18.650. Statistics for Applications
Fall 2016. Problem Set 7

Due Friday, Oct. 28 at 12 noon

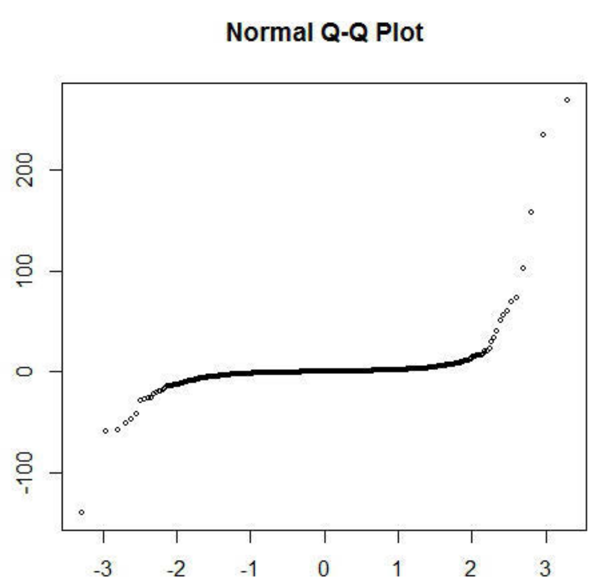
Problem 1 QQ-plots

Recall that the Laplace distribution with parameter $\lambda > 0$ is the continuous probability measure with density $f_\lambda(x) = \frac{\lambda}{2} e^{-\lambda|x|}, \quad x \in \mathbb{R}$ and the Cauchy distribution is the continuous probability measure with density $g(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \quad x \in \mathbb{R}$.

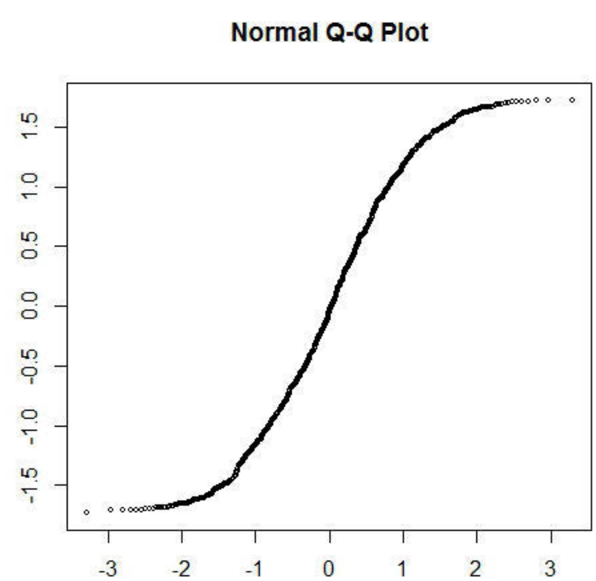
Consider five samples of i.i.d. random variables with the following distributions:

- The standard Gaussian distribution;
- The uniform distribution on $[-\sqrt{3}, \sqrt{3}]$;
- The Cauchy distribution;
- The exponential distribution with parameter 1;
- The Laplace distribution with parameter $\sqrt{2}$.

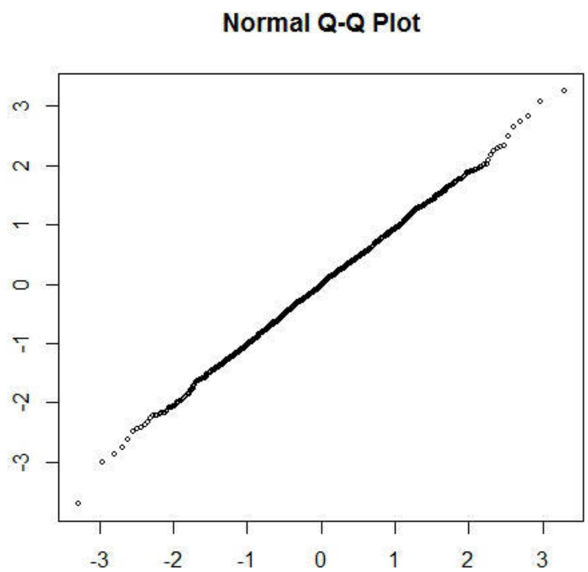
For each of these samples, we have drawn the normal QQ-plots below. Identify which plot corresponds to which sample.



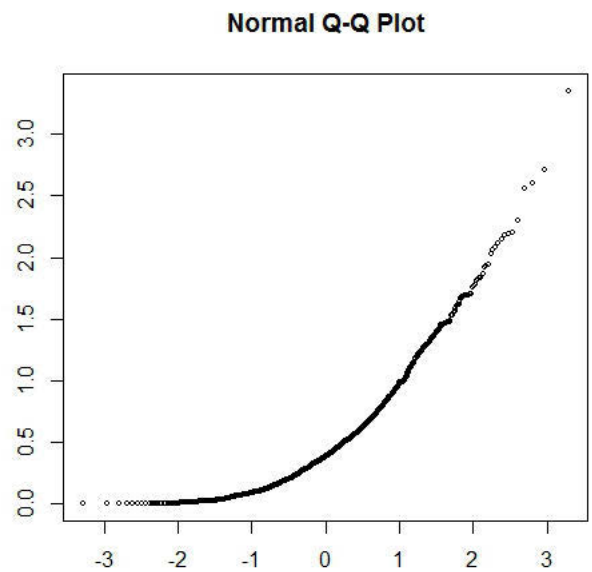
QQ-Plot 1



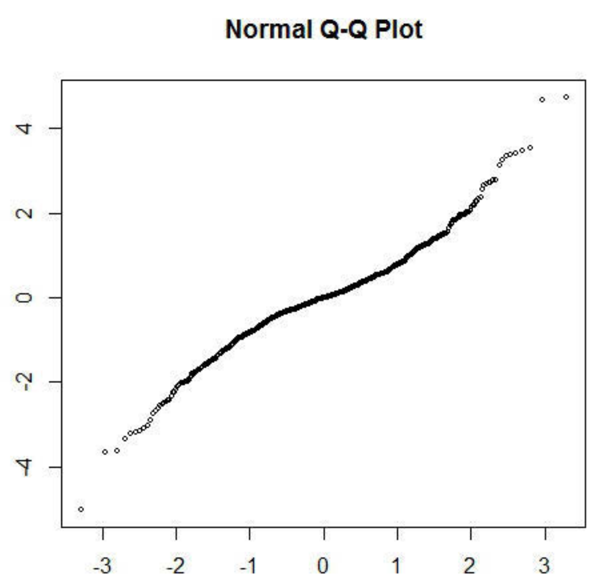
QQ-Plot 2



QQ-Plot 3



QQ-Plot 4



QQ-Plot 5

Problem 2 Kolomogorov-Smirnov test for two samples

Consider two independent samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ of independent real valued continuous random variables, and assume that the X_i's are iid with some cdf F and that the Y_i's are iid with some cdf G. Note that the two samples may have different sizes (if $n \neq m$). We want to test whether $F = G$. We consider the following hypotheses:

$$H_0 : \text{"}F = G\text{"} \quad \text{and} \quad H_1 : \text{"}F \neq G\text{"}.$$

For simplicity, we will assume that in addition to be continuous, F and G are increasing.

1. Propose an example of experiment in which testing whether two samples are generated by the same distribution would be of interest.
2. For $i = 1, \dots, n$, denote by $U_i = F(X_i)$ and for $j = 1, \dots, m$, let $V_j = G(Y_j)$. What are the distributions of the U_i's and the V_j's ?
3. Let F_n be the empirical cdf of the sample $\{X_1, \dots, X_n\}$ and G_m be the empirical cdf of $\{Y_1, \dots, Y_m\}$.
   a) Let $T_{n,m} = \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)|$. Prove that $T_{n,m}$ can be written as the maximum value of a finite collection of numbers.
   b) If H_0 is true, show that
   $$T_{n,m} = \sup_{0 \le x \le 1} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i \le x} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{V_j \le x} \right|.$$
   c) If H_0 is true, what is the joint distribution of the $n + m$ random variables $U_1, \dots, U_n, V_1, \dots, V_m$?
   d) Conclude that the test statistic $T_{n,m}$ is pivotal, i.e., if H_0 is true, the distribution of $T_{n,m}$ does not depend on the unknown distribution of the samples.
   e) Let $\alpha \in (0, 1)$ and let q_α be the $(1 - \alpha)$-quantile of the distribution of $T_{n,m}$ under H_0. In practice, even if this quantile may be available on tables for some values of n and m, you may not be able to find it online for your values of n and m. Describe an algorithm that you could run on a software (e.g., R) in order to get an approximate value of q_α, for a given α.
   f) Define a test with non asymptotic level α for the hypotheses H_0 v.s. H_1.

Problem 3 Test of independence for samples with continuous cdf

Consider n i.i.d. pairs of real random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ with some continuous distribution. We would like to test whether X_1 and Y_1 are independent. Define the following hypotheses:

$$H_0 : \text{"}X_1 \perp\!\!\!\perp Y_1\text{"} \quad \text{and} \quad H_1 : \text{"}X_1 \text{ and } Y_1 \text{ are not independent"}.$$

For $i = 1, \ldots, n$, define R_i as the rank of X_i in the sample $X_1, \ldots, X_n$: E.g., if X_i is the smallest number of this sample, then $R_i = 1$; If X_i is the largest, then $R_i = n$. In a similar fashion, define Q_i as the rank of Y_i in the sample $Y_1, \ldots, Y_n$.

1. Propose an example of experiment in which testing independence of two samples would be of interest.
2. Without a rigorous proof, explain why $R_1, \ldots, R_n$ are not independent random variables.
3. Prove that the distribution of $(R_1, \ldots, R_n)$ does not depend on the (unknown) distribution of the X_i's. Similarly, the distribution of $(Q_1, \ldots, Q_n)$ does not depend on that of the Y_i's.
4. Prove that if H_0 is true, then the two vectors of ranks $(R_1, \ldots, R_n)$ and $(Q_1, \ldots, Q_n)$ are independent.
5. Conclude that if H_0 is true, then the joint distribution of the $2n$ random variables $R_1, \ldots, R_n, Q_1, \ldots, Q_n$ is known and does not depend on the distribution of the original sample.
6. Consider the following test statistic:

$$T_n = \frac{\sum_{i=1}^n (R_i - \bar{R}_n)(Q_i - \bar{Q}_n)}{\sqrt{\sum_{i=1}^n (R_i - \bar{R}_n)^2 \sum_{i=1}^n (Q_i - \bar{Q}_n)^2}}.$$

   T_n is the empirical correlation between the R_i's and the Q_i's. If H_0 is true, then T_n should be very close to zero. We are first going to show that T_n has a very simpler expression.

   a) Prove that

$$\bar{R}_n = \bar{Q}_n = \frac{n+1}{2},$$

$$\sum_{i=1}^n (R_i - \bar{R}_n)^2 = \sum_{i=1}^n (Q_i - \bar{Q}_n)^2 = \frac{n(n^2-1)}{12}.$$

   Hint: Recall that $\sum_{i=1}^n i = \frac{n(n+1)}{2}$ *and* $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$.

   b) Conclude that

$$T_n = \frac{12}{n(n^2-1)} \sum_{i=1}^n R_i Q_i - \frac{3(n+1)}{n-1}.$$

7. Using all the previous questions, prove that if H_0 is true, then T_n has the same distribution as

$$S_n = \frac{12}{n(n^2-1)} \sum_{i=1}^n R_i' Q_i' - \frac{3(n+1)}{n-1},$$

where $(R'_1, \dots, R'_n)$ and $(Q'_1, \dots, Q'_n)$ are the respective rank vectors of two independent samples of i.i.d. uniform random variables in $[0, 1]$.

8. Let $\alpha \in (0, 1)$. Denote by q_α the $(1 - \alpha)$-quantile of S_n. Describe an algorithm that you could run on the software R in order to get an approximate value of q_α, for a given value of n.
9. Define a test for H_0 v.s. H_1 that has non asymptotic level α.