MS Module 1 Normal probability plots – practice problems

(The attached PDF file has better formatting.)

Probability distributions used in this course include normal distributions, *t* distributions (heavy-tailed), uniform distributions (light-tailed), *F* distributions (positively skewed), χ^2 distributions (positively skewed), and others (binomial distributions, Poisson distributions).

Normal probability plots test if a sample of values is normally distributed, heavy-tailed, light-tailed, positively skewed, negatively skewed, or other (such as a distribution with two or more modes). The test is qualitative, not quantitative. These plots are used throughout the textbook.

Exercise 1.1: Distributions

A. What does the cumulative distribution function (CDF) show?
B. How does the shape of the CDF differ for discrete vs continuous distributions?
C. What is the probability mass function for a discrete distribution?
D. What is the probability density function for a continuous distribution?
E. What does the quantile function show?
F. What do location, scale, and shape mean?
G. What does heavy tailed vs thin tailed mean?
H. What is meant by positively skewed (right-skewed) vs negatively skewed (left skewed)?

Part A: The cumulative distribution function (CDF), or F(x), is the probability that a random variable is equal to or less than x. Page 105 (equation 3.3) of the textbook explains:

> The cumulative distribution function (cdf) F(x) of a discrete random variable X with probability mass function p(x) is defined for every number x by $F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y)$. For any number x, F(x) is the probability that the observed value of X will be at most x.

Page 163 adds:

> The cumulative distribution function (cdf) F(x) for a discrete random variable X gives, for any specified number x, the probability $P(X \leq x)$. It is obtained by summing the probability mass function p(y) over all possible values y satisfying $y \leq x$. The cdf of a continuous random variable gives the same probabilities $P(X \leq x)$ and is obtained by integrating the probability density function f(y) between the limits $-\infty$ and x.

Illustration: A two-side 95% significance level for a normal distribution is 1.96 standard deviations. For a standard normal distribution with a mean of zero and a standard deviation of one, the probability that a random draw is between $0 - 1.96 \times 1 = -1.96$ and $0 + 1.96 \times 1 = 1.96$ is 95%. The normal distribution is symmetric, so the probability is 2.5% that the random draw is below –1.96 and 2.5% that it is above 1.96. The CDF for the standard normal distribution is 2.5% at –1.96 and 97.5% at +1.96. (Two sided confidence intervals are explained in a later module.)

Part B: The CDF for a continuous distribution is a smooth curve rising from zero at the lower limit of the range to one at the upper limit of the range.

For discrete distributions, the CDF is a step function. For example, a Poisson distribution is defined on the non-negative integers. The CDF is a horizontal line at y=0 from $-\infty$ until the smallest value x=0, jumps up to the probability of drawing a zero at x=0, is a flat horizontal line until x=1, jumps up to the probability of drawing a zero or a one at the point x=1, and so forth.

Part C: For discrete distributions, the probability mass function (pdf) at x is the probability of drawing a value x from the distribution.

Part D: For continuous distributions, the probability of drawing any value is zero. The probability density function is defined from the CDF. The probability of drawing a value between x and $x + \partial x$ is the CDF at $x + \partial x$ minus the CDF at x. The pdf is defined so that this value equals the integral of the pdf from x to $x + \partial x$, so pdf(x) is the limit as $\partial x \rightarrow 0$ of

$$(\text{CDF}(x + \partial x) - \text{CDF}(x - \partial x)) / (2 \times \partial x)$$

Intuition: The likelihood of a random variable at the point x_0 is the probability of a random draw being in a small interval around x_0 divided by the size of that interval as the size of that interval approaches zero.

Part E: The quantile function is the inverse of the CDF. The CDF is defined on the real numbers and takes values from 0 to 1. The quantile function is defined on the interval [0.1] and takes values from the range of the CDF.

Illustration: Consider the standard normal distribution. Its range is the real numbers, with CDF($-\infty$) = 0 and CDF($+\infty$) = 1, so the quantile of 0 is $-\infty$ and the quantile of 1 is $+\infty$. The standard normal distribution is symmetric about zero, so the quantile of 50% is zero. CDF(97.5%) = 1.96, so the quantile of 1.96 is 97.5%.

Part F: We compare the shapes of distributions by comparing their quantiles. To compare the shapes of distributions, we must adjust for location and scale. Location is the mean of the distribution. For example, $N(0,1^2)$ and $N(10,1^2)$ are the same distribution, but $N(10,1^2)$ is shifted 10 units to the right of $N(0,1^2)$.

The scale is the standard deviation of the distribution. $N(0,1^2)$ and $N(0,2^2)$ are both normal distributions, but $N(0,2^2)$ is stretched out. If we change the units of measurement to make each unit half as large, $N(0,2^2)$ looks like $N(0,1^2)$.

The textbook explains on page 185:

> When $X \sim N(\mu, \sigma^2)$, probabilities involving X are computed by "standardizing." The standardized variable is $(X - \mu) / \sigma$. Subtracting μ shifts the mean from μ to zero, and then dividing by σ scales the variable so that the standard deviation is 1 rather than σ.

Distributions with different shapes look different no matter the location or the units of measurement. Different distributions (normal, uniform, Poisson, lognormal) have different shapes. Even distributions of the same type, such as Gamma distributions and *F* distributions, may have different shapes depending on the parameters.

This statistics course deals with standardized variables and standardized residuals, which are tested on the final exam. A normal probability plot tests the shape of the distribution.

Part G: The textbook explains heavy- vs light-tailed as

- heavy-tailed: the density curve declines less rapidly out in the tails than does a normal curve
- light-tailed: the density curve declines more rapidly out in the tails than does a normal curve

Part H: An intuitive explanation is that a distribution is skewed if one of its tails is longer than the other.

- A positively skewed distribution has a longer tail in the positive direction.
- A negatively skewed distribution has a longer tail in the negative direction.

This explanation is not a rigorous definition; many skewed distributions have tails that extend to $-\infty$ and $+\infty$, but the probability density function drops off more slowly in one of the tails.

Exercise 1.2: Normal probability plots

A statistician forms a normal probability plot.

A. What does the horizontal axis of the plot represent?
B. What does the vertical axis of the plot represent?
C. What do we infer from points that lie above or below the comparison line in the tails of the distribution?

What type of empirical distribution is implied by

D. Points that lie above the comparison line in both tails.
E. Points that below the comparison line in both tails.
F. Points that lie above the comparison line in the upper tail and below the comparison line in the lower tail.
G. Points that lie below the comparison line in the upper tail and above the comparison line in the lower tail.

Part A: The horizontal axis represents the theoretical quantiles. The theoretical quantiles are those of the type of distribution to which the sample is being compared.

Illustration: The cumulative distribution function plots quantiles on the vertical axis against the values of the distribution on the horizontal axis. The quantiles range from zero to one. The normal probability plots assume the quantiles are for a normal distribution.

Question: What do you mean by the *type of distribution* to which the sample is being compared?

Answer: The type of distribution means normal, lognormal, uniform, binomial, Poisson, χ^2, and so forth, no the values of the mean or variance for a particular distribution. Probability plots can be formed for any distribution, but the final exam problems test only normal probability plots.

Part B: The vertical axis represents the sample quantiles.

Question: How do we form sample quantiles?

Answer: Suppose the sample has N observations. We first rank the observations by size. For example, if the sample has the five points 3, –1, 0, 11, 2, we rank them as –1, 0, 2, 3, 11.

We divide the cumulative distribution function into N equal bands: 0 to 1/N, 1/N to 2/N, …, (N-1)/N to 1. For example, if the sample has five observations, we use five bands: 0% to 20%, 20% to 40%, 40% to 60%, 60% to 80%, and 80% to 100%.

The quantiles for the observations are the midpoints of the bands.

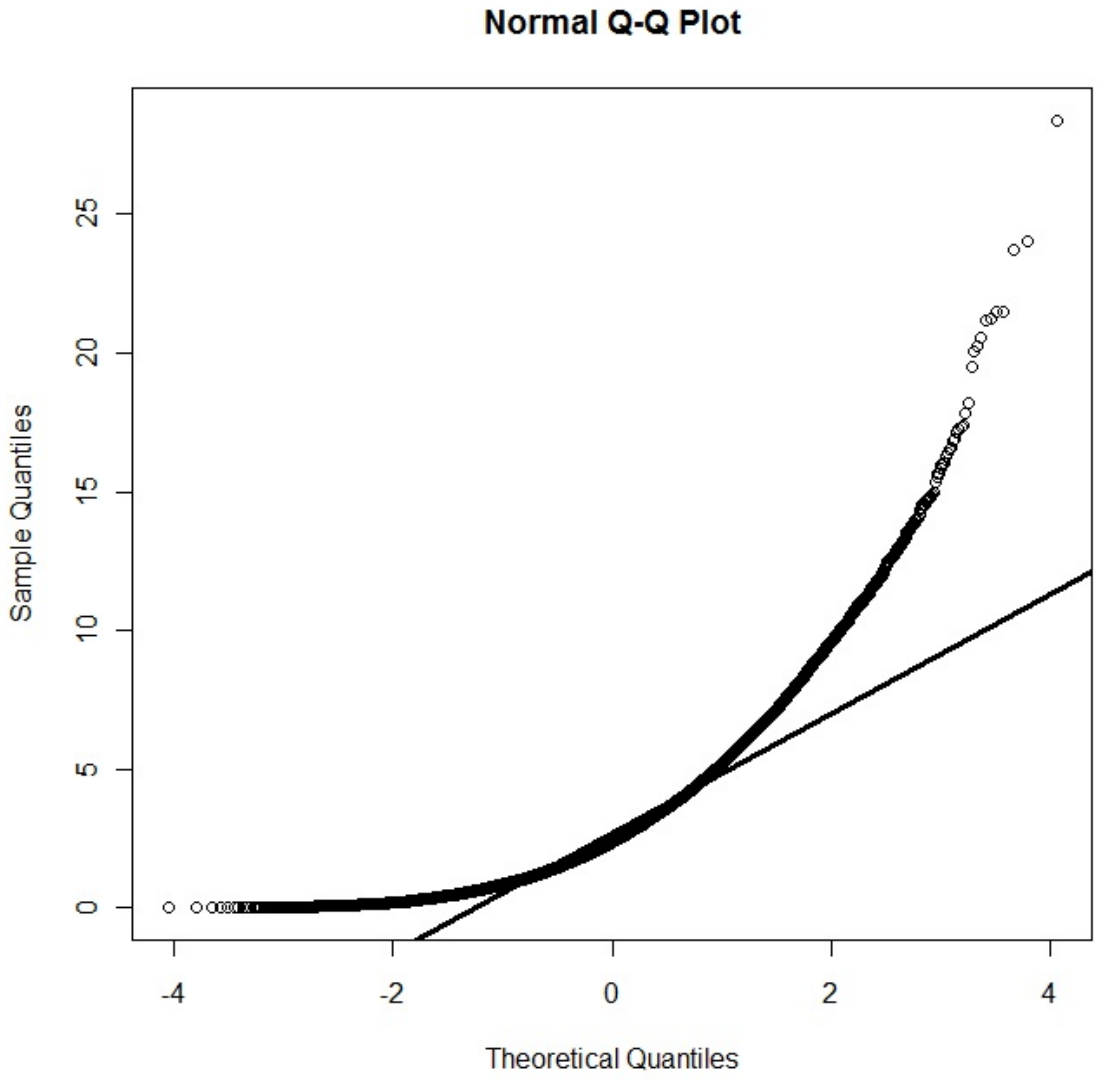
- The smallest observation is the quantile 1/2N.
- The next smallest observation is the quantile 3/(2N).
- The largest observation is the quantile (2N-1)/(2N).

If the sample has five observations, the five quantiles are 10%, 30%, 50%, 70%, and 90%. The formula is (i – 0.5) / N, where *i* is the ordinal value of the observation.

For large samples, this procedure works well. For small samples, it is not exact. The textbook mentions better approximations used by some statistical packages. Final exam problems do not test the other approximations.

Part C: The normal probability plot is a scatterplot of the sample quantiles against the theoretical quantiles.

Part D: Positively skewed distributions have points that lie above the comparison line in both tails. Below is a normal probability plot from a χ-squared distribution with three degrees of freedom.



Part E: Negatively skewed distributions have points that below the comparison line in both tails.

Part F: Heavy tailed distributions have points that lie above the comparison line in the upper tail and below the comparison line in the lower tail.

Part G: Light tailed distributions have points that lie below the comparison line in the upper tail and above the comparison line in the lower tail.