STA437/2005 - Methods for Multivariate Data

Lecture 4

Gun Ho Jang

September 29, 2014

Assessment of Normality

Most of data analysis methods were developed under normality assumption of observed data. Hence it should be checked whether or not observed data is normally distributed. Simple check-ups.

- Whether or not marginal distributions follow normal distribution
- Whether or not pair-wise distributions follow normal distribution
- Whether or not there are wild observations which are quite different from normal distributions.

Univariate normal check-ups

- Let $Z \sim N(\mu, \sigma^2)$.
- $P(X \in (\mu - k\sigma, \mu + k\sigma)) = \Phi(k) - \Phi(-k) = 2(\Phi(k) - 1/2) = 2\Phi(k) - 1$.
- The probability is 0.6827 if $k = 1$ and 0.9545 if $k = 2$.
- Normality is violated if the proportions of samples contained in $\widehat{\mu} \pm k\widehat{\sigma}$ is quite different from normal distribution.

- Let $Z_1, \ldots, Z_n$ be i.i.d. $N(\mu, \sigma^2)$.
- Let $C_j = \sum_{i=1}^{n} I(Z_i \in \widehat{\mu} \pm k\widehat{\sigma}) \sim$ Binomial(n, p_k) where $p_k = \Phi(k) - \Phi(-k)$.
- Approximately, $\sqrt{n}(C_k - np_k) \approx N(0, p_k(1 - p_k))$.
- Big $|\sqrt{n}(C_k - np_k)|$ indicates departure from normal distribution.
- If $|\sqrt{n}(C_k - np_k)| > 3$, then departure from normality with confidence more than 99.7%.

Q-Q plot

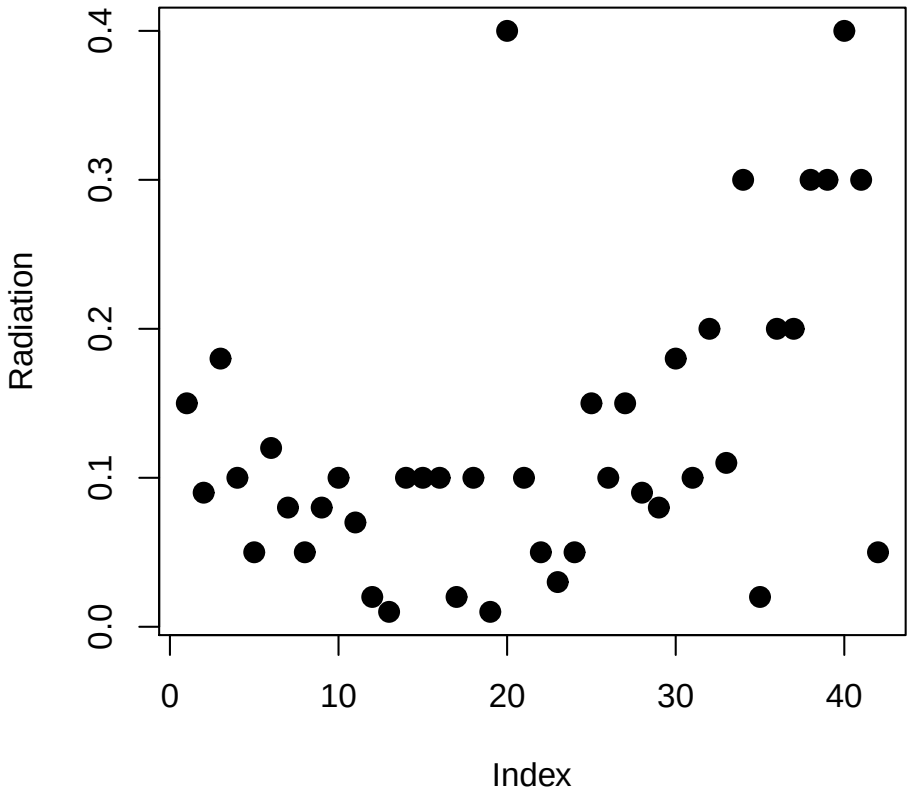
- Let $Z_1, \ldots, Z_n \sim i.i.d.\ N(0,1)$.
- Order statistics $Z_{(1)}, \ldots, Z_{(n)}$
- Each $Z_{(k)}$ is expected to be centred around $q_{(k)}$.
- Such $q_{(k)}$ is approximated by the $(k-1/2)/n$th (or $k/(n+1)$th or $(k-3/4)/(n+1/4)$th) quantile of the standard normal.
- If the data is normally distributed, then the plot of $(q_{(k)}, z_{(k)})$ should be closed to a line.
- The correlation coefficient of $(q_{(k)}, z_{(k)})$ is given by

$$r_Q = \frac{\sum_{i=1}^n (z_{(i)} - \bar{z})(q_{(i)} - \bar{q})}{\left[\sum_{i=1}^n (z_{(i)} - \bar{z})^2 \sum_{i=1}^n (q_{(i)} - \bar{q})^2\right]^{1/2}}$$

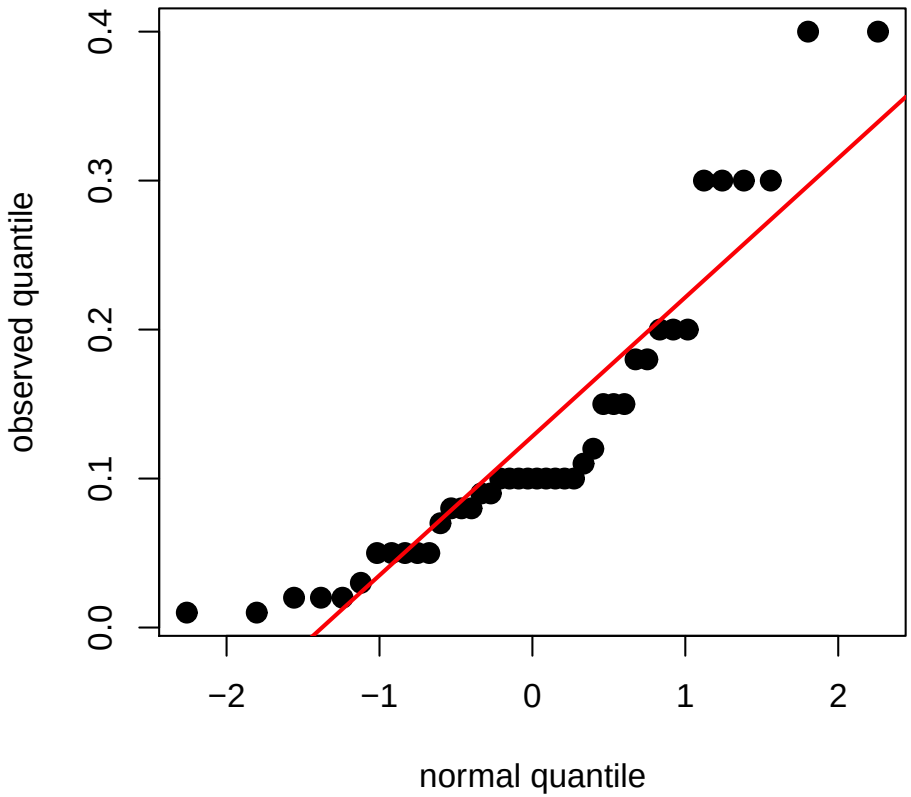
- This coefficient is closed to 1 if the observed data is normally distributed.
- A few tests were developed on ordered numbers.
- Shapiro-Wilks test is very popular which is implemented in R as `shapiro.test`.
- Jarque-Bera test is popular in time series data analysis which is based on the limit behaviour of third and fourth moment.

Example

Radiation data in textbook

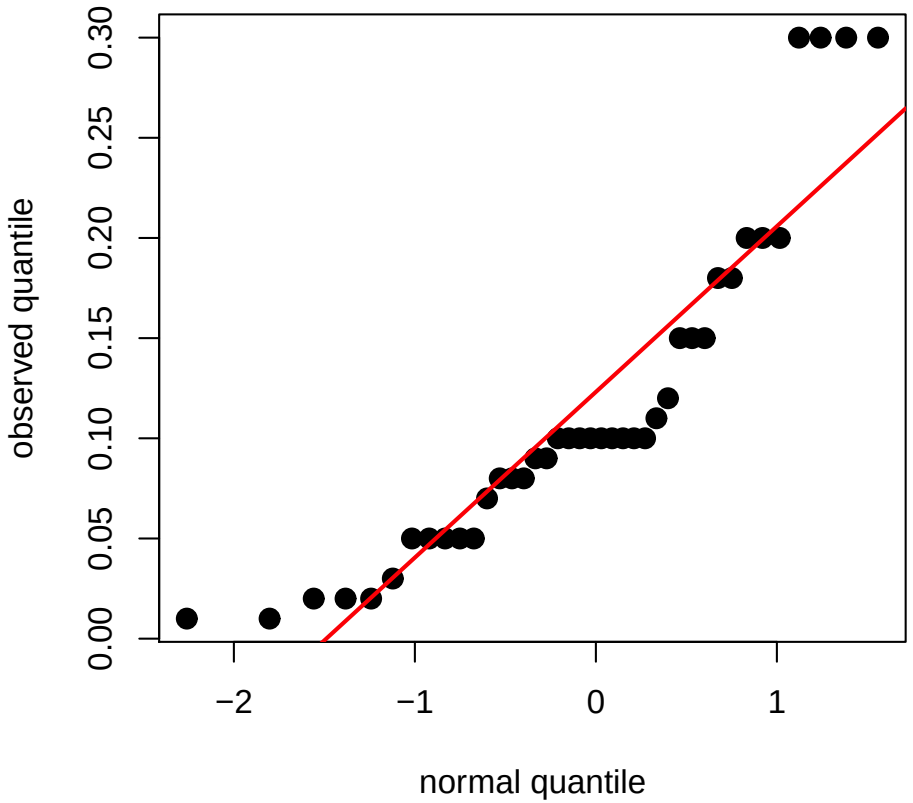


Example



Shapiro-Wilks test p-value is $9.902e-5$.

Example



Shapiro-Wilks test p-value is $4.956e-4$.

Jarque-Bera Test

- One of the most popular normality test is Shapiro-Wilks test which assesses linearity of sample quantile and normal quantile
- Most statistics packages implement Shapiro-Wilks test. In `R`, `shapiro.test` tests normality up to of size 5,000
- One of the most common normality test in time series is Jarque-Bera test which is asymptotic test based on third and fourth moments, that is,

$$JB = \frac{n}{6} Skewness^2 + \frac{n}{24}(Kurtosis - 3)^2 \quad \xrightarrow{d} \chi^2(2)$$

  where $Skewness = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3/\hat{\sigma}^3$ and $Kurtosis = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4/\hat{\sigma}^4$, and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
- The asymptotic distribution might be a bit different from the sample distribution if the sample size is small.
- In `R`, Jarque-Bera test is implemented in `jarque.bera.test` in `tseries` package.

Normality of Multivariate

- Normality of a multivariate random vector can be assessed through χ^2 distribution.
- $(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu) \leq \chi^2_\gamma(k)$ can be a basis for normality test.
- For example, $\#\{\mathbf{x}_i : (\mathbf{x}_i-\bar{\mathbf{x}})^\top S^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}) \leq \chi^2_\gamma(k)\}$ can be approximated by Binomial(n,γ).

Assessment of Multivariate Normality

- Compute $d_j^2 = (\mathbf{x}_i-\bar{\mathbf{x}})^\top S^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})$
- Sort and get order statistics $d^2_{(1)} \leq d^2_{(2)} \leq \cdots \leq d^2_{(n)}$
- Plot $(\chi^2_{(j-1/2)/n}(p), d^2_{(j)})$
- Assess whether the plot is linear or not.

Outlier Detection

- Many data sets contain a few unusual data point which may not belong to the pattern of the most observation.
- Such *outliers* may hinder recognition of the pattern of the most observed data.
- With appropriate justifications, outliers can be removed and the pattern of most observed data can be efficiently recognized.

Outlier Detection

- In most cases, outliers can be visually detectable

Steps for Detecting Outliers

(1) Make a univariate data plot such as histogram or density.

(2) Make a scatter plot for each pair of variables

(3) Compute standardized value $z_{jk} = (x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$. *Large* values of $|z_{jk}|$ could be resulted in outliers. Interpretations of "large" are depend on the dimension of data. A recommended cutoff is 3.5 for moderate sample size.

(4) Compute χ^2 statistics $(\mathbf{x}_j - \bar{\mathbf{x}})^\top S^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$. If the value is large, then the data point could be an outlier. A cutoff should be extremely high quantile of $\chi^2(p)$.

Variance Stabilization Transformation

Estimators of certain types show different characteristics in variance. For example, the variance of count data like Poisson random variables is proportional to the mean. Which can be stabilized using a square-root transformation.

Variance Stabilization

count data, y	$\sqrt{y}$
proportion, $\hat{p}$	$\text{logit}(\hat{p}) = \frac{1}{2}\log(\frac{\hat{p}}{1-\hat{p}})$
correlation, $\hat{r}$	Fisher's $z(\hat{r}) = \frac{1}{2}\log(\frac{1+\hat{r}}{1-\hat{r}})$

Box-Cox Transformation

A generalized power transformation given by

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(x) & \lambda = 0 \end{cases}$$

for positive $x > 0$. The tuning parameter λ can be chosen to maximize

$$\ell(\lambda) = -\frac{n}{2} \log \left(\frac{1}{n} \sum_{j=1}^{n} (x_j^{(\lambda)} - \overline{x^{(\lambda)}})^2 \right) + (\lambda - 1) \sum_{j=1}^{n} \log x_j$$

where $\overline{x^{(\lambda)}} = \frac{1}{n} \sum_{j=1}^{n} x_j^{(\lambda)}$.
The function `boxcox` can be used in `MASS` package of `R`.

Inference about Mean Vector

Consider univariate case, that is, $x_1, \ldots, x_n \sim i.i.d.\ N(\mu, \sigma^2)$. Then $\sqrt{n})(\bar{x} - \mu)/\sqrt{(n-1)s^2} \sim t(n-1)$. Hence a hyperthesis assessment $H_0 : \mu = \mu_0$ can base on the statistic $\sqrt{n})(\bar{x} - \mu)/\sqrt{(n-1)s^2} \sim t(n-1)$.
A natural generalization of univariate t-statistic for multivariate is

$$T^2 = n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu)$$

which is called *Hotelling's T^2-statistic* since it is square of t-statistic for univariate case.
It is known that $T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p}$. A long proof is given by Harold Hotelling.

Example

Note that $t(k) \sim N(0,1)/[\chi^2(k)/k]^{1/2}$. Then $t(k)^2 \sim \chi^2(1)/(\chi^2(k)/k) \sim kF_{1,k}$.

Note that $n(\bar{\mathbf{x}} - \mu)^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mu) \sim \chi^2(p)$. Roughly speaking, $\Sigma^{-1/2} S \Sigma^{-1/2} \sim (n-1)^{-1} W_p(I_p, n-1)$ contributes $[(n-p)/(p(n-1))]\chi^2(n-p)$ along with $\Sigma^{-1/2}(\bar{\mathbf{x}} - \mu)$.