Hypothesis Tests
Part 5: Tests of Normality

DASC 512

Why Test for Normality?

Most statistical tests we've discussed assume normality in data. That assumption won't go away in the second half of the course.

We must apply due diligence to confirm this assumption. Some strategies:

- Visual: QQ-Plots
- CDF Comparison Tests: Kolmogorov-Smirnov test, Lilliefors test, Anderson-Darling test, Shapiro-Wilk test
- Moment-based Tests: D'Agostino-Pearson Omnibus test

Details on how these tests work are beyond the scope of this course, but let's look at how to use them.

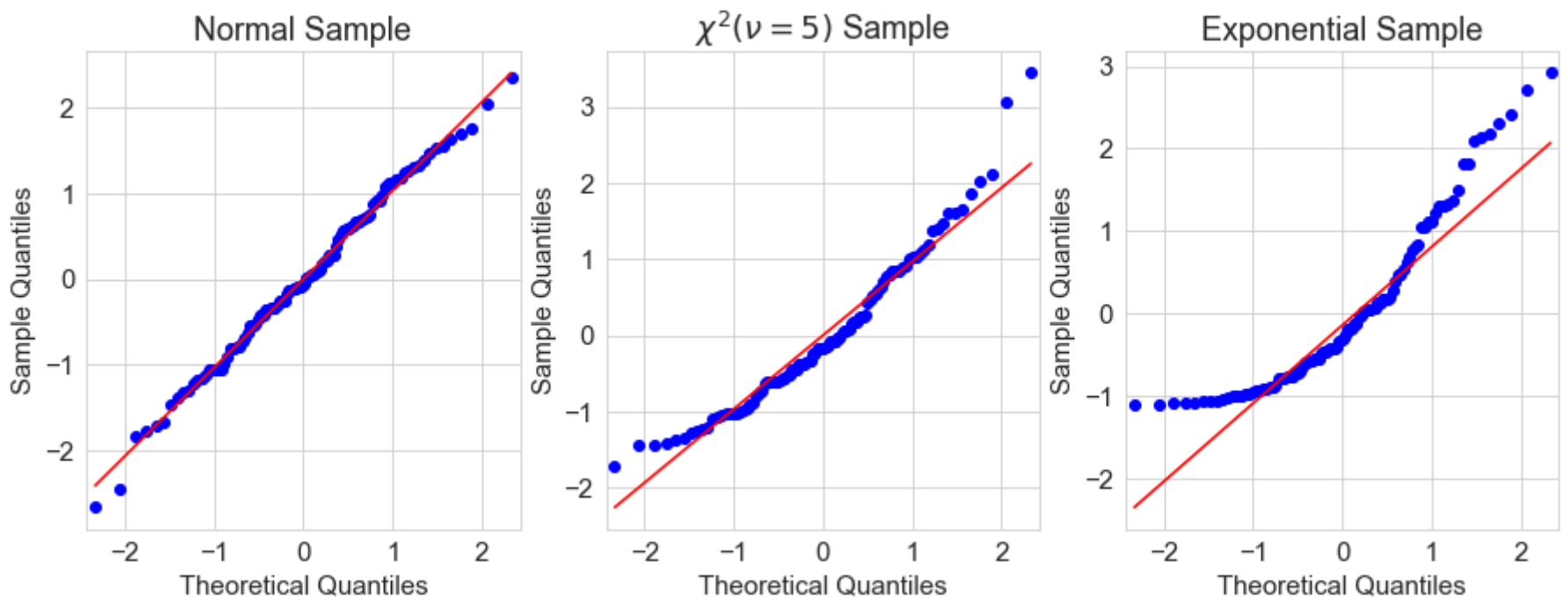
When to Use Each Test

- Visual – **always**
- Shapiro-Wilk test – Small samples $(n < 100)$
  - High power in even small samples
  - May be overly sensitive especially for large samples
- Omnibus test – Large samples $(n > 100)$
  - Creates a test statistic from mean, variance, skewness, and kurtosis
  - Not overly sensitive for large samples
- Kolmogorov-Smirnov test – For distributions other than normal
  - Anderson-Darling can also be used.

Visual Inspection for Normality

One of the consistently best ways to assess normality

- QQ-plot is a scatterplot of observed vs. expected quantiles of that distribution. This can be done for any distribution.



Python tests

- It is good practice to test normality computationally, especially for small samples, but remember that statistical significance is not the same as practical significance.
- Small samples: stats.shapiro(data)
- Large samples: stats.normaltest(data)
- Exponential, Logistic, Gumbel distributions: stats.anderson((data-np.mean(data))/np.std(data, ddof=1)), dist='dist')
- Other distributions: stats.kstest((data-np.mean(data))/np.std(data, ddof=1)), cdf='dist')
  - Note: 'dist' should be any of the dist.* distributions, i.e., 'norm', 'expon', 'binom', etc.

Recap

- **Always** look at a QQ-plot and ask “is this good enough?”
- Consider your sample size in choosing an analytical test
  - <100 – Shapiro-Wilk
  - >100 – Omnibus
- If you need to test for a distribution other than the normal
  - Anderson-Darling
  - Kolmogorov-Smirnov (K-S)