6.6 QQ-plot

Example 6.6.1 (GDP per capita)**.** *(a) Does the GDP follow normal distribution?*

(b) Does the logarithm of GDP follow a normal distribution?

(c) Does the GDP follow gamma distribution with parameters estimated based on MLE?

One way to test whether the sample follows a certain distribution is using the quantile plots (Q-Q plots).

(i). Sort the observed data increasingly to get $y_{(i)}$;

(ii). Find the quantiles of the distribution by looking up the fractions $(i - 0.5)/n$ and calcualte the inverse cumulative distribution function to get $q_i = F^{-1}((i - 0.5)/n)$;

(iii). Plot $y_{(i)}$ against the quantile q_i.

Example 6.6.2. *Try the Q-Q plot for t distribution with various degrees of freedom.*

6.7 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (KS test) is used to formally assess hypothesis statements concerning probability distributions. The KS one-sample test tests whether a random sample is coming from a hypothesized null distribution.

$$H_0 : X \sim F_0(x), H_a : X \text{does not follow } F_0(x).$$

The KS statistic is defined as

$$D_n = \sup_x F_n(x) - F(x).$$

Null distribution: under the null hypothesis, if $F_0(x)$ is continuous, then

$$\sqrt{n} D_n \to K,$$

where K is the Kolmogorov distribution defined as

$$K = \sup_{t \in [0,1]} |B(t)|,$$

with the cumulative distribution function given by

$$P(K \leq x) = \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} \exp(-(2k-1)^2 \pi^2 / (8x^2)).$$

Consequently, we set the rejection region as $\mathcal{R} = \{\sqrt{n} D_n > K_\alpha\}$ where $P(K \leq K_\alpha) = 1 - \alpha$. In R, the package *kolmim* provides function to calculate the critical values based on the Kolmogorov distribution.

Example 6.7.1 (GDP)**.**

Example 6.7.2 (Microarray data set)**.**

6.8 Permutation test

When conducting a z-test or t-test, we are actually assuming that the data follow a normal distribution. We can check the normality assumption by, for instance, the Kolmogorov-Smirnov test. If this assumption holds, we know the distribution of the test statistic under the null hypothesis. We can then calculate the p-values and/or rejection regions. This can be refereed to as "Parametric approaches".

What if the distributional assumption does not hold? We still want to test if there exists a significance difference between two means. How should we proceed?

- Transformation of data to make the data normal;
- Choose some tests that do not make such distribution assumptions – "nonparametric approaches".

Permutation tests can be used without the normal assumption for the distribution of data. It is a nonrparametric approach to establish the null distribution of a test statistic.
Idea of permutation test: under the null, assume that the data are exchangeable. We permute the data by shuffling their labels of treatments, and then calculate the test statistic on each permutation. The collection of all test statistic constructs the null distribution.
Steps of Permutation test

(I). Identify the hypothesis;

(II). Choose a test statistic and rejection rule that distinguishes the null form the alternative;

(III). Compute the test statistic for the original observations;

(IV). Rearrange the observations, compute the test statistic for the rearranged data;

(V). Compare the original test statistic with the ones from re-arranged data and calculate the p-values;

(VI). Make conclusion.

Example 6.8.1. *Consider the gene from the prostate cancer data.*