find the survival function of a **lower k^{th}** order statistic than its CDF, but the reverse is true for a **higher k^{th}** order statistic. Since $k = 2$, let's derive the survival function of $X_{(2)}$.

To focus on survival functions, let N be the number of observations that are greater than x . Realize that

$$N \sim \text{Binomial} \left(m = 4, q = S_X(x) = 1 - e^{-0.1/x} \right)$$

The second smallest observation will be greater than x when at least 3 of the 4 observations are greater than x .



Therefore,

$$\begin{aligned}
\Pr(X_{(2)} > x) &= \Pr(N = 3) + \Pr(N = 4) \\
&= \binom{4}{3} \left(1 - e^{-0.1/x}\right)^3 \left(e^{-0.1/x}\right) + \left(1 - e^{-0.1/x}\right)^4 \\
&= \left(1 - e^{-0.1/x}\right)^3 \left(4e^{-0.1/x} + 1 - e^{-0.1/x}\right) \\
&= \left(1 - e^{-0.1/x}\right)^3 \left(3e^{-0.1/x} + 1\right)
\end{aligned}$$

Solve for the answer as

$$\begin{aligned}
\Pr(X_{(2)} < 0.25) &= 1 - \Pr(X_{(2)} > 0.25) \\
&= 1 - \left(1 - e^{-0.1/0.25}\right)^3 \left(3e^{-0.1/0.25} + 1\right) \\
&= \mathbf{0.892}
\end{aligned}$$



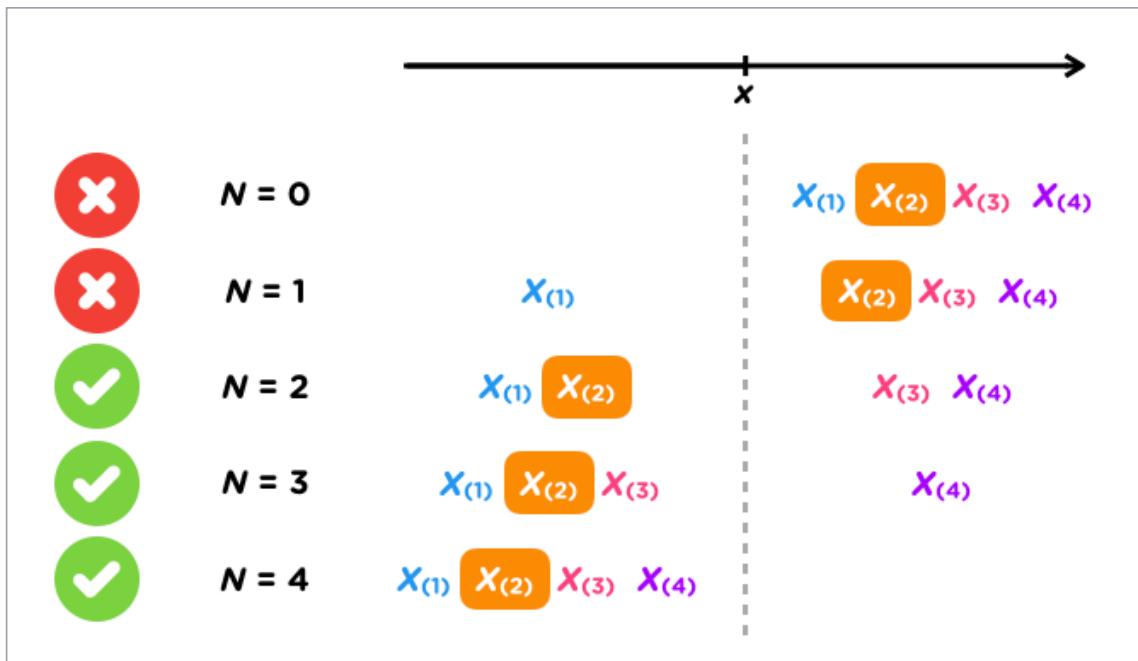
Alternative Solution

In the previous solution, note that it is because we wished to determine the **survival function** of $X_{(2)}$ that we defined N as the number of observations that are **greater than x** .

Now, let's solve for the CDF of $X_{(2)}$ directly. Let N be the number of observations that are **at most x** instead, so

$$N \sim \text{Binomial} \left(m = 4, q = F_X(x) = e^{-0.1/x} \right)$$

The second smallest observation will be at most x when at least 2 of the 4 observations are at most x .



Therefore,

$$\begin{aligned}\Pr(X_{(2)} \leq x) &= \Pr(N = 2) + \Pr(N = 3) + \Pr(N = 4) \\ &= 6\left(e^{-0.1/x}\right)^2\left(1 - e^{-0.1/x}\right)^2 + 4\left(e^{-0.1/x}\right)^3\left(1 - e^{-0.1/x}\right) + \left(e^{-0.1/x}\right)^4\end{aligned}$$

As a result,

$$\begin{aligned}\Pr(X_{(2)} < 0.25) &= 6\left(e^{-0.1/0.25}\right)^2\left(1 - e^{-0.1/0.25}\right)^2 + 4\left(e^{-0.1/0.25}\right)^3\left(1 - e^{-0.1/0.25}\right) + \left(e^{-0.1/0.25}\right)^4 \\ &= 0.892\end{aligned}$$

In the previous solution, we mentioned that finding the survival function of $X_{(2)}$ is easier than finding its CDF; notice that we need to sum more binomial probabilities to obtain the CDF. But either way, we arrive at the same answer because we simply swap what a binomial "success" represents ("greater than x " versus "at most x "), so there is no genuine difference.



Example 2.7.1.3

Let X_1, X_2 , and X_3 be the annual claims of three newly written policies drawn from a random sample with PDF:

$$f(x) = \frac{5}{x^6}, \quad x > 1$$

Calculate the expected value of the sample median of the three claims.

Solution

For an odd-numbered n , the sample median is unambiguously $X_{(m)}$, where m is the median of the integers 1 to n . For an even-numbered n , the definition of the sample median is ambiguous; we expect exam problems will avoid this ambiguity.

Hence, the goal is to calculate $E[X_{(2)}]$. Let's first solve for the PDF of $X_{(2)}$. As seen in Equation 2.7.1.3, we require the PDF, CDF, and survival function of the distribution from which the sample is taken.

From the given PDF, realize that the sample is taken from a single-parameter Pareto with $\alpha = 5$ and $\theta = 1$. Use the exam table to determine the single-parameter Pareto's CDF and survival function. As a result,

$$\begin{aligned} f_{X_{(2)}}(x) &= \frac{3!}{(2-1)!(3-2)!} \cdot \left[1 - \left(\frac{1}{x}\right)^5\right]^{2-1} \cdot \frac{5}{x^6} \cdot \left[\left(\frac{1}{x}\right)^5\right]^{3-2} \\ &= 6 \left(\frac{x^5 - 1}{x^5}\right) \left(\frac{5}{x^6}\right) \left(\frac{1}{x^5}\right) \\ &= 30 \left(\frac{1}{x^{11}} - \frac{1}{x^{16}}\right), \quad x > 1 \end{aligned}$$

Solve for the answer as

$$\begin{aligned} E[X_{(2)}] &= \int_{-\infty}^{\infty} x \cdot f_{X_{(2)}}(x) dx \\ &= 30 \int_1^{\infty} \left(\frac{1}{x^{10}} - \frac{1}{x^{15}}\right) dx \\ &= 30 \left[-\frac{1}{9x^9} + \frac{1}{14x^{14}}\right]_1^{\infty} \\ &= 30 \left[0 - \left(-\frac{1}{9} + \frac{1}{14}\right)\right] \\ &= \mathbf{1.19} \end{aligned}$$



Alternative Solution

Recall that an alternative to calculating an expected value is to integrate the survival function, i.e. Equation 1.0.4.2. Let's solve for the survival function of $X_{(2)}$ using the binomial approach. Define N as the number of observations that are greater than x , i.e.

$$N \sim \text{Binomial} \left(3, S(x) = \frac{1}{x^5} \right)$$

The sample median will be greater than x when at least 2 of the 3 observations are greater than x . Therefore,

$$\begin{aligned} S_{X_{(2)}}(x) &= \Pr(N = 2) + \Pr(N = 3) \\ &= 3 \left(\frac{1}{x^5} \right)^2 \left(1 - \frac{1}{x^5} \right) + \left(\frac{1}{x^5} \right)^3 \\ &= \frac{3}{x^{10}} - \frac{3}{x^{15}} + \frac{1}{x^{15}} \\ &= \frac{3}{x^{10}} - \frac{2}{x^{15}}, \quad x > 1 \end{aligned}$$

It is important to note that $S_{X_{(2)}}(x) = \Pr(X_{(2)} > x) = 1$ when $x \leq 1$. This is because x would be less than the smallest possible value of the distribution, meaning the interval $X_{(2)} > x$ covers the entire domain.

Finally, calculate the answer using Equation 1.0.4.2.

$$\begin{aligned} \mathbb{E}[X_{(2)}] &= \int_0^\infty S_{X_{(2)}}(x) dx \\ &= \int_0^1 1 dx + \int_1^\infty \left(\frac{3}{x^{10}} - \frac{2}{x^{15}} \right) dx \\ &= [x]_0^1 + \left[-\frac{3}{9x^9} + \frac{2}{14x^{14}} \right]_1^\infty \\ &= [1 - 0] + \left[0 - \left(-\frac{3}{9} + \frac{2}{14} \right) \right] \\ &= \mathbf{1.19} \end{aligned}$$



2.7.2 Special Cases

Uniform Sample

Assume the random sample is drawn from a uniform distribution on the interval $[a, b]$. Then, recall from Section 1.4.3 that

$$\mathbb{E}[X_{(k)}] = a + \frac{k(b-a)}{n+1} \quad (1.4.3.1)$$

If we further assume that $a = 0$ and $b = \theta$, note that

$$\begin{aligned} f_{X_{(k)}}(x) &= \frac{n!}{(k-1)!(n-k)!} \cdot \left(\frac{x}{\theta}\right)^{k-1} \cdot \frac{1}{\theta} \cdot \left(1 - \frac{x}{\theta}\right)^{n-k} \\ &= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} \left(\frac{x}{\theta}\right)^k \left(1 - \frac{x}{\theta}\right)^{n-k} \frac{1}{x}, \quad 0 \leq x \leq \theta \end{aligned}$$

meaning

$$X_{(k)} \sim \text{Beta}(k, n-k+1, \theta)$$

Coach's Remarks

With $a = 0$ and $b = \theta$, realize that $X_{(n)}$ is an asymptotically unbiased estimator of θ . Moreover, recall that $X_{(n)}$ is

- a sufficient statistic for θ (from Example 2.2.4.5), and
- the maximum likelihood estimator of θ (from Section 2.1.5).

Exponential Sample

Assume the random sample is drawn from an exponential distribution with mean θ . Then,

$$\mathbb{E}[X_{(k)}] = \theta \sum_{i=n-k+1}^n \frac{1}{i} \quad (2.7.2.1)$$

k-out-of-n System

To avoid confusion with the k in " k^{th} order statistic", let K be the required minimum number of functioning components in a K -out-of- n system.

More precisely, let X_1, \dots, X_n be the lifetimes of n components in a system. Then, the system lifetime for a K -out-of- n system is $X_{(n-K+1)}$. This means the two topics are connected by $k = n - K + 1$. A simple way to remember this is to notice that they represent consecutive integers moving in opposite directions.

K	k
1	n
2	$n - 1$
\vdots	\vdots
n	1

Assuming that the component lifetimes are i.i.d., then a K -out-of- n system problem can be solved using order statistic concepts already discussed.

Coach's Remarks

Notice that Equation 2.7.2.1 is the same formula shown in the Coach's Remarks that concludes Example 1.5.7.2.

$$\mathbb{E}[X_{(n-K+1)}] = \theta \sum_{i=K}^n \frac{1}{i}$$

Example 2.7.2.1

A random sample of size 8 is drawn from a uniform distribution valid on $[0, \theta]$ to test the hypotheses

$$H_0 : \theta = 2 \quad H_1 : \theta > 2$$

Determine the critical region for the uniformly most powerful test of size 0.05.

Solution

This UMP test has a critical region based on the Neyman-Pearson theorem. Let h_1 be a constant that satisfies a θ value in H_1 , i.e. h_1 is some constant greater than 2.

In Example 2.2.4.5, we constructed the joint PDF carefully using the indicator function $I(\cdot)$, $x_{(1)} = \min(x_1, \dots, x_n)$, and $x_{(n)} = \max(x_1, \dots, x_n)$. We do the same to express the likelihood function precisely.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^8 \frac{I(0 \leq x_i \leq \theta)}{\theta} \\ &= \frac{I(x_{(1)} \geq 0) \cdot I(x_{(8)} \leq \theta)}{\theta^8} \\ &= \begin{cases} \frac{1}{\theta^8}, & x_{(1)} \geq 0, \quad x_{(8)} \leq \theta \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

Thus,

$$L(2) = \begin{cases} \frac{1}{256}, & x_{(1)} \geq 0, \quad x_{(8)} \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

$$L(h_1) = \begin{cases} \frac{1}{h_1^8}, & x_{(1)} \geq 0, \quad x_{(8)} \leq h_1 \\ 0, & \text{otherwise} \end{cases}$$

Let $\Lambda = L(2) \div L(h_1)$. Similar to Example 2.4.1.4, calculating Λ requires using the likelihood value that corresponds to the proper range. Keep in mind that $h_1 > 2$. As a result,

$$\begin{aligned}\Lambda &= \begin{cases} \frac{1 \div 256}{1 \div h_1^8}, & x_{(1)} \geq 0, \quad x_{(8)} \leq 2 \\ 0, & x_{(1)} \geq 0, \quad 2 < x_{(8)} \leq h_1 \end{cases} \\ &= \begin{cases} \frac{h_1^8}{256}, & x_{(1)} \geq 0, \quad x_{(8)} \leq 2 \\ 0, & x_{(1)} \geq 0, \quad 2 < x_{(8)} \leq h_1 \end{cases}\end{aligned}$$

Observe that Λ is smaller when $x_{(8)}$ is larger. Therefore, the critical region has the form $x_{(8)} \geq c$. Rejecting H_0 for large values of $x_{(8)}$ should make sense; for example, if $x_{(8)}$ is anything greater than 2, then it is impossible for H_0 to be true.

Given $\alpha = 0.05$, we want to solve for c . Use Equation 2.7.1.5 to determine the CDF of $X_{(8)}$.

$$\begin{aligned}\Pr(X_{(8)} \geq c \mid \theta = 2) &= 0.05 \\ \Pr(X_{(8)} \leq c \mid \theta = 2) &= 0.95 \\ \left[\frac{c}{2}\right]^8 &= 0.95 \\ c &= 2 \cdot 0.95^{1/8} \\ &= 1.987\end{aligned}$$

Therefore, the critical region for the UMP test of size 0.05 is

$$x_{(8)} \geq \underline{\mathbf{1.987}}$$



Example 2.7.2.2

A random sample of size 6 is drawn from an exponential distribution with mean 10. Calculate the probability that second largest of the sample is less than its expected value.

Solution

The second largest of the sample refers to $X_{(5)}$, so the goal is to calculate

$$\Pr(X_{(5)} < \mathbb{E}[X_{(5)}])$$

By Equation 2.7.2.1,

$$\begin{aligned}\mathbb{E}[X_{(5)}] &= 10 \sum_{i=6-5+1}^6 \frac{1}{i} \\ &= 10 \left(\frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{6} \right) \\ &= 14.5\end{aligned}$$

Next, derive the CDF of $X_{(5)}$. Let N be the number of observations that are at most x . Thus,

$$N \sim \text{Binomial} \left(6, 1 - e^{-x/10} \right)$$

The second largest observation will be at most x when at least 5 of the 6 observations are at most x . Therefore,

$$\begin{aligned}\Pr(X_{(5)} \leq x) &= \Pr(N = 5) + \Pr(N = 6) \\ &= 6 \left(1 - e^{-x/10} \right)^5 \left(e^{-x/10} \right) + \left(1 - e^{-x/10} \right)^6\end{aligned}$$

The answer is

$$\Pr(X_{(5)} < 14.5) = 6 \left(1 - e^{-14.5/10}\right)^5 \left(e^{-14.5/10}\right) + \left(1 - e^{-14.5/10}\right)^6$$

$$= \mathbf{0.57}$$



Example 2.7.2.3

You are given the following information about a one-out-of-four system:

- All components in the system are independent.
- The reliability for each component follows an inverse exponential distribution with $\theta = 0.1$.

Calculate the 30th percentile of the system's lifetime.

Solution

With $K = 1$, the system's lifetime is $X_{(4-1+1)} = X_{(4)}$. This is inherently the same problem as Example 2.7.1.1, now packaged in a reliability theory context. So, the answer is the same as before: the 30th percentile of the system's lifetime is **0.332**.

As a reminder, a 1-out-of- n system is also a parallel system.



Coach's Remarks

Likewise, if asked for the probability that a 3-out-of-4 system fails before 0.25 (with all else unchanged), we would solve it exactly as shown in Example 2.7.1.2. It should be intuitive how $K = 3$ aligns with the binomial approach in deriving the survival function of $X_{(2)}$.

2.7 Summary

$X_{(k)}$ is the k^{th} order statistic, i.e. the k^{th} observation in ascending order, for $k = 1, \dots, n$.

- $X_{(1)} = \min(X_1, \dots, X_n)$ is the first order statistic
- $X_{(n)} = \max(X_1, \dots, X_n)$ is the n^{th} order statistic
- For a random sample,

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)! (n-k)!} \cdot [F_X(x)]^{k-1} \cdot f_X(x) \cdot [S_X(x)]^{n-k}$$

$$\Pr(X_{(1)} > x) = [S_X(x)]^n$$

$$\Pr(X_{(n)} \leq x) = [F_X(x)]^n$$

- If the random sample is drawn from Uniform (a, b) , then

$$\mathbb{E}[X_{(k)}] = a + \frac{k(b-a)}{n+1}$$

- If the random sample is drawn from Uniform $(0, \theta)$, then

$$X_{(k)} \sim \text{Beta}(k, n-k+1, \theta)$$

- If the random sample is drawn from Exponential (θ) , then

$$\mathbb{E}[X_{(k)}] = \theta \sum_{i=n-k+1}^n \frac{1}{i}$$

- For a K -out-of- n system, the system lifetime is $X_{(n-K+1)}$.

