Assessing Normality – The Univariate Case

• In general, most multivariate methods will depend on the distribution of $\mathbf{X}$ or on distances of the form

$$n(\mathbf{X} - \mu)'S^{-1}(\mathbf{X} - \mu)$$

• Large sample theory tells us that if the sample observations $X_1,\ldots,X_n$ are iid from some population with mean $\mu$ and positive definite covariance $\Sigma$, then for large $n - p$

$$\sqrt{n}(\mathbf{X} - \mu) \text{ is approx. } N_p(0, \Sigma)$$

$$n(\mathbf{X} - \mu)'S^{-1}(\mathbf{X} - \mu) \text{ is approx. } \chi^2_p.$$
Assessing Normality (cont’d)

- This holds regardless of the form of the distribution of the observations.

- In making inferences about mean vectors, it is not crucial to begin with MVN observations if samples are large enough.

- For small samples, we will need to check if observations were sampled from a multivariate normal population.
Assessing Normality (cont’d)

- Assessing multivariate normality is difficult in high dimensions.

- We first focus on the univariate marginals, the bivariate marginals, and the behavior of other sample quantities. In particular:
  1. Do the marginals appear to be normal?
  2. Do scatter plots of pairs of observations have an elliptical shape?
  3. Are there 'wild' observations?
  4. Do the ellipsoids 'contain' something close to the expected number of observations?
Assessing Normality (cont’d)

• One other approach to check normality is to investigate the behavior of conditional means and variances. If \( X_1, X_2 \) are jointly normal, then

1. The conditional means \( E(X_1|X_2) \) and \( E(X_2|X_1) \) are linear functions of the conditioning variable.

2. The conditional variances do not depend on the conditioning variables.

• Even if all the answers to questions appear to suggest univariate or bivariate normality, we cannot conclude that the sample arose from a MVN normal distribution.
Assessing Normality (cont’d)

• If $X \sim MVN$, all the marginals are normal, but the converse is not necessarily true. Further, if $X \sim MVN$, then the conditionals are also normal, but the converse does not necessarily follow.

• In general, then, we will be checking whether necessary but not sufficient conditions for multivariate normality hold or not.

• Most investigations in the book use univariate normality, but we are also going to present some practical and recent work on assessing multivariate normality.
Univariate Normal Distribution

• If $X \sim N(\mu, \sigma^2)$, we know that

  Probability for interval $(\mu - \sqrt{\sigma^2}, \mu + \sqrt{\sigma^2}) \approx 0.68$

  Probability for interval $(\mu - 2\sqrt{\sigma^2}, \mu + 2\sqrt{\sigma^2}) \approx 0.95$

  Probability for interval $(\mu - 3\sqrt{\sigma^2}, \mu + 3\sqrt{\sigma^2}) \approx 0.99$.  

• In moderately large samples, we can count the proportion of observations that appear to fall in the corresponding intervals where sample means and variances have been plugged in place of the population parameters.

• We can implement this simple approach for each of our $p$ variables.
Normal Q-Q plots

- Quantile-quantile plots can also be constructed for each of the $p$ variables.

- In a Q-Q plot, we plot the sample quantiles against the quantiles that would be expected if the sample came from a standard normal distribution.

- If the hypothesis of normality holds, the points in the plot will fall along a straight line.
Normal Q-Q plots

- The slope of the estimated line is an estimate of the population standard deviation.

- The intercept of the estimated line is an estimate of the population mean.

- The sample quantiles are just the sample order statistics. For a sample $x_1, x_2, ..., x_n$, quantiles are obtained by ordering sample observations

$$x(1) \leq x(2) \leq ... \leq x(n),$$

where $x(j)$ is the $j$th smallest sample observation or the $j$th sample order statistic.
Normal Q-Q plots (cont’d)

• When the sample quantiles are distinct (as can be expected from a continuous variable), exactly \( j \) observations will be smaller than or equal to \( x(j) \).

• The proportion of observations to the left of \( x(j) \) is often approximated by \((j - 0.5)/n\). Other approximations have also been suggested.

• We need to obtain the quantiles that we would expect to observe if the sample observations were sampled from a normal distribution. For a standard normal random variable, quantiles are computed as

\[
\Pr(Z \leq q(j)) = \int_{-\infty}^{q(j)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)dz = j - \frac{1}{2} = p(j).
\]
Normal Q-Q plots (cont’d)

• For example, if $p(j) = 0.5$, then $q(j) = 0$ (or the median), and if $p(j) = 0.95$, then $q(j) = 1.645$.

• Given the sample size $n$, we can compute the expected standard normal quantile ($q(j)$) for each ordered observation using $p(j) = (j - 0.5)/n$. SAS uses the Blom approximation with $p(j) = (j - 0.375)/(n + 0.25)$.

• If the plot of the pairs $(x(j), q(j))$ shows a straight line, we do not reject the hypothesis of normality.

• If observations are tied, the associated quantile is the average of the quantiles that would have corresponded to slightly different values.
### Example

<table>
<thead>
<tr>
<th>Ordered Observations $x(j)$</th>
<th>Probability Level $(j - 0.5)/n$</th>
<th>Standard Normal Quantiles $q(j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.00</td>
<td>0.05</td>
<td>-1.645</td>
</tr>
<tr>
<td>-0.10</td>
<td>0.15</td>
<td>-1.036</td>
</tr>
<tr>
<td>0.16</td>
<td>0.25</td>
<td>-0.674</td>
</tr>
<tr>
<td>0.41</td>
<td>0.35</td>
<td>-0.385</td>
</tr>
<tr>
<td>0.62</td>
<td>0.45</td>
<td>-0.125</td>
</tr>
<tr>
<td>0.80</td>
<td>0.55</td>
<td>0.125</td>
</tr>
<tr>
<td>1.26</td>
<td>0.65</td>
<td>0.385</td>
</tr>
<tr>
<td>1.54</td>
<td>0.75</td>
<td>0.674</td>
</tr>
<tr>
<td>1.71</td>
<td>0.85</td>
<td>1.036</td>
</tr>
<tr>
<td>2.30</td>
<td>0.95</td>
<td>1.645</td>
</tr>
</tbody>
</table>
Example (cont’d)
Example (cont'd)

• The normal quantiles can be computed with SAS using the *probit* function or the *RANK* procedure.

• Note that

\[ q(j) = \Phi^{-1}\left(\frac{j - 0.5}{n}\right) = \text{probit}\left(\frac{j - 0.5}{n}\right) \]

with \( \Phi(a) \) the standard normal cumulative distribution function evaluated at \( a \).

• SAS uses a different (Blom) approximation to the probability levels when the "normal" option is executed in the Rank procedure:

\[ q(j) = \Phi^{-1}\left(\frac{j}{n + \frac{3}{8}}\right). \]
Microwave ovens: Example 4.10

- Microwave ovens are required by the federal government to emit less than a certain amount of radiation when the doors are closed.

- Manufacturers regularly monitor compliance with the regulation by estimating the probability that a randomly chosen sample of ovens from the production line exceed the tolerance level.

- Is the assumption of normality adequate when estimating the probability?

- A sample of $n = 42$ ovens were obtained (see Table 4.1, page 180). To assess whether the assumption of normality is plausible, a Q-Q plot was constructed.
Microwaves (cont’d)
Goodness of Fit Tests
Correlation Test

• In addition to visual inspection, we can compute the correlation between the $x_{(j)}$ and the $q_{(j)}$:

$$r_Q = \frac{\sum_{i=1}^{n}(x(i) - \bar{x})(q(i) - \bar{q})}{\sqrt{\sum_{i=1}^{n}(x(i) - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(q(i) - \bar{q})^2}}.$$ 

• We expect values of $r_Q$ close to one if the sample arises from a normal population.

• Note that $\bar{q} = 0$, so the above expression simplifies.
Correlation Test (cont’d)

• The sampling distribution of $r_Q$ has been derived (see Looney and Gulledge, *The American Statistician* 39:75-79) and percentiles of its distribution have been tabulated (see Table 4.2 in book).

• Using the tabled values, we can test the hypothesis of normality and for a sample of size $n$ can reject it at level $\alpha$ if $r_Q$ falls below the corresponding table value.

• The critical values for $r_Q$ depend on both $n$ and $\alpha$. 
Correlation Test (cont’d)

• For the earlier example with \( n = 10 \), we have \( r_Q = 0.994 \).

• The critical value from Table 4.2 in book for \( \alpha = 0.05 \) and \( n = 10 \) is 0.9198.

• Since \( r_Q > 0.9198 \) we fail to reject the hypothesis that the sample was obtained from a normal distribution.
Shapiro-Wilks’ Test

• A weighted correlation between the $x_{(j)}$ and the $q_{(j)}$:

$$W = \frac{\sum_{i=1}^{n} a_j (x_i - \bar{x})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^{n} a_j^2 (x_i - \bar{x})^2 \sqrt{\sum_{i=1}^{n} (q_i - \bar{q})^2}}}.$$

• We expect values of $W$ close to one if the sample arises from a normal population.

• SAS has stored values of $a_j$ for $n < 2000$
Empirical Distribution Function (EDF) Tests

• Compare the EDF

\[ F_n(x) = \frac{\text{Number of observations } \leq x}{n} \]

to an estimate of the hypothesized distribution.

• For the hypothesized family of normal distributions, compare with

\[ F(x; \hat{\mu}, \hat{\sigma}^2) = \Phi \left( \frac{x - \bar{x}}{s} \right) \]
EDF Tests: Anderson-Darling Test

• Order the observation from smallest to largest:

\[ x(1) \leq x(2) \leq \cdots \leq x(n) \]

• Anderson-Darling Statistic

\[
A_n^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x, \hat{\theta})]^2}{F(x, \hat{\theta})[1 - F(x, \hat{\theta})]}dF(x, \hat{\theta})
\]

\[
= -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1) \left[ \ln(p_i) + \ln(1 - p_{n+1-i}) \right]
\]

where \( p_i = \Phi \left( \frac{x(i) - \bar{x}}{s} \right) \)
EDF Tests: Kolmogorov-Smirnov Test

- Order the observation from smallest to largest:

\[ x(1) \leq x(2) \leq \cdots \leq x(n) \]

- Kolmogorov-Smirnov Statistic

\[ D_n = \max \left( D^{-}, D^{+} \right) \]

where \( D_{n}^{-} = \max_{1 \leq i \leq n} |p_{i} - \frac{i-1}{n}| \)

and \( D_{n}^{+} = \max_{1 \leq i \leq n} |p_{i} - \frac{i}{n}| \)

and \( p_{i} = \Phi \left( \frac{x(i) - \bar{x}}{s} \right) \)
EDF Tests

- Reject normality for large values of $A_{n}^{2}$ or $D_{n}$
- Approximate upper percentiles for $D_{n}$ are

$$D_{n,0.05} = 0.895 \left[ \sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}} \right]^{-1}$$

$$D_{n,0.01} = 1.035 \left[ \sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}} \right]^{-1}$$

- Approximate upper percentiles for $A_{n}^{2}$ are

$$A_{n,0.05}^{2} = 0.7514 \left[ 1 - \frac{0.795}{n} - \frac{0.89}{n^{2}} \right]$$

$$A_{n,0.01}^{2} = 1.0348 \left[ 1 - \frac{1.013}{n} - \frac{0.93}{n^{2}} \right]$$