Theoretical Foundations of the Analysis of Large Data Sets

Report 1

Magdalena Potok

Prepared on:
October 27, 2024

Excercise 1

Given a simple random variable $X_1, \ldots, X_n$ from a distribution with the density funcion $f(x, \alpha) = (\alpha+1)x^\alpha$ for $x \in (0, 1)$ and $\alpha > -1$, we observe that $X_1, \ldots, X_n \sim Beta(\alpha + 1, 1)$.
The **maximum likelihood estimator** is the value of the parameter that maximizes the **likelihood function**, here it is $L(\alpha) = \prod_{i=1}^n f(x_i, \alpha) = (\alpha + 1)^n \prod_{i=1}^n x_i^\alpha$. To simplify the optimization, we work with the **log-likelihood function** $l(\alpha) = nlog(\alpha + 1) + \alpha \sum_{i=1}^n logx_i$. To find the MLE, we take the derivative of the log-likehood function with respect to α and set it to zero: $\frac{dl}{d\alpha} = n\frac{1}{\alpha+1} + \sum_{i=1}^n logx_i = 0$. Solving this equation yields the MLE of α:

$$\hat{\alpha}_{MLE} = -1 - \frac{n}{\sum_{i=1}^n logX_i}.$$

To confirm that $\hat{\alpha}_{MLE}$ maximizes the likelihood, we check the second derivative of the log-likelihood function $\frac{d^2l}{d\alpha^2} = -\frac{n}{(\alpha+1)^2}$. Since this is negative is confirms, that calculated estimator is a maximum.
Fisher Information quantifies the amount of information that the observed data provide about the parameter α. It plays a key role in the asymptotic properties of the MLE, especially in determining the covariance matrix of the MLE. It is also essential in formulating test statistics, such as the Wald test. The Fisher information is computed as the negative expected value of the second derivative of the log-likelihood function, for this model, we find that:

$$I(\alpha) = -\mathbb{E}\big(\frac{\partial^2 logf(x, \alpha)}{\partial \alpha^2}\big) = \mathbb{E}\big(\frac{1}{(\alpha + 1)^2}\big) = \frac{1}{(\alpha + 1)^2}.$$

The **asymptotic distribution** of the MLE describes its behavior as the sample size n becomes large. The central limit theorem implies that the MLE is approximately normally distributed for large n $\sqrt{n}(\hat{\alpha_n} - \alpha_0) \xrightarrow{d} N(0, \frac{1}{I(\alpha)})$, so here it means that $\sqrt{n}(\hat{\alpha}_n - \alpha_0) \xrightarrow{d} N(0, (\alpha + 1)^2)$. In this case, the asymptotic distribution od $\hat{\alpha}_n$ is:

$$\hat{\alpha}_n \sim N(\alpha, \frac{(\alpha + 1)^2}{n}).$$

The **method of moments estimator** is based on equation the theoretical moments of the distribution to the sample moments. The first moment (mean) of the distribution $f(x, \alpha)$ is $\frac{\alpha+1}{\alpha+2}$. By solving for α in terms of the sample mean $\bar{X}$ we get the equation $\frac{\alpha+1}{\alpha+2} = \frac{1}{n}\sum_{i=1}^n X_i$, we derieve the MoM estimator as:

$$\hat{\alpha}_{mom} = \frac{2\bar{X} - 1}{1 - \bar{X}},$$

where $\bar{X}$ is the sample mean.

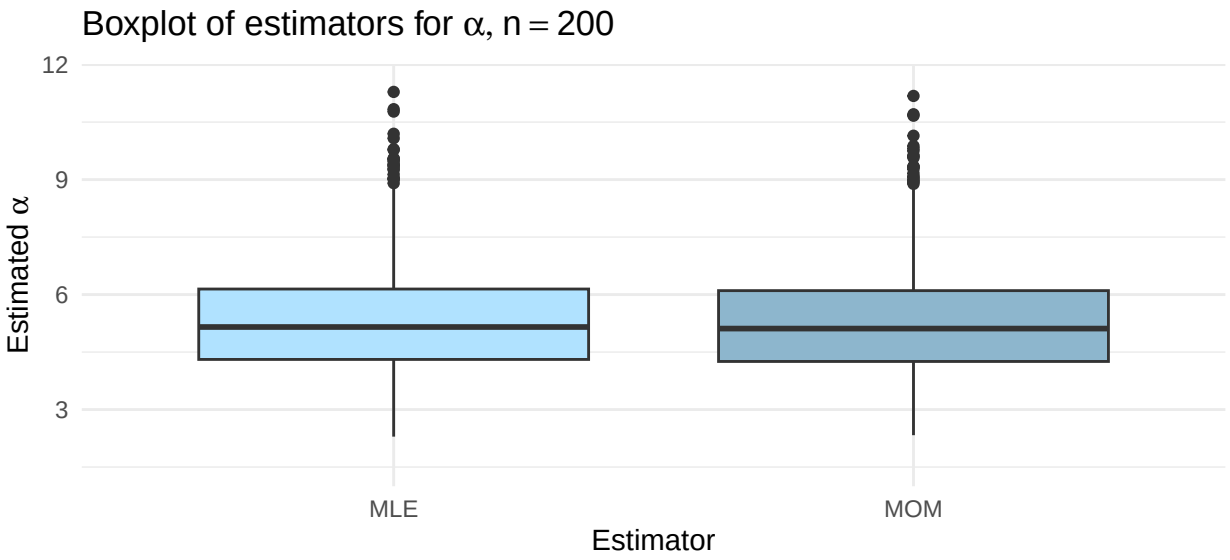
Now we fix $\alpha = 5$ and generate one random sample of the size $n = 20$ to calculate both estimators and the respective values of $\alpha - \hat{\alpha}$ and $(\alpha - \hat{\alpha})^2$.

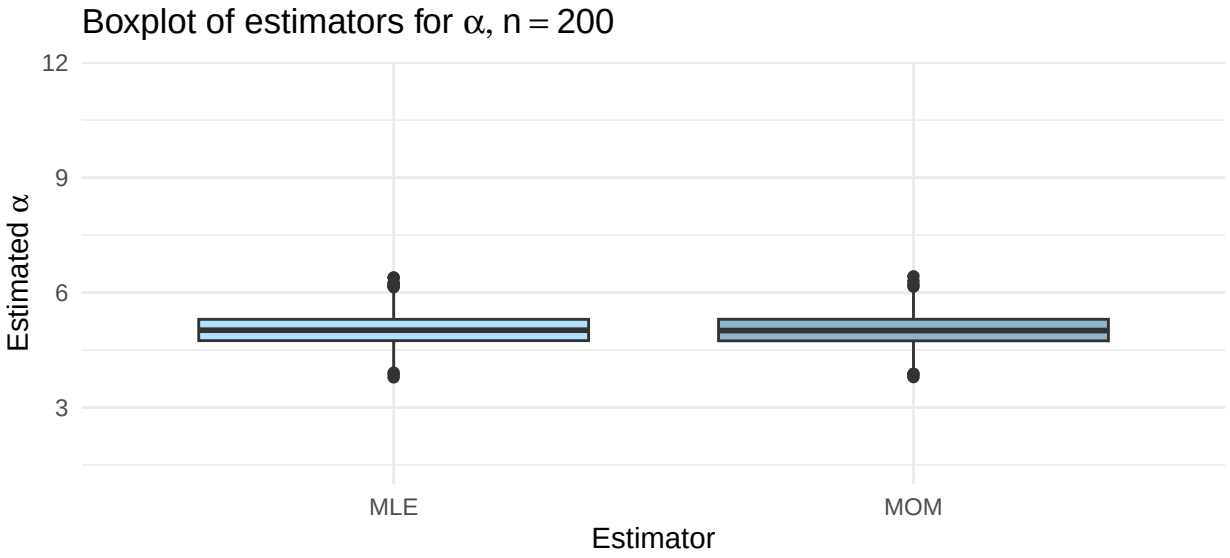
	$\hat{\alpha}_{MLE}$	$\hat{\alpha}_{MoM}$
$\alpha - \hat{\alpha}$	-0.951	-0.760
$(\alpha - \hat{\alpha})^2$	0.905	0.578

The method of moments estimator is more accurate because the difference between this estimator and the true value of α is smaller than the value of the maximum likelihood estimator.

The previous experiment was based on a single sample, but we know that using more samples provides more reliable results. To gain better insights, we will evaluate which estimator performs better by using 1000 samples for two different values of n (20 and 200), and then compare the results.

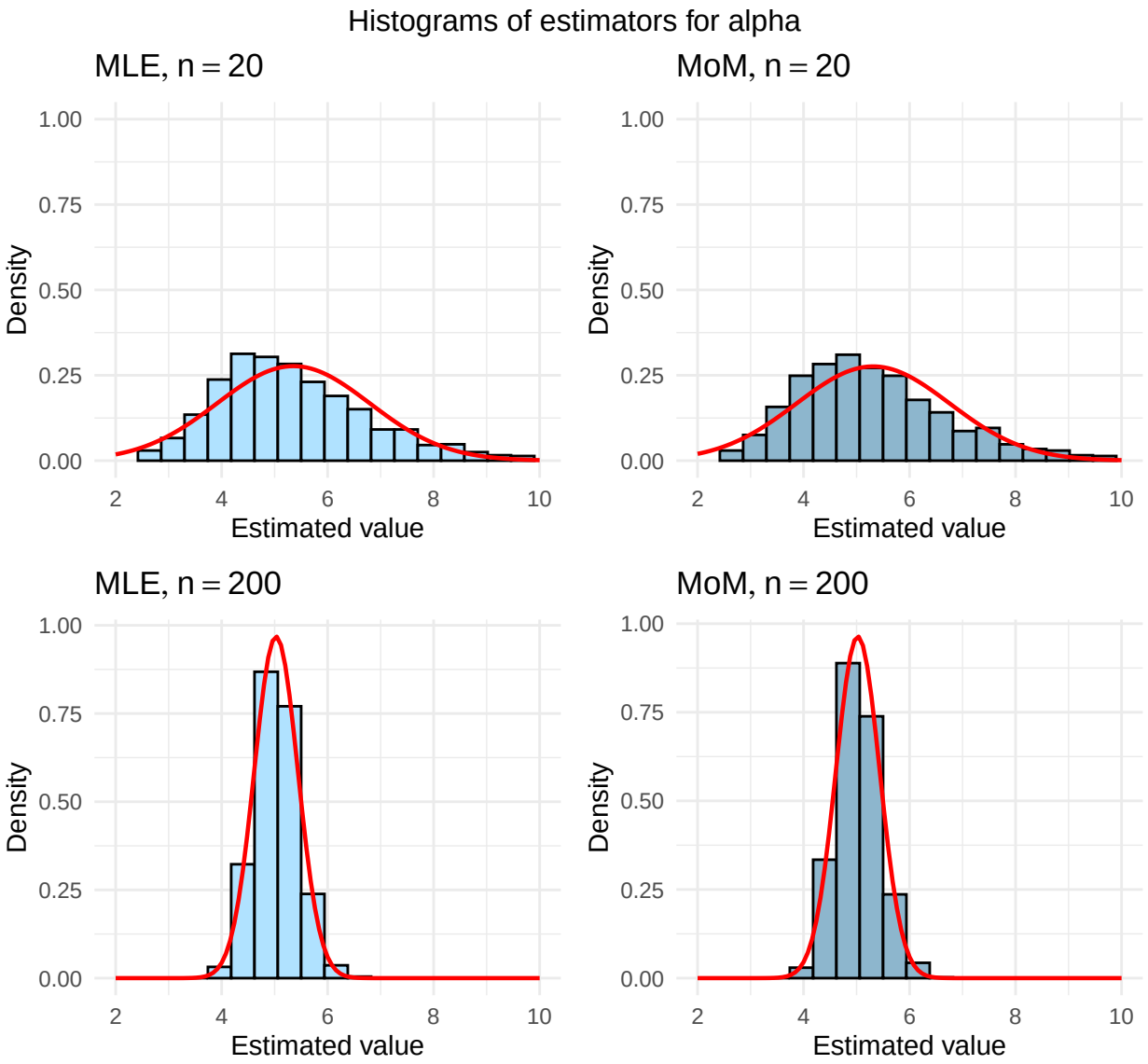
1. Comparing Boxplots





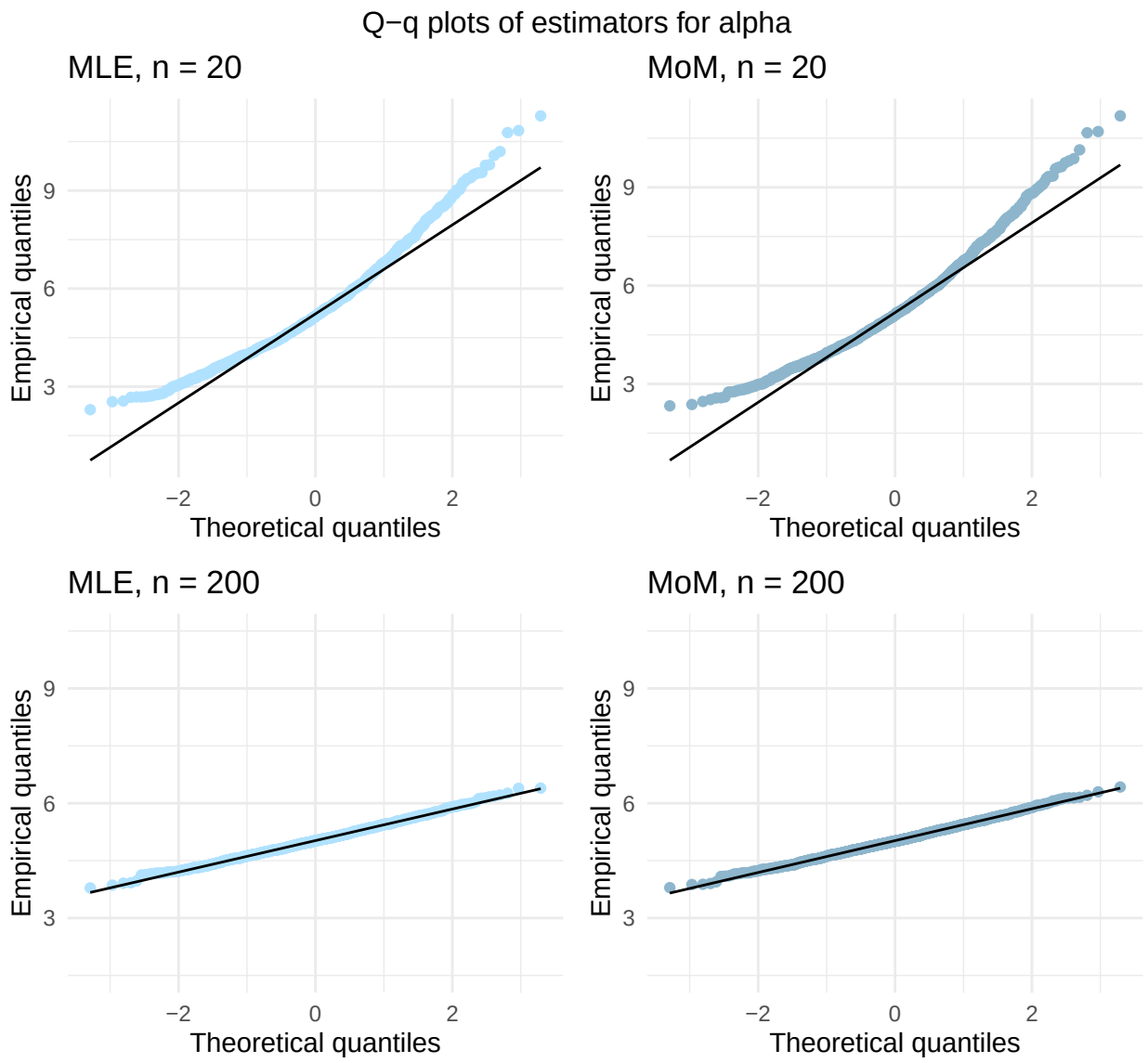
When sample size is $n = 20$ we can see that both the MLE and MoM estimators have the same median (around 5.5), but the MoM estimator has a little wider spread, which indicates higher variability. We can see that there are several outliers for both estimators, suggesting more extreme values are common. When sample size is $n = 200$ both estimators show much reduced variability compared to the situation, when $n = 20$. The median is closer to the real value of $\alpha = 5$. There are fewer outliers in larger sample, which is expected as the estimators stabilize with more data. The wishers are much shorter compared to $n = 20$, suggesting lower variability. For situation $n = 20$ and $n = 200$ boxplots for both estimators looks almost the same, and it is challenging to determine which one is better.

2. Comparing histograms



For $n = 20$, both MLE and MoM estimators show that most observations have a value close to 5, both of them are spread-out distributions, indicating high variability. The red line in each plot represents the normal distribution curve fitted based on the mean and standard deviation of the data, providing a visual reference for how closely the estimated values follow a normal distribution. In both cases for $n = 20$, the histograms show a right-skewed pattern, meaning there are more values to the left of the distribution's peak, with a longer tail to the right. For $n = 200$ MLE and MoM estimators give much tighter and more concentrated histograms around the true value 5. This reflects the asymptotic properties of these estimators: as the sample size increases, the estimators become more precise, with lower variance. We can see that more observations are close to 5 for MLE estimator, which suggest the MLE estimator is more efficient.

3. Comparing QQ-plots



When sample size is $n = 20$ for both estimators the plot deviates from the straight line in both tails, especially in upper tail. This indicates that these estimators have heavier tails than normal distribution when the sample size is small. For sample size $n = 200$ both Q-Q plots follows the straight line across the entire range, showing perfect agreement with the theoretical normal quantiles. This indicates that both of estimators is approximately normally distributed as the sample size increases.

4. Comparing estimated bias, variance and MSE

	n = 20			n = 200		
	Bias	Var	MSE	Bias	Var	MSE
$\hat{\alpha}_{MLE}$	0.355	2.075	2.201	0.027	0.170	0.170
$\hat{\alpha}_{MoM}$	0.312	2.089	2.187	0.023	0.171	0.172

For small sample size ($n = 20$) the MLE estimator has a slightly higher bias compared to MoM, which means that the MLE estimates are slightly more off from the true value on average. The variance for MLE is marginally lower than for MoM, indicating slightly less variability in the estimates for MLE. Both estimators ave almost identical MSE, though MoM has a slightly lower MSE, is combines both bias and variance, so this suggest that, overall, MoM performs very slightly better for $n = 20$. When the size of sample is larger $n = 200$ MSE for both estimators are almost the same, this suggest that with large samples MLE and MoM perform almost identically, but MLE very slightly performs better, because it has lower value of MSE.

Theoretical parameters provided by asymptotic distribution of MLE.

	Bias	Var	MSE
$\hat{\alpha}_{MLE}, n = 20$	0	1.80	1.80
$\hat{\alpha}_{MLE}, n = 200$	0	0.18	0.18

For $n = 20$ the theoretical variance, MSE and bias are lower that the empirical values, indicating that the small-sample performance of MLE does not match the asymptotic results. For large sample size theoretical values of parameters closely match the empirical values, suggesting that the MLE aligns well with its asymptotic properties when $n = 200$.

5. 95% Confidence intervals

	n = 20			n = 200		
	Var	Bias	MSE	Var	Bias	MSE
$\hat{\alpha}_{MLE}$	[1.904,2.270]	[0.265, 0.444]	[2.111, 2.290]	[1.917,2.285]	[0.222, 0.402]	[2.097, 2.276]
$\hat{\alpha}_{MoM}$	[0.156,0.186]	[0.001, 0.052]	[0.145, 0.196]	[0.157,0.187]	[-0.002, 0.049]	[0.146, 0.197]

In the table, we observe that for both sample sizes $n = 20$ and $n = 200$, the MLE estimator has much wider confidence intervals compared to the MoM estimator. This suggests greater variability and uncertainty in the MLE estimates. For $n = 20$, the MLE intervals are especially spread out, indicating that MLE is less reliable with smaller sample sizes. As the sample size increases to $n = 200$, both estimators show more precision, with narrower intervals, but MLE still exhibits more spread than MoM. The MoM estimator consistently provides tighter confidence intervals, reflecting higher stability and precision across both sample sizes. This highlights the better performance of MoM, particularly in smaller samples.

Excercise 2

Simple random sample $X_1, \ldots, X_n$ from the distribution with the density $f(x, \lambda) = \lambda e^{-\lambda x}, \quad x > 0, \lambda > 0$ is a random sample from the exponential distribution with parameter λ.
To find the uniformly most powerful test at level $\alpha = 0.05$ for testing the hypothesis

$$H_0 : \lambda = 5 \quad \text{against} \quad H_1 : \lambda = 3$$

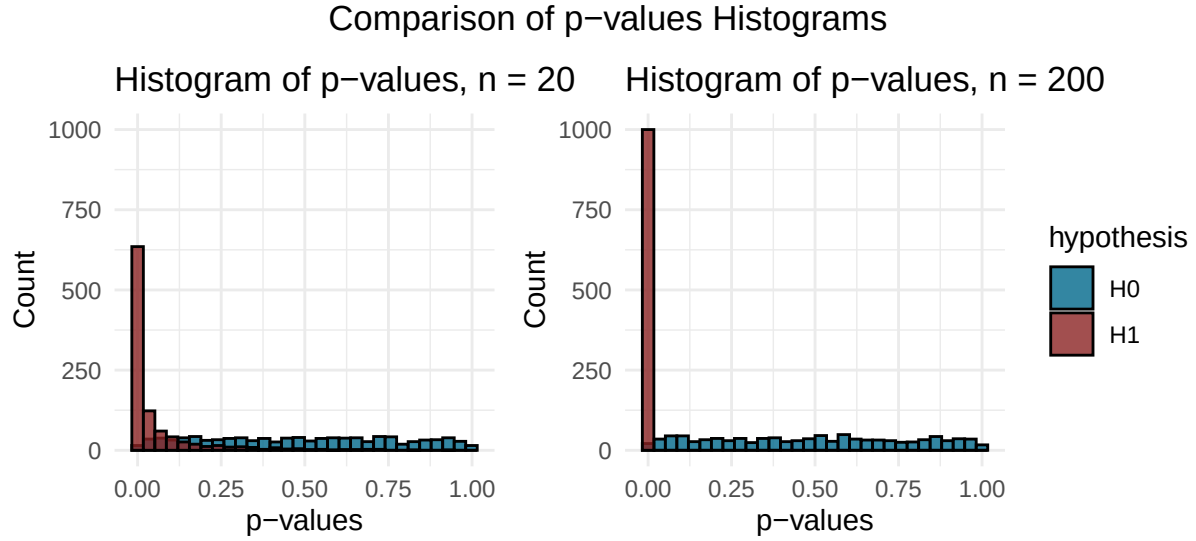
we use the **Neyman-Pearson lemma**.
The **critical value** for this test can be determined using the following inequality:

$$\frac{\prod_{i=1}^{n} f_{H_1}(x, \lambda)}{\prod_{i=1}^{n} f_{H_0}(x, \lambda)} > k,$$

where α, the significance level, represents the probability of committing a Type I error. This inequality provides a threshold for $\sum_{i=1}^{n} x_i$ is our statistic. Since we know that the sum of independent exponential distributions follows a Gamma distribution with parameters n and λ, the critical value condition in terms of Gamma CDF is: $\alpha = \mathbb{P}_{H_0}(\sum_{i=1}^{n} x_i > k^*) \implies k^* = F^{-1}_{Gamma(n,5)}(1 - \alpha)$.
To calculate the **power** of the test, which is the complement of the probability of making a Type II error, the expression becomes: $\mathbb{P}_{H_1}(X \in C) = 1 - F_{Gamma(n,3)}\left(F^{-1}_{Gamma(n,5)}(1 - \alpha)\right)$.
P-value is the probability of observing a test statistic as extreme or more extreme that the observed value under H_0.

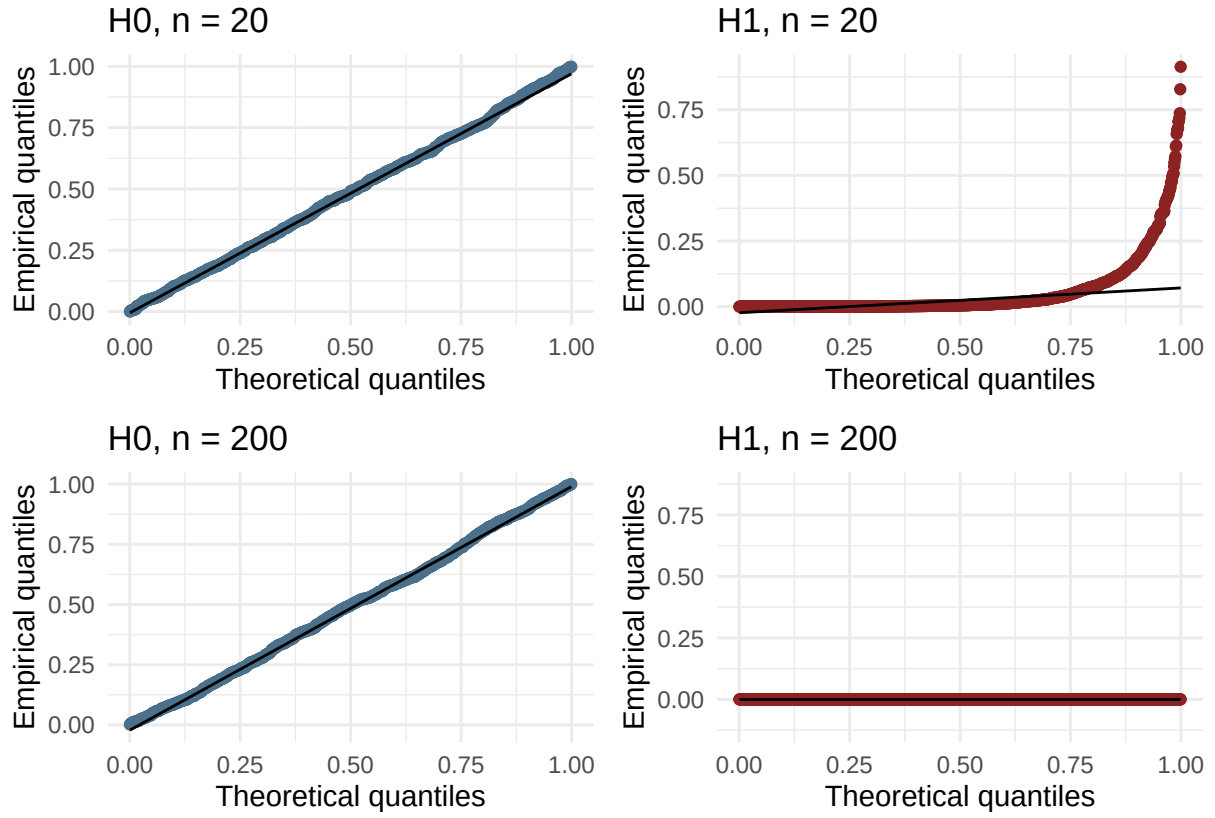
$$p = \mathbb{P}(T > \sum_{i=1}^{n} x_i | T \sim Gamma(n, 5)) = 1 - F_{Gamma(n,5)}(\sum_{i=1}^{n} x_i)$$

For $n = 20$ we generate one random sample from H_0 and another from H_1 and the respective p-values are 0.229 and 0.025. In the case where the sample is drawn from H_0, we fail to reject the null hypothesis since the p-value is greater than α. However, in the second case, where the sample is drawn from H_1, the p-value allows us to reject the null hypothesis, which is the expected outcome.
The p-value follows a uniform distribution on $[0, 1]$ when the data come from H_0 because, under the null hypothesis, all p-values are equally likely. This means the probability of observing any particular p-value between 0 and 1 is the same.
Now we will generate 1000 samples of the size $n = 20$ and $n = 200$ from H_0 and H_1 and calculate respective p-values. We will compare the distribution of these p-values to the distribution derieved in.



For both sample sizes, the p-values from H_0 follow a uniform distribution, as expected. This reflects that

when the null hypothesis is true, p-values are equally likely across the interval $[0, 1]$. Also for both sample sizes, the p-values under H_1 are concentrated around smaller values, indicating that most of them arre under the significance level $\alpha = 0.05$. This suggests that when the alternative hypothesis is true, the test correctly rejects H_0 most of the time, as we would expect. We can also notice that the power of the test increases as the sample size grows, because when $n = 200$ it is more frequent to reject H_0 when it is false.



The Q-Q plots illustrate that uder H_0, p-values follow a uniform distribution for both n values, aligning with theoretical expectationa and ensuring controlled type I error rates. Under H_1, the distribution is highly skewed towards zero, the skew increasing as the sample size grows. This indicated higher test power with larger samples, as expected.
Using these simulations we construct the 95% confidence interval for the type I error of the test when data is from distribution from H_0 and for the power of the test, when data is from H_1. The calculated confidence interval for the power of the test we compare with the theoretically calculated power.

	95% confidence interval type I error	95% confidence interval power	Theoretical power
$n = 20$	[0.032, 0.058]	[0.728, 0.782]	0.758
$n = 200$	[0.040, 0.068]	[1.000, 1.000]	1.000

From the table we can read that the test maintains the correct type I error rate, with confidence intercals for both n sizes including the nominal level $\alpha - 0.05$. For power, the simluated 95% confidence intervals align closely with the theoretical power values. For $n = 20$, the observed interval is $[0.728, 0.782]$ is close to the calculated theoretical power of 0.758, and for $n = 200$, both the simulated and theoretical power reach 1, demonstrating that the test's effectiveness as sample size increases.