University of Applied Science - Online

Study-branch: data science

ADVANCED STATISTICS WORKBOOK

Dimitri MARCHAND

Matrikelnummer: 92131807

Realized with input of the Parameter Generator: ¡773c7a544ec570b813041c9425516a5fb8f5fa3e¿

Delivery date: December 24, 2024

Pledge

I hereby certify to have done this workbook by myself.

....
Liege, December 24, 2024 Dimitri MARCHAND

Chapter 1

Task 1: Basic Probabilities and Visualizations (1)

ξ values

- $\xi_1 = 1$
- $\xi_2 = 40$

1.1 problem statement

The number of meteorites falling into an ocean in a given year can be modeled by:

$$P(x) = \frac{e^{-\xi_2} \xi_2^x}{x!} \tag{1.1}$$

As seen in figure 1.1, the expected value of this distribution is ξ_2 and the median is ξ_2 too.

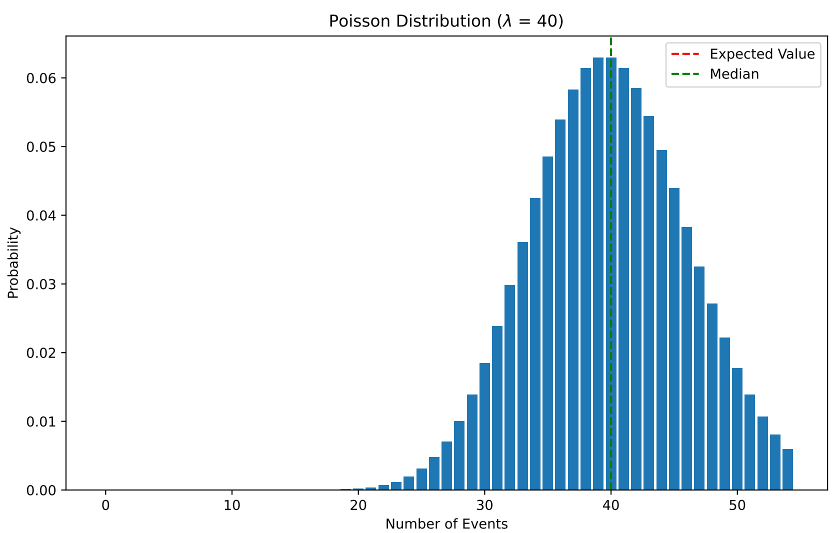

Figure 1.1: event distribution, with expected and median values

Chapter 2

Assignment 2: Basic Probabilities and Visualizations (2)

2.1 ξ values

- $\xi_4 = 1$
- $\xi_5 = \frac{2}{3}$
- $\xi_6 = 8$
- $\xi_7 = \frac{1}{3}$
- $\xi_8 = 3$

2.2 Problem statement

apparition time of an owl

Let Y be the random variable with the time it takes to hear an owl (in hours). The probability density function of Y is given by:

event probability

From $Y(t)$, we can derive the probability of the event 'hearing an owl', and more precisely its probability density function $P(t)$::

$$pdf_{event}(t) = \frac{d(1 - P(t))}{dt} \tag{2.1}$$

2.3 probability to wait 2-4 hours to see the event

this probability is given by substracting $Y(4)$ from $Y(2)$. The result is 0.0048.

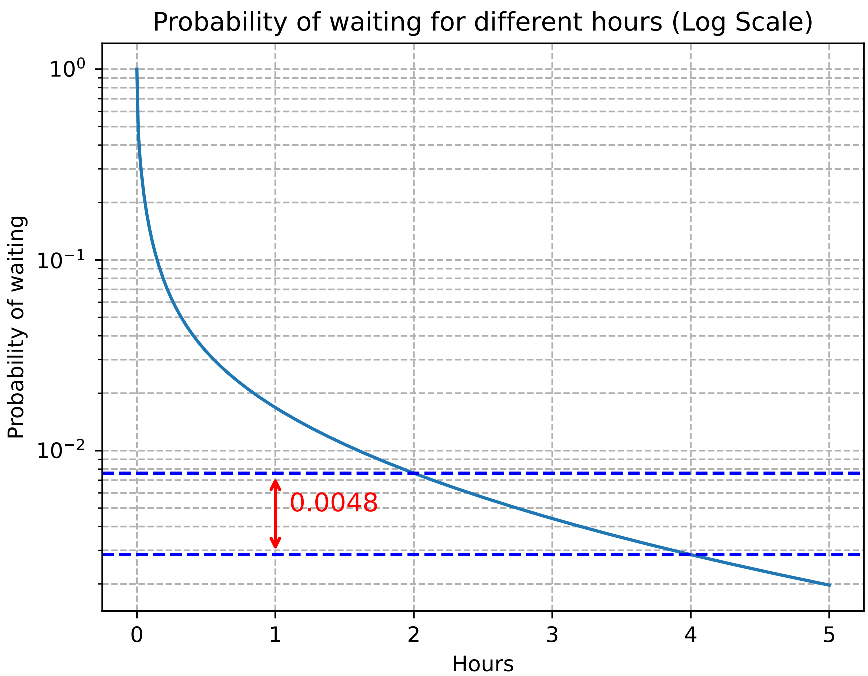

Figure 2.1: the probability to wait between 2 and 4 hours is 0.0048

2.4 event probability density graph

The graph of the probability density function of the event is given in figure 2.2.

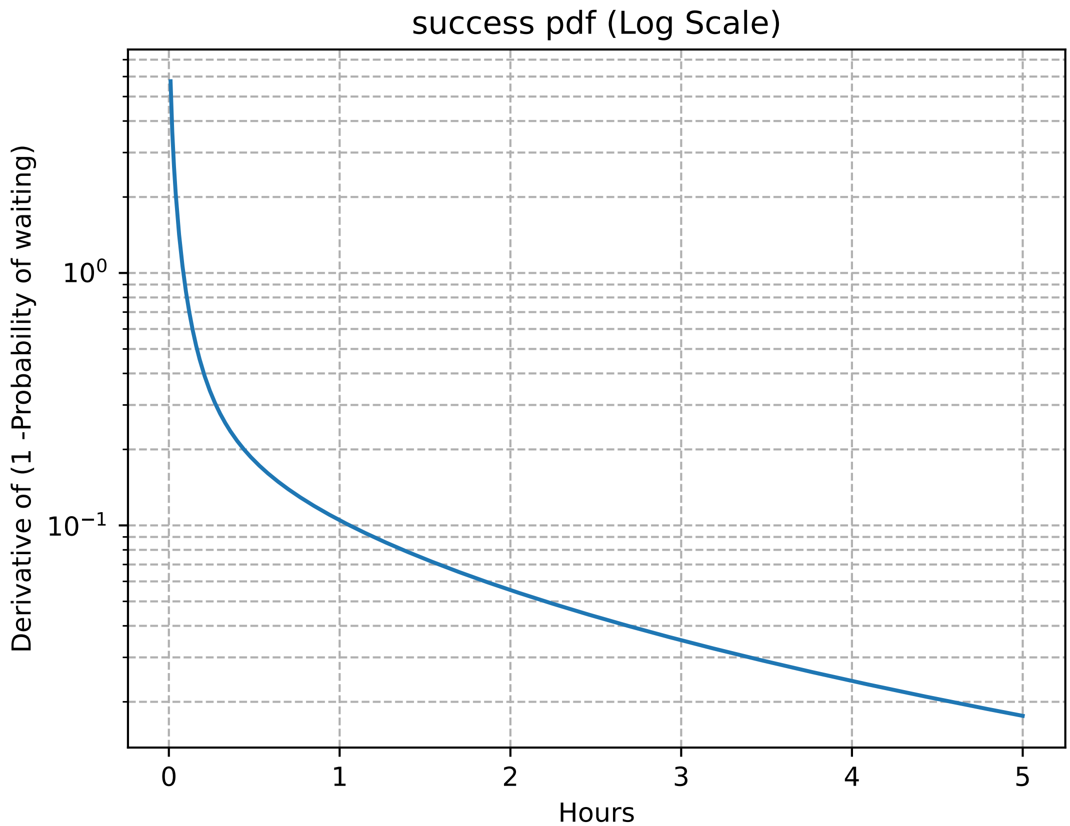

Figure 2.2: the probability density function of the event

2.5 distribution metrics

The metrics of the distribution are given in the table below. There is a large variance, which means that the distribution is not very regular.

metric	value	unit
mean	5.7	hours
variance	1167	hours*hours
q1	4.2	minutes
q2 (median)	30.5	minutes
q3	2.5	hours

Table 2.1: metrics of the distribution

2.6 distribution graph

The graph of the distribution is given in figure 2.3.

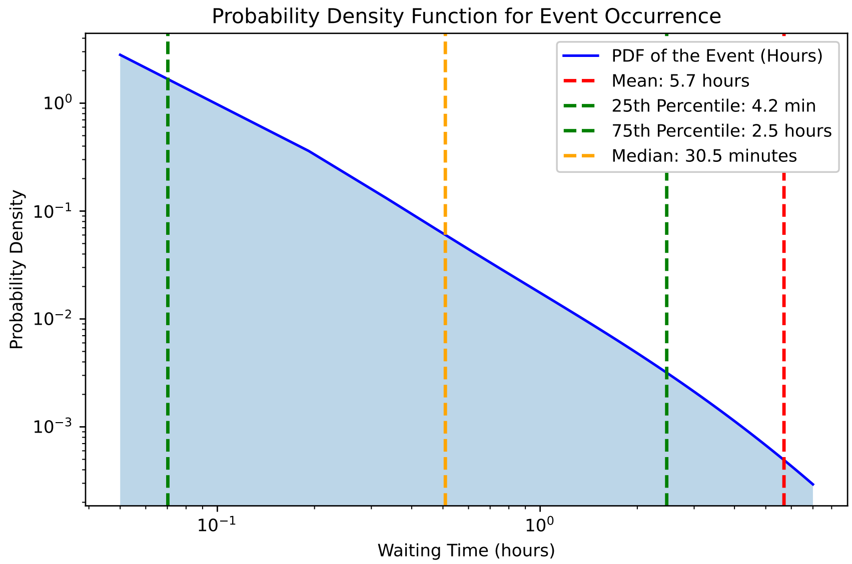

Figure 2.3: the distribution of the event

2.7 histogram of probability of hearing the owl at given minutes (around 3 hours)

The histogram of the probability of hearing the owl at given minutes (around 3 hours) is given in figure 2.4.

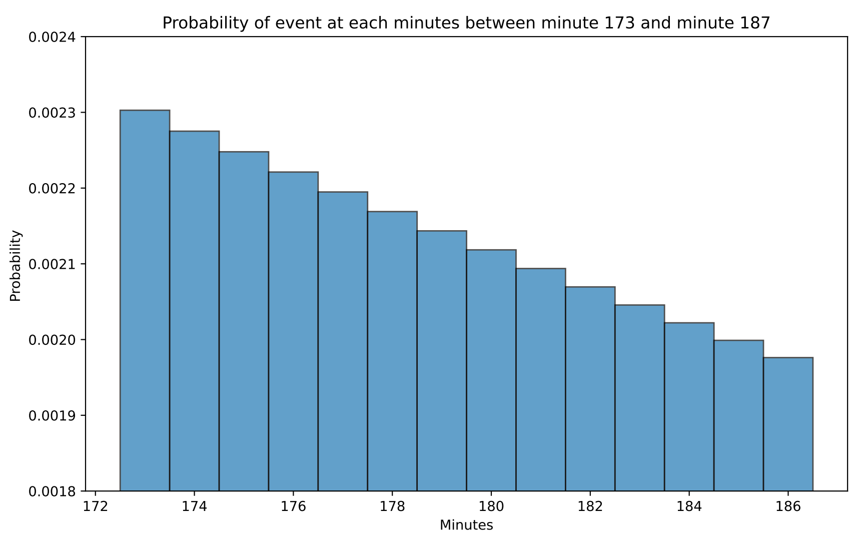

Figure 2.4: the histogram of the probability of hearing the owl at given minutes (around 3 hours)

Chapter 3

Assignment 3 : Transformed Random Variables

ξ values

- $\xi_9 = 0$
- $\xi_10 = 8, 7, 6, 17, 12$

3.1 Problem statement

The total bandwidth to failure S of a single router follows an exponential distribution with density:

$$f_S(s) = \frac{1}{\theta} \exp\left(-\frac{s}{\theta}\right), \quad s > 0,\, \theta > 0, \tag{3.1}$$

where θ is the mean failure bandwidth for a single router.

For a dual-router system, the total bandwidth to failure T can be expressed as:

$$T = S_1 + S_2 \tag{3.2}$$

where S_1 and S_2 are independent and identically distributed random variables representing the bandwidth totals to failure of each router.

3.2 Density Function of T

Given S_1 and S_2 both follow an exponential distribution of parameter $\frac{1}{\theta}$, the sum $T = S_1 + S_2$ follows a *Gamma distribution* with shape parameter $k = 2$ and rate $\lambda = \frac{1}{\theta}$. The probability density function of T with $\lambda = \frac{1}{\theta}$ is:

$$f_T(t) = \frac{t\lambda^2 e^{-\lambda t}}{1} = \frac{t}{\theta^2} \exp\left(-\frac{t}{\theta}\right) \tag{3.3}$$

3.3 Likelihood Function

Given an independent sample $T_1, T_2, \ldots, T_n$ of T, the likelihood function for the parameter θ is:

$$L(\theta) = \prod_{i=1}^{n} f_T(T_i) = \prod_{i=1}^{n} \frac{T_i}{\theta^2} \exp\left(-\frac{T_i}{\theta}\right) \tag{3.4}$$

Simplifying:

$$L(\theta) = \frac{1}{\theta^{2n}} \prod_{i=1}^{n} T_i \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} T_i\right) \tag{3.5}$$

3.4 Simplification of the Likelihood Function

To maximize the likelihood function, we simplify using the log-likelihood:

$$\ell(\theta) = \ln L(\theta) = -2n\ln\theta + \sum_{i=1}^{n} \ln T_i - \frac{1}{\theta}\sum_{i=1}^{n} T_i \tag{3.6}$$

The derivative of $\ell(\theta)$ with respect to θ is:

$$\frac{\partial \ell}{\partial \theta} = -\frac{2n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} T_i. \tag{3.7}$$

Setting this equal to zero gives:

$$\hat{\theta} = \frac{1}{2n}\sum_{i=1}^{n} T_i. \tag{3.8}$$

3.5 Estimation and Expectation for the Experiment

Given the sample $[8, 7, 6, 17, 12]$, I compute $\hat{\theta}$ and the expectation of T as:

$$E[T] = 2\hat{\theta} \tag{3.9}$$

from the given sample, the model with the maximum likelihood has $\theta = 5$. The expectation of T is 10.

Chapter 4

Assignment 4: Hypothesis Test

ξ values

- $\xi_{11} = 912$
- $\xi_{12} = 36.6$
- $\xi_{13} = 2$
- $\xi_{14} = [879, 842, 954, 842, 885, 918, 989, 768, 867, 1022]$

4.1 Problem statement

A statistical hypothesis testing is needed to determine if a new production system for hammers yields higher weights than the current system. The weights of hammers produced in the factory are normally distributed with a mean of 912 grams and a standard deviation of 36.6 grams. The analysis steps include setting up the hypotheses, performing the test, and making a decision based on the results.

Proposed Model for the Hammer Weights

The weights of hammers produced in the factory can be modeled using a normal distribution based on the observed long-term data:

$$W \sim \mathcal{N}(\mu, \sigma^2) \tag{4.1}$$

where:

- $\mu = \xi_{11}$ is the mean weight of the hammers.
- $\sigma = \xi_{12}$ is the standard deviation of the weights.

Assumptions

The following assumptions are made for this model to hold:

- The weights of the hammers are independent .
- mean and standard deviation are not varying in time.
- The underlying distribution of weights is approximately normal

Model Parameters

The parameters of the model are:

- μ: Mean of the distribution.
- σ^2: Variance of the distribution (or σ: standard deviation).

4.2 Hypothesis Testing

chosen statistical test and Decision Rule

I perform a one-sample one-tailed t-test

The test statistic is given by:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{4.2}$$

where:

- $\bar{x}$ is the sample mean,
- s is the sample standard deviation,
- n is the sample size,
- $\mu_0 = 912$ is the null hypothesis mean.

The decision rule is to reject H_0 if $t > t_c$, where t_c is determined from the t-distribution table at a chosen significance level α with $n - 1 = 9$ degrees of freedom.

To determine if the new production system yields higher weights, we set up the following hypotheses:

- H_0 and $\mu = 912 :$ the mean weight remains unchanged
- H_a and $\mu > 912 :$ the mean weight is higher under the new system

Error Probabilities

- Type I error (α): Rejecting H_0 when H_0 is true. Choose $\alpha = 0.05$.
- Type II error (β): Failing to reject H_0 when H_a is true. Can be estimated if the true mean under H_a is known.

test computation

The sample weights are:

$$[879, 842, 954, 842, 885, 918, 989, 768, 867, 1022] \tag{4.3}$$

Calculate the sample mean:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{879 + 842 + \cdots + 1022}{10} = 896.6 \tag{4.4}$$

Calculate the sample standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \approx 71.5 \tag{4.5}$$

Compute the test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{896.6 - 912}{71.5/\sqrt{10}} \approx -0.69 \tag{4.6}$$

Decision and Conclusion

For $\alpha = 0.05$ and $df = 9$, $t_c \approx 1.8$. Since $t = -0.69 < t_c$, I fail to reject H_0.

Conclusion: There is insufficient evidence to suggest that the new system produces hammers with higher weights.

Chapter 5

Assignment 5: Regularized Regression

5.1 problem statement

Given data points (x_i, y_i), $i = 1, \dots, n$, we aim to fit a polynomial model:

$$f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{10} x^{10} \tag{5.1}$$

where the parameters $\alpha = [\alpha_0, \alpha_1, \dots, \alpha_{10}]^\top$ are determined using Ordinary Least Squares (OLS) and ridge-regularized OLS.

5.2 Procedure

Step 1: Matrix Formulation

we define

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{10} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{10} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{10} \end{bmatrix} \tag{5.2}$$

Then the polynomial model can be written as:

$$\mathbf{y} = \mathbf{X}\alpha + \epsilon \tag{5.3}$$

where ϵ is the error term.

Step 2: OLS Estimate

The OLS estimate minimizes the sum of squared residuals:

$$\hat{\alpha}_{OLS} = \arg\min_{\alpha} \|\mathbf{y} - \mathbf{X}\alpha\|_{2^2} \tag{5.4}$$

The solution is given by:

$$\hat{\alpha}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{5.5}$$

Step 3: Ridge-Regularized OLS Estimate

Ridge regularization adds a penalty term to control the magnitude of α:

$$\hat{\alpha}_{Ridge} = \arg\min_{\alpha} \|\mathbf{y} - \mathbf{X}\alpha\|_2^2 + \lambda\|\alpha\|_2^2 \tag{5.6}$$

where $\lambda > 0$ is the regularization weight.

The closed-form solution is:

$$\hat{\alpha}_{Ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \tag{5.7}$$

where $\mathbf{I}$ is the identity matrix.

Step 4: Computation

1. Construct $\mathbf{X}$ from the input data by calculating x_i^k for $k = 0, \ldots, 10$.
2. Compute $\hat{\alpha}_{OLS}$ using the formula $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.
3. Select a penalty weight λ. A common practice is to test multiple values ($\lambda = 0.1, 1, 10, \ldots$) and evaluate the solutions.
4. Compute $\hat{\alpha}_{Ridge}$ using the formula $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$.

Weight for the Penalties (λ)

- Ridge regression shrinks the coefficients, reducing the risk of overfitting. The choice of λ is crucial:
- Small λ: Minimal regularization, closer to OLS.
- Large λ: Higher regularization, biasing coefficients toward zero.

Results

Qualities of Solutions

- OLS provides the best fit to the training data, but it may overfit if the model is too complex or if multicollinearity is present.
- Ridge regression balances the fit and model complexity, improving generalization by controlling the magnitude of α.

Conclusion

- Use OLS for cases with low multicollinearity and sufficient training data.
- Use Ridge regression when multicollinearity exists or to reduce overfitting for high-degree polynomial models.

Chapter 6

Assignment 5: Regularized Regression

ξ values

-

6.1 Problem statement

We want to make a bayesian modelling of a gamma distribution whose scale parameter is itself a random variable following a gamma distribution.

Part 1: Posterior Distribution of θ

The likelihood function for X given θ is:

$$f(x \mid \theta) = \frac{\beta^{\alpha} \cdot x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{6.1}$$

where $\beta = \frac{1}{\theta}$ Substituting $\beta = \frac{1}{\theta}$, we get:

$$f(x \mid \theta) = \frac{\left(\frac{1}{\theta}\right)^3 \cdot x^2 \cdot e^{-\frac{x}{\theta}}}{\Gamma(3)}. \tag{6.2}$$

The prior distribution for θ is:

$$\pi(\theta) = \frac{\Xi_{18}^{\Xi_{17}} \cdot \theta^{\Xi_{17}-1} \cdot e^{-\Xi_{18}\theta}}{\Gamma(\Xi_{17})}. \tag{6.3}$$

The posterior distribution $\pi(\theta \mid x)$ is proportional to the product of the likelihood and the prior:

$$\pi(\theta \mid x) \propto f(x \mid \theta) \cdot \pi(\theta). \tag{6.4}$$

Substituting the expressions for the likelihood and prior:

$$\pi(\theta \mid x) \propto \left(\frac{1}{\theta}\right)^3 \cdot x^2 \cdot e^{-\frac{x}{\theta}} \cdot \theta^{\Xi_{17}-1} \cdot e^{-\Xi_{18}\theta}. \tag{6.5}$$

Simplifying:

$$\pi(\theta \mid x) \propto \theta^{\Xi_{17}-4} \cdot e^{-\Xi_{18}\theta - \frac{x}{\theta}}. \tag{6.6}$$

The posterior distribution has the form of a Gamma distribution. For a Gamma distribution, the shape and rate parameters are updated as:

$$\tilde{\alpha} = \Xi_{17} + \alpha = \Xi_{17} + 3, \quad \tilde{\beta} = \Xi_{18} + x = \Xi_{18} + \Xi_{19}. \tag{6.7}$$

Thus, the posterior distribution is:

$$\theta \mid x \sim \Gamma(\tilde{\alpha} = \Xi_{17} + 3, \tilde{\beta} = \Xi_{18} + \Xi_{19}). \tag{6.8}$$

Part 2: Bayes Estimate with Square-Error Loss

The Bayes estimate under the square-error loss function is the **mean** of the posterior distribution.
For a Gamma distribution $\Gamma(\alpha, \beta)$, the mean is given by:

$$Mean = \frac{\alpha}{\beta}. \tag{6.9}$$

From Part (a), the posterior parameters are:

$$\tilde{\alpha} = \Xi_{17} + 3, \quad \tilde{\beta} = \Xi_{18} + \Xi_{19}. \tag{6.10}$$

Thus, the Bayes estimate is:

$$\hat{\theta}_{mean} = \frac{\tilde{\alpha}}{\tilde{\beta}} = \frac{\Xi_{17} + 3}{\Xi_{18} + \Xi_{19}}. \tag{6.11}$$

6.1.1 Part 3: Bayes Estimate Using the Mode

The Bayes estimate under the mode of the posterior distribution is the **mode** of the posterior Gamma distribution.
For a Gamma distribution $\Gamma(\alpha, \beta)$, the mode is given by:

$$Mode = \frac{\alpha - 1}{\beta} \tag{6.12}$$

Using the posterior parameters from Part (a):

$$\tilde{\alpha} = \Xi_{17} + 3, \quad \tilde{\beta} = \Xi_{18} + \Xi_{19}. \tag{6.13}$$

The mode is:

$$\hat{\theta}_{mode} = \frac{\tilde{\alpha} - 1}{\tilde{\beta}} = \frac{\Xi_{17} + 2}{\Xi_{18} + \Xi_{19}}. \tag{6.14}$$