CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 11
Q-Q and Normal Probability Plots

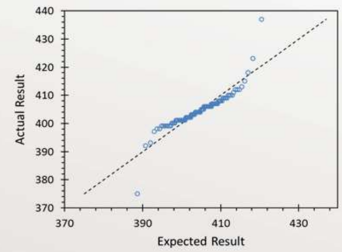
Chris A. Mack

Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

Q-Q Plots

- The normal probability plot is one example of a quantile-quantile (Q-Q) plot



Steps for Normal Probability Plots

- Rank order the data x (k = 1 is the smallest)
- For sample size n, calculate rank quantile as

$$Emperical\ CDF\colon\ U_k = \frac{k}{n}$$ (or a variation of this)

- Find expected data point as

$$InverseCDF(U_k) = \bar{x} + s_x * InverseStandardNormalCD \quad _k)$$

- Plot actual data point vs. expected data point

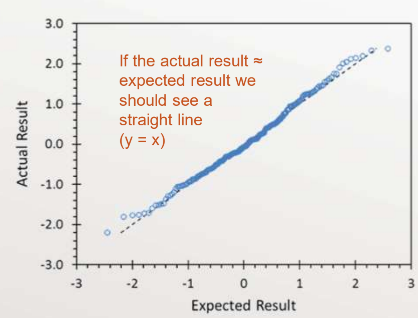
Rank Quantiles, U_k

- Simplest approach: $U_k = k/(n + 1)$, which avoids $U_k = 1$
- Symmetric approach: $(k - a)/(n + 1 - 2a)$ for a in the range from 0 to 1/2.
  - (k − 0.3) / (n + 0.4) [1]
  - (k − 0.3175) / (n + 0.365) [2] ⟵ We'll use this one
  - (k − 0.326) / (n + 0.348) [3]
  - (k − ⅓) / (n + ⅓) [4]
  - (k − 0.375) / (n + 0.25) [5] ⟵ This one is also popular
  - (k − 0.4) / (n + 0.2) [6]
  - (k − 0.44) / (n + 0.12) [7]
  - (k − 0.567) / (n − 0.134) [8]
  - (k − 1) / (n − 1) [9]
- For reasonably large n, the choice doesn't matter much

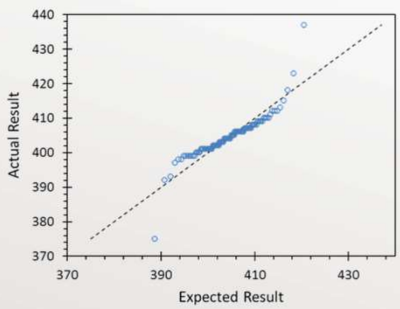
References

1. A. Benard and E.C. Bos-Levbach, "The plotting of observations on probability paper", *Statistica Neerlandica*, **7**, 163-173 (1953).
2. *Engineering Statistics Handbook: Normal Probability Plot* – Note this also uses a different expression for the first & last points. See also Filliben 1975. This expression is an estimate of the medians of U(k). See http://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm.
3. G.-H Yu and C.-C. Huang, "Distribution free plotting position", *Stochastic environmental research and risk assessment*, **15**(6), 462-476 (2001).
4. J.W. Tukey, "The future of data analysis", *Annals of Mathematical Statistics*, **33** (1962).
5. Blom, G. (1958), *Statistical estimates and transformed beta variables*, New York: John Wiley and Sons (used in MINITAB). This expression is an estimate of the means of U(k).
6. C. Cunnane, "Unbiased plotting positions: A Review", *J. of Hydrology*, **37**, 205-222 (1978).
7. I. I. Gringorten, "A Plotting rule for extreme probability paper", *Journal of Geophysical Research*, **68**, 813-814 (1963). Used to plot points in tests for the Gumbel distribution.
8. R.I. Larsen, T.C. Curran, J.R.W.F. Hunt, "An air quality data analysis system for interrelating Effects, standards and needed source reductions: Part 6. calculating concentration reductions", *J. of the Air Pollution Control Association*, **30**, 662-669 (1980).
9. J. J. Filliben, "The Probability Plot Correlation Coefficient Test for Normality", *Technometrics*, **17**(1), 111–117 (February 1975). These are equal to the modes of U(k).

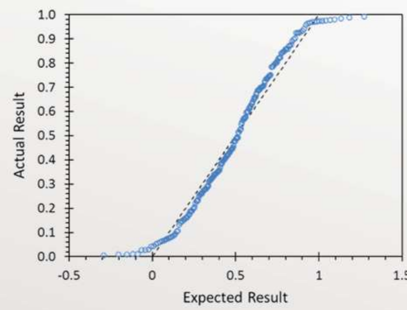
True Normal Distribution



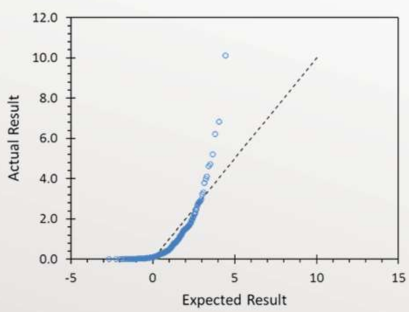
Heavy Tails



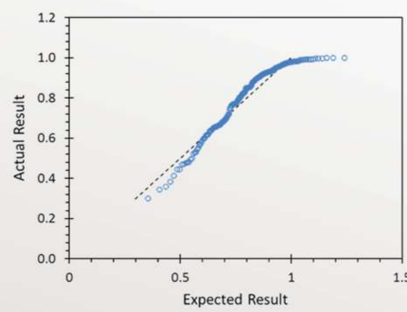
Light Tails



Skewed Right



Skewed Left



Critical Correlation Coefficient

- Calculate the correlation coefficient between the actual result and the expected result from the normal probability plot
- If the correlation coefficient is above the critical value (see table), we cannot reject the hypothesis that the true distribution is normal

	Significance Level alpha				
n	0.005	0.01	0.025	0.05	0.1
5	0.807	0.826	0.856	0.880	0.903
6	0.820	0.838	0.866	0.888	0.910
7	0.828	0.850	0.877	0.898	0.918
8	0.840	0.861	0.887	0.906	0.924
9	0.854	0.871	0.894	0.912	0.930
10	0.862	0.879	0.901	0.918	0.934
12	0.876	0.892	0.912	0.928	0.942
14	0.890	0.905	0.923	0.935	0.948
16	0.899	0.913	0.929	0.941	0.953
18	0.908	0.920	0.935	0.946	0.957
20	0.916	0.926	0.940	0.951	0.960
25	0.929	0.939	0.951	0.959	0.966
30	0.939	0.947	0.957	0.964	0.971
35	0.947	0.954	0.962	0.969	0.974
40	0.953	0.959	0.966	0.972	0.977
50	0.961	0.966	0.972	0.977	0.981
60	0.967	0.971	0.976	0.980	0.984
70	0.971	0.975	0.979	0.983	0.986
80	0.975	0.978	0.982	0.985	0.987
100	0.979	0.982	0.985	0.987	0.989

Generalized Q-Q Plot

- For a normal probability plot, we use the inverse normal CDF of the empirical CDF to calculate the expected result
- For a general Q-Q plot, we can use any CDF that can be inverted to calculate the expected result
- By plotting the actual result versus the expected result, parameter values that give the best match can be found
  - For example, use the Gamma distribution and vary the alpha and beta parameters

Finding a Distribution with the Q-Q Plot

- Beta Distribution: support over [0,1] and parametrized by positive shape parameters α and β
  - Used for a random variable that is a fraction (weight fraction, frequency of population with some characteristic, fraction of time that some characteristic is present, etc.)
- Gamma Distribution: support over [0,∞] and parametrized by a shape parameter k and a scale parameter θ (or α and β)
  - Used for waiting times, size of insurance claims, rainfall, etc.
  - Others with same support: log-normal, F-distribution, Weibull, Rayleigh, Maxwell-Boltzmann
- Symmetric distributions with support over [-∞,∞]:
  - Normal, Cauchy, Double Exponential, Student's t, Logistic

Lecture 11: What have we learned?

- What are the steps for generating a normal probability plot?
- How does one interpret the shape of the normal probability plot?
- How can we test for normality using a normal probability plot?
- What other distributions can we test against a given data set?