Chapter 1: Preliminary

Suppose that X is the random variable (rv, in short) of our interest in a random experiment. Since there exists an uncertainty in the experiment, or a randomness of X, we can only describe X with some probabilisitic statements, say $P(1.2 < X \leq 3.5) = 0.75$ or $P(X > 3) = 0.1$. To find the probability about X, its distribution must be known. However, in practice, we have never known or can just partially known it. To get the information about the distribution of X, we collect the data/ actual values/ observations/ realizations of X because they contain some information about the distribution of X.

1 What is Statistical Inference?

Statistical inference is a statistical process that we investigate how to use the information from data to make an inference about the distribution of the random variable of our interest.

Throughout this course, we will focus on two core concepts of statistical inference:

i. Point Estimation, and
ii. Hypothesis Testing.

1.1 Random Sample

To make a statistical inference about the distribution of X, we need to draw a sample of data. So, how the sample should be collected? Of particular importance is the so-called a **random sample** (rs, inshort).

Definition (Random Sample): A random sample is a set of random variables which are independent and identically distributed, which is often shortened to **i.i.d.**

Under random sampling, each X_i, the i^{th} copy of X for $i = 1, \dots, n$, has the same probability distribution (described as f_X or p_X) as X, , where n represents the sample size. After sampling, the realizations of each X_i are known, denoted by x_i, for $i = 1, \dots, n$.

It is obvious to see that under random sampling the joint pdf of a rs of a continuous rv X with a common distribution given by f_X is

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} f_X(x_i).$$

Similarly, we have the joint pmf $p_{X_1,\dots,X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} p_X(x_i)$ for a discrete rv X.

1.2 PARAMETRIC DISTRIBUTION

One of the central problems in statistics is **where the data come from**, i.e. what is the distribution of X. Practically speaking, the distribution is never known, but we can estimate it statistically. One common way of doing the estimation is to assume that it belongs to a parametric family.

If a distribution has a given form of pdf/pmf with an unknown parameter(s), then it is said to be **a parametric distribution**, or we can say that the distribution belongs to a parametric family with a pdf in $\{f_X: f_X = f_X(\cdot\,|\theta)\}$ or a pmf in $\{p_X: p_X = p_X(\cdot\,|\theta)\}$, where the parameter θ is unknown. The typical examples of a parametric distribution are

- Normal distribution: $\{f_X: f_X = f_X(\cdot\,|\theta)\}$, where $\theta = \binom{\mu}{\sigma^2}$ is unknown.
- Binomial distribution: $\{p_X: p_X = p_X(\cdot\,|\theta)\}$, where $\theta = p$ is unknown.

Throughout this course, we consider all statistical problems with parametric distributions, and the parameter is assumed to be UNKNOWN.

Why do we focus on parametric distributions?

Under the parametric setting, we can see that all unknown of the distribution ONLY comes from the unknown of the parameter(s). In other words, if the parameter is known, then the parametric distribution will also be completed known. Therefore, we can **reduce the problem of estimating the distribution** (a function f_X or p_X) **into an estimation problem of the parameter(s)** (a number(s) θ), like μ and σ^2 for a normal distribution, and λ for a Poisson distribution.

How do we estimate a parameter?

There is no any better way than the one using the "information" from the distribution itself to estimate its unknown parameter. However, how do we get this information about the distribution? The answer is "**DATA**".

What should we do next? We need **a statistic**!

Definition (Statistic): If $\boldsymbol{X} = (X_1, \dots, X_n)^T$ represents a set of random variables and $T(\cdot)$ is a real-valued (or vector-valued) function such that for all $\boldsymbol{X}$, $T(\boldsymbol{X})$ does not contain any unknown terms. Then, we would say that T is *a statistic*.

Example: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are statistics. However, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ is not a statistic unless the value of μ is known.

In particular, when a statistic $T(\boldsymbol{X})$ is used to estimate θ, we would call it ***an estimator of*** θ. When the actual values $\boldsymbol{x} = (x_1, \dots, x_n)^T$ of the random variables, i.e. data, are used, we would have the actual value of the statistic ---- ***an estimate***.

Note that we would also commonly use the notation $\hat{\theta}$ as an estimate/estimator of θ.

What's next?

Under the parametric setting, our central problem becomes how to determine which estimator will be the ***best*** one to estimate θ. This will be formulated in more details in Chapter 2.

2 MULTIVARIATE NORMAL DISTRIBUTION AND ITS PROPERTIES

Recall that the univariate normal distribution with mean $\mu \in (-\infty, \infty)$ and variance $\sigma^2 \in (0, \infty)$ has a pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{(x-\mu)}{\sigma}\right]^2},$$

for $x \in (-\infty, \infty)$.

Proposition 1: If $X \sim N(\mu, \sigma)$, then $\frac{(X-\mu)^2}{\sigma^2} \sim \chi^2(1)$, the χ^2distribution with 1 degree of freedom.

Proposition 2: If $Y_i \sim \chi^2(r_i)$ for $i = 1, \dots, k$, where $r_1, \dots, r_k$ are any positive integers, and $Y_1, \dots, Y_k$ are independent, then $\sum_{i=1}^{k} Y_i \sim \chi^2(\sum_{i=1}^{k} r_i)$.

Note that the term $\left[\frac{(x-\mu)}{\sigma}\right]^2$ in the exponent of the pdf above indeed measures the square distance between x and μ in unit standard deviation. This can be generalized for a $p \times 1$ vector $\boldsymbol{x} = (x_1, \dots, x_p)^T$ --- an observation of a random vector $\boldsymbol{X} = (X_1, \dots, X_p)^T \in R^p$--- as $(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$, where $\boldsymbol{\mu} = E(\boldsymbol{X})$ and Σ is a positive-definite variance-covariance matrix of $\boldsymbol{X}$ (i.e. for any non-zero vector $\boldsymbol{a}$ with real entries, $\boldsymbol{a}^T \Sigma^{-1} \boldsymbol{a} > 0$).

Correspondingly, we have a p-dimensional multivariate normal distribution of $\boldsymbol{X}$ with its pdf

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})},$$

denoted by $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$.

Proposition 3: Let $\binom{X_1}{X_2} \sim N_2\left(\binom{\mu_1}{\mu_2}, \begin{pmatrix}\sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22}\end{pmatrix}\right)$, then X_1 and X_2 are independent if and only if $\sigma_{12} = 0$ (or $\sigma_{21} = 0$).

Lemma 1: If $\boldsymbol{X} = (X_1, \dots, X_p)^T \sim N_p(\boldsymbol{\mu}, \Sigma)$, then for any $q \times p$ matrix A, we have

$$A\boldsymbol{X} \sim N_q(A\boldsymbol{\mu}, A\Sigma A^T).$$

Lemma 2: If $X_1 \sim N(\mu_1, \sigma_{11})$, $X_2 \sim N(\mu_2, \sigma_{22})$, and they are independent, then

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{pmatrix}\right).$$

Theorem 1: Consider a random sample $\{X_1, \dots, X_n\}$ from $N(\mu, \sigma^2)$ with $n > 1$. Then,

(i) $\bar{X}_n \sim N(\mu, \ \sigma^2/n)$.

(ii) $\bar{X}_n$ and S_n^2 are independent.

(iii) $\frac{(n-1)S_{n-1}^2}{\sigma^2} = \frac{nS_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$.

Proofs of (i) and (ii): Using Lemma 2, we can easily get the result that

$$\boldsymbol{X} = (X_1, \dots, X_n)^T \sim N_n(\boldsymbol{\mu}, \sigma^2 I_{n\times n}),$$

where $\boldsymbol{\mu} = (\mu, \dots, \mu)^T$ is a $n \times 1$ vector and $I_{n\times n}$ is a $n \times n$ identity matrix.

For the first result, set $A = (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$ to be a $1 \times n$ vector. Thus, $A\mathbf{X} = \overline{X}$. Then by Lemma 1, we have

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

For the second result, for any $i = 1, 2, \dots, n$, set $A = \begin{pmatrix} \frac{1}{n} & \dots & \frac{1}{n} & \dots & \frac{1}{n} \\ -\frac{1}{n} & \dots & 1-\frac{1}{n} & \dots & -\frac{1}{n} \end{pmatrix}$ to be a $2 \times n$ matrix, where the $(2,i)^{th}$ element is $1 - \frac{1}{n}$, and $(2,j)^{th}$ elements are $-\frac{1}{n}$, for $j = 1, 2, \dots, i-1, i+1, \dots, n$. Thus, $A\mathbf{X} = \begin{pmatrix} \overline{X} \\ X_i - \overline{X} \end{pmatrix}$. Then by Lemma 1, we have

$$\begin{pmatrix} \overline{X} \\ X_i - \overline{X} \end{pmatrix} \sim N_2\left(A\boldsymbol{\mu}, \sigma^2 A I_{n\times n} A'\right), \qquad (2)$$

where $\sigma^2 A I_{n\times n} A' = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{n-1}{n}\sigma^2 \end{pmatrix}$. Therefore, by Lemma 2, $\overline{X}$ and $X_i - \overline{X}$ are independent for any $i = 1, \dots, n$, which implies that $\overline{X}$ and S_n^2 are also independent because S_n^2 is a function of $X_1 - \overline{X}, X_2 - \overline{X}, \dots, X_n - \overline{X}$.

Two common distributions defined by a normal distribution:

In statistical inference, there are two distributions which is as common as normal distributions. They are t and F distributions. Here we can have their formal definitions and their related results.

Definition (t distribution): If $Z \sim N(0,1)$, $U \sim \chi^2(r)$, and they are independent, then the distribution of

$$\frac{Z}{\sqrt{U/r}}$$

is called a t distribution with r degrees of freedom, denoted by $t(r)$.

Note that the pdf of $T \sim t(r)$ can be shown to be

$$f_T(t) = \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\,\Gamma(r/2)}(1 + t^2/r)^{-(r+1)/2},$$

for $t \in (-\infty, \infty)$. As $r \to \infty$, f_T tends to a pdf of N(0,1). In R, the function **qt** can be used to find a quantile of t distribution.

Lemma 3: Consider a random sample $\{X_1, \ldots, X_n\}$ from $N(\mu, \sigma^2)$ with $n > 1$. Then,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S_{n-1}} \sim t(n-1).$$

Note that a t distribution is more spread than a standard normal distribution. In Lemma 3, we can see that the denominator is S_{n-1} rather than σ. Since S_{n-1} is a random variable changing with different samples, the variability in t is more, thus resulting in more spread.

Definition (*F* distribution): If $X \sim \chi^2(r_1)$, $Y \sim \chi^2(r_2)$, and they are independent, then the distribution of

$$\frac{X/r_1}{Y/r_2}$$

is called a *F* distribution with r_1 and r_2 degrees of freedom, denoted by $F(r_1, r_2)$.

Note that the pdf of $F \sim F(r_1, r_2)$ can be shown to be

$$f_F(w) = \frac{\Gamma[(r_1 + r_2)/2]}{\Gamma(r_1/2)\,\Gamma(r_2/2)} \left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}} w^{\frac{r_1}{2}-1}(1 + r_1 w/r_2)^{-(r_1+r_2)/2},$$

for $w \in (0, \infty)$. In R, the function **qf** can be used to find a quantile of F distribution.

Lemma 4: Consider two independent random samples $\{X_1, \ldots, X_n\}$ from $N(\mu_X, \sigma_X^2)$ and $\{Y_1, \ldots, Y_m\}$ from $N(\mu_Y, \sigma_Y^2)$ with $n > 1$ and $m > 1$. Then,

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(n-1, m-1),$$

where $S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ and $S_Y^2 = \frac{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}{m-1}$.

3 POPULATION AND SAMPLE MOMENTS

In this section, we would study one particular estimator for an unknown population quantity --- population moment.

Definition (Population moments): For a positive integer k,

(i) the k^{th} population moment (about 0) of X is denoted by μ_k' and defined as

$$\mu_k' = E(X^k) \text{ if it exists.}$$

Note that μ_1' is the population mean μ of X.

(i) the k^{th} population CENTRAL moment of X is denoted by μ_k and defined as

$$\mu_k = E(X-\mu)^k \text{ if it exists.}$$

Note that $\mu_1 = 0$. For population variance, we have $\sigma^2 = E(X-\mu)^2 = \mu_2$.

For the sake of convenience, we define $\sigma^k = (\sigma^2)^{k/2}$ for any positive integer k.
Caution: Except for $k = 2$, $\sigma^k \neq \mu_k$.

Some useful population moments used to describe a distribution

In the following, we comment on how the first four moments of a random variable are used as measures of various characteristics of the corresponding density (or distribution).

1. Mean: μ_1', a measure of the central tendency or location of the density of a random variable.
2. Variance: $\sigma^2 = \mu_2$, a measure of the dispersion of the density of a random variable.
3. Skewness: μ_3 (the 3^{rd} central moment), a measure of asymmetry or skewness. Symmetrical distributions can be shown to have $\mu_3 = 0$. If a density has a tail to the left (or right), i.e. the curve of the density function is skewed to the left (or right), then it can be shown to have $\mu_3 < 0$ (or $\mu_3 > 0$). The ratio μ_3/σ^3, which is unitless, is called the *coefficient of skewness*.
4. Kurtosis: μ_4 (the 4^{th} central moment) is used to compare the tail behavior of the distribution with a normal distribution. Note that kurtosis tells us nothing about the peakedness of the distribution [See Am Stat. 2014; 68(3): 191-195]. The term $\mu_4/\sigma^4 - 3$ is called the *coefficient of excess kurtosis*. The normal distribution can be shown to have $\mu_4/\sigma^4 - 3 = 0$. For the distribution having thicker (thinner) tails than a normal distribution, $\mu_4/\sigma^4 - 3 > 0$ (< 0).

Definition (Sample moments): Let $\{X_i : i = 1, \dots, n\}$ be a random sample of size n. For a positive integer k, the k^{th} sample moment about a of X is defined as

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - a)^k.$$

In particular,

(i) when $a = 0$, we have $\frac{1}{n}\sum_{i=1}^{n} X_i^k$, which is denoted by $\overline{X^k}$. Note that the sample mean $\bar{X}$ will be obtained if $k = 1$.

(ii) when $a = \bar{X}$, we have $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$, which is denoted by M_k .

Note that $M_1 = 0$. For sample variances, $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = M_2$ and $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are commonly used in practice.

For the sake of convenience, we define $S_n^k = (S_n^2)^{k/2}$ for any positive integer k.
Caution: Except for $k = 2$, $S_n^k \neq M_k$.

Lemma 5: Consider a random sample $\{X_1, \dots, X_n\}$. For any positive integer k, the population mean of the k^{th} sample moment about 0 is equal to the k^{th} population moment about 0, i.e.,

$$E\left(\overline{X^k}\right) = \mu_k' \ \ (assuming\ that\ \mu_k'\ exists).$$

Furthermore,

$$Var\left(\overline{X^k}\right) = \frac{1}{n}[\mu_{2k}' - (\mu_k')^2] \ \ (assuming\ that\ \mu_{2k}'\ exists).$$

3.1 Moment Generating Functions

Normally, there is no particular way to find population moments. However, if the moment generating function (mgf) of a random variable exists, then we can get the result easily.

Definition (Moment generating function): The moment generating function (mgf) of a random variable X is denoted by $M_X(t)$ and is defined as $M_X(t) = E(e^{tX})$, if the expectation exists for t in a neighborhood of 0. To be more precise, there is a positive h such that, for all t in $(-h, h)$, $E(e^{tX})$ exists.

More explicitly, we write the mgf of X as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

if the random variable X is continuous and $f(\cdot)$ is the pdf of X, or

$$M_X(t) = \sum_x e^{tx} p(x)$$

if X is discrete and $p(\cdot)$ is the pmf of X.

Note that the mgf of X does not always exist. However, if it exists, then $M_X(t)$ is continuously differentiable in some neighborhood of the origin. If we differentiate the mgf k times with respect to t, then we have

$$\frac{d^k}{dt^k} M_X(t) = \int_{-\infty}^{\infty} x^k e^{tx} f(x)\, dx,$$

and letting $t = 0$, we find

$$\frac{d^k}{dt^k} M_X(0) = E(X^k) = \mu_k'.$$

That is, the k^{th} moment of X is equal to the k^{th} derivative of $M_X(t)$ evaluated at $t = 0$.

Remark that if we replace e^{tX} by its series expansion in (3), we obtain the series expansion of $M_X(t)$ in terms of the moments of X; thus

$$M_X(t) = E\Big[\sum_{i=0}^{\infty} \frac{(tX)^i}{i!}\Big] = \sum_{i=0}^{\infty} \frac{t^i}{i!} E(X^i) = \sum_{i=0}^{\infty} \frac{t^i}{i!} \mu_i'$$

In addition, mgf can also be used to determine the distribution of a random variable. This property can lead to some extremely powerful results when used properly, and it is shown by the following theorem without proofs.

Theorem 2: Let X and Y be two random variables. Suppose that their mgfs, $M_X(t)$ and $M_Y(t)$, both exist and are equal for all t in $(-h, h)$ for some $h > 0$. Then, the distributions of X and Y are equal.

This theorem is particularly useful when the distribution of an independent sum of random variables has to be determined.

For instance, we can use it to prove Propositions 1 and 2, and (iii) of Theorem 1.

For the third result, since $\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n (X_i - \mu + \mu - \overline{X})^2 = \sum_{i=1}^n (X_i-\mu)^2 - n(\overline{X}-\mu)^2$, $\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2} = n\frac{S_n^2}{\sigma^2} + n(\frac{\overline{X}-\mu}{\sigma})^2$. Note that $\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2}$ and $n(\frac{\overline{X}-\mu}{\sigma})^2$ have a $\chi^2(n)$ and $\chi^2(1)$ distribution, respectively, and by the second result, $n\frac{S_n^2}{\sigma^2}$ and $n(\frac{\overline{X}-\mu}{\sigma})^2$ are independent. Thus, for $t < 1/2$,

$$\begin{aligned}(1-2t)^{-n/2} &= M_{\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2}}(t) \\ &= M_{n\frac{S_n^2}{\sigma^2}}(t) M_{n(\frac{\overline{X}-\mu}{\sigma})^2}(t) \\ &= M_{n\frac{S_n^2}{\sigma^2}}(t) \times (1-2t)^{-1/2}\end{aligned}$$

Therefore, we have

$$M_{n\frac{S_n^2}{\sigma^2}}(t) = (1-2t)^{-(n-1)/2},$$

a mgf of a $\chi^2(n-1)$ distribution. By the uniqueness of mgf, we have

$$n\frac{S_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

The following result even indicates how the mgf can be used in the problems of limiting distributions.

Theorem 3: Suppose that $\{X_n: n = 1,2, \dots\}$ is a sequence of random variables, each with an existing mgf $M_{X_n}(t)$, and $\lim_{n\to\infty} M_{X_n}(t) = M_Y(t)$, for all t in a neighborhood of 0, where $M_Y(t)$ is an mgf of Y. Then, the limiting distribution of X_n is equal to the distribution function of Y, i.e.

$$\lim_{n\to\infty} F_{X_n}(y) = F_Y(y), \qquad for\ all\ y,$$

where $F_Y(y)$ is the distribution function of Y and is continuous at y.

Remark that

- The proofs of Theorem 2 and 3 reply on the (uniqueness) theory of Laplace transforms. So, the proofs omit here.
- Theorem 3 is commonly used for the proofs of the Poisson approximation to the Binomial distribution and the Central Limit Theorem.

3.2 CENTRAL LIMIT THEOREM

The central limit theorem (CLT) is one of the most important and powerful theorems in both statistics and probability. It gives us an approximated/limiting distribution of sample mean $\bar{X}_n$ without any distribution assumption (other than independence and finite variance) of the sample.

Theorem 4 (Central Limit Theorem, standard version): Suppose that $\{X_n : n = 1,2, \dots\}$ is a sequence of i.i.d. random variables with a positive variance, each with an existing mgf $M_{X_n}(t)$, for all t in a neighborhood of 0. Denote by $\bar{X}_n$ the sample mean of $X_1, \dots, X_n$. Then, the limiting distribution of

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

is a standard normal distribution, where μ is the common population mean of $X_1, \dots, X_n$.

Note that both μ and σ^2 are finite since the mgfs exist.

Taylor's Theorem: Consider a function g, where its derivative of order r at a exists, i.e. $g^{(r)}(a)$ exists. Then, $\lim_{x \to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0$, where $T_r(x) = \sum_{k=0}^{r} \left[\frac{g^{(k)}(a)}{k!} (x-a)^k \right]$. **Proof omitted.**

Proof of Theorem 4:

Proof of Theorem 4 (cont'd)

Question: Why is CLT so important in statistics?

Answer: CLT provides us with large-sample solutions to problems in statistical inference, when the finite-sample solutions cannot be found easily.

4 LIMIT THEOREM

In mathematical analysis, for a sequence of real numbers $\{x_n: n = 1,2, \dots, \}$, we have

$$x_n \text{ converges to } x \leftrightarrow \lim_{n\to\infty} x_n = x$$

In probability theory, we can also define the convergence of a sequence of random variables $\{X_n: n = 1,2, \dots, \}$, so that we can say something about the limiting behavior of the random variables.

Unlike the mathematical analysis, there are several modes of convergence in probability. In the following, we study two common modes of convergence for this course.

4.1 CONVERGENCE IN PROBABILITY

(A sequence of) X_n converges in probability to a random variable X if for every $\epsilon > 0$,

$$P(|X_n - X| < \epsilon) \to 1 \text{ (or equivalently, } P(|X_n - X| \geq \epsilon) \to 0)$$

Notationally, we have $X_n \xrightarrow{p} X$.

One well-known result in probability and statistics, weak law of large numbers (WLLN), tells us that $\bar{X}_n$ converges in probability to μ. Details will be provided later.

4.2 CONVERGENCE IN DISTRIBUTION

(A sequence of) X_n converges in distribution to a random variable X if

$$F_{X_n}(t) \to F_X(t) \text{ for all continuity points } t \text{ of } F_X,$$

where F_{X_n} and F_X are the respective cdfs of X_n and X.

Notationally, we have $X_n \xrightarrow{d} X$.

Thus, CLT indeed tells us the result that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1).$$

4.3 Relationship between convergences in probability and in distribution

In gerenal,

$$X_n \xrightarrow{p} X \text{ implies } X_n \xrightarrow{d} X.$$

However, the converse is not true, except that X is a constant. That is,

$$X_n \xrightarrow{d} c \text{ implies } X_n \xrightarrow{p} c.$$

4.4 Weak Law of Large Numbers (WLLN)

Lemma 6: Consider a r.s. $\{X_i : i = 1, \dots, n\}$ with a finite common population mean μ and variance σ^2. Then, $\bar{X}_n \xrightarrow{p} \mu$.

We need Chebyshev's inequality to prove the above result.

Chebyshev's inequality: For any random variable Y with $\mu_Y = E(Y) < \infty$, then for any $k > 0$, we have $P(|Y - \mu_Y| \geq k) \leq \frac{\sigma_Y^2}{k^2}$.

Proof of Chebyshev's inequality: For simplicity, consider continuous cases. Then, we have

$$\sigma_Y^2 = E[Y - \mu_Y]^2 = \int_{\{y:\, |y-\mu_Y| \geq k\}} [|y - \mu_Y|^2 f_Y(y)]dy + \int_{\{y:\, |y-\mu_Y| < k\}} [|y - \mu_Y|^2 f_Y(y)]dy$$

$$\geq \int_{\{y:\, |y-\mu_Y| \geq k\}} [|y - \mu_Y|^2 f_Y(y)]dy \; + \; 0$$

$$\geq k^2 \int_{\{y:\, |y-\mu_Y| \geq k\}} [f_Y(y)]dy \; = k^2 \, P(|Y - \mu_Y| \geq k).$$

Proof of WLLN: By Chebyshev's inequality, we can see that for every $\epsilon > 0$,

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0.$$

Recall that if a r.s. of X with size n is considered and $X \sim N(\mu, \sigma^2)$, then we have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1).$$

What if we remove the normality assumption? The exact distribution of $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ cannot be found easily, but by CLT we know that as $n \to \infty$, we have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1).$$

Note that practical speaking, the CLT alone now is NOT powerful enough to deal with most general cases. For instance, if σ is unknown in the above case, then CLT fails to construct a confidence interval for μ or to formula a test statement for a hypothesis of the true value of μ.

Thus, we need some add-on theorems. They are **Slutsky's Theorem** and **Delta method**.

4.5 Slutsky's Theorem (proof omitted)

If $Y_n \xrightarrow{d} V$ and $W_n \xrightarrow{p} c$, then

(i) $Y_n \pm W_n \xrightarrow{d} V \pm c$

(ii) $W_n Y_n \xrightarrow{d} cV$

(iii) $\frac{Y_n}{W_n} \xrightarrow{d} \frac{V}{c}$ if $c \neq 0$.

Note that if we can show the convergence of S_{n-1} to σ in probability, then by Slutsky's theorem (ii) with CLT, we can get the limiting distribution of $\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_{n-1}}$.

Convergence of S_{n-1} to σ in probability

Consider a r.s. $\{X_i : i = 1,2,\dots,n\}$ with a finite common population mean μ and variance σ^2. By Chebyshev's inequality, we first have

$$P(|S_{n-1}^2 - \sigma^2| \geq \epsilon) \leq \frac{Var(S_{n-1}^2)}{\epsilon^2}.$$

If $Var(S_{n-1}^2) \to 0$ as $n \to \infty$, then $S_{n-1}^2 \xrightarrow{p} \sigma^2$.

According to the following result in Assignment 1,

> 3. Consider a r.s. $\{X_1, X_2, \dots, X_n\}$ of size $n > 1$ from a distribution with mean μ and variance σ^2. We have already known that S_{n-1}^2 has a mean σ^2. Here, we look at its variance. Please show that the variance of S_{n-1}^2 is
>
> $$\frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right).$$

We only need a finite fourth moment about μ, i.e. $\mu_4 < \infty$.

Continuous Mapping Theorem: if $Y_n \xrightarrow{p} c$ and h is a continuous function, then $h(Y_n) \xrightarrow{p} h(c)$.

Proof omitted.

Under the assumption of finite μ_4, we have $S_{n-1}^2 \xrightarrow{p} \sigma^2$. Then, we can use **Continuous Mapping Theorem** to get $S_{n-1} \xrightarrow{p} \sigma$. Similarly, we also have $\frac{1}{S_{n-1}} \xrightarrow{p} \frac{1}{\sigma}$ or $\frac{\sigma}{S_{n-1}} \xrightarrow{p} 1$. Therefore, by CLT with Slutsky's theorem (ii), we have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_{n-1}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\frac{\sigma}{S_{n-1}} \xrightarrow{d} Z \sim N(0,1).$$

This result can then be used to construct a confidence interval for μ or to formula a test statement for a hypothesis of the true value of μ when a normality assumption is not used and σ is unknown (for large n).

4.6 DELTA METHOD

Another add-on method with CLT is the **delta method**. It is a very powerful statistical approach when we want to get a limiting distribution of a function of an estimator, like $1/\bar{X}_n$.

Theorem 5 (Delta Method): Let $\{Y_n : n = 1,2, \dots\}$ be a sequence of random variables that for a constant a and a positive constant b,

$$\sqrt{n}(Y_n - a) \xrightarrow{d} N(0, b^2), \qquad as\ n \to \infty.$$

Then, for any given function g that $g'(a)$ exists and $g'(a) \neq 0$, we have

$$\sqrt{n}\big(g(Y_n) - g(a)\big) \xrightarrow{d} N(0, [g'(a)b]^2), \qquad as\ n \to \infty.$$

Proof (outline): The 1st order Taylor approximation of g about the point a, i.e.

$g(Y_n) \approx g(a) + g'(a)(Y_n - a)$, yields $\sqrt{n}\big(g(Y_n) - g(a)\big) \xrightarrow{d} g'(a)N(0, b^2)$.

Remark that a more careful study of Taylor's approximation with remainder is required to justify all steps for the above result.

In particular, combining the delta method with CLT, we can get the following asymtotic result of a function of a sample mean.

Corollary: If $\bar{X}_n$ denotes the sample mean of a random sample of size n from a distribution with a finite mean μ and a finite positive variance σ^2. Then, for any given function g that $g'(\mu)$ exists and $g'(\mu) \neq 0$, when $\mu \neq 0$, we have $\sqrt{n}\big(g(\bar{X}_n) - g(\mu)\big) \xrightarrow{d} N(0, [g'(\mu)\sigma]^2),\ as\ n \to \infty.$

For instance, consider $g(t) = 1/t$. Then, we have

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \frac{1}{\mu}\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\mu^4}\right), \qquad as\ n \to \infty.$$

We can also use Slutsky's theorem to replace $\frac{\sigma^2}{\mu^4}$ by $\frac{S_{n-1}^2}{\bar{X}_n^4}$ if we want to make an inference for $\frac{1}{\mu}$.

Theorem 6 (2nd order Delta Method): Let $\{Y_n : n = 1,2, \dots\}$ be a sequence of random variables that for a constant a and a positive constant b,

$$\sqrt{n}(Y_n - a) \xrightarrow{d} N(0, b^2), as\ n \to \infty.$$

Then, for any given function g that $g'(a) = 0,\ g"(a)$ exists and $g"(a) \neq 0$, we have

$$n\big(g(Y_n) - g(a)\big) \xrightarrow{d} b^2 \frac{g"(a)}{2} \chi^2(1), \qquad as\ n \to \infty.$$

Proof omitted.

Note that the above result can also be written as

$$n\big(g(Y_n) - g(a)\big) \xrightarrow{d} b^2 \frac{g"(a)}{2} Gamma\left(\frac{1}{2}, \frac{1}{2}\right), \qquad as\ n \to \infty.$$

Moreover, according to the property of gamma distribution, we can further simplify the result as

$$n\big(g(Y_n) - g(a)\big) \xrightarrow{d} Gamma\left(\frac{1}{2}, \frac{1}{b^2 g"(a)}\right), \qquad as\ n \to \infty.$$

Remark that the pdf of the gamma distribution $Gamma(\alpha, \beta)$ we use here is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} I_{\{x>0\}}.$$

5 ORDER STATISTICS

In statistical inference, there are a lot of parametric distributions whose parameter(s) appears in the range of the rv, e.g. $U[0,\theta]$. Such a parameter is often estimated by **an order statistic**.

Definition (Order statistics): The order statistics of a rs $\{X_i: i = 1,2,\dots,n\}$ are the rvs whose sample values are placed in ascending order. They are denoted by $X_{(1)}, X_{(2)}, \dots, X_{(n)}$, where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$.

Remark that we usually call $X_{(i)}$ the ith smallest order statistic of the rs, and $X_{(1)}$ and $X_{(n)}$ are the minimum and maximum order statistics, respectively. Moreover, according to the above definition, it is clear to see that order statistics are NOT independent.

For simplicity, we only derive the pdf for an order statistic(s) when the rs is from a CONTINUOUS distribution.

Theorem 7: Consider order statistics $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ from a rs of size n from a continuous distribution with pdf $f_X(\cdot)$ and cdf $F_X(\cdot)$. Then, for any $i = 1, 2, \dots, n$, the pdf of $X_{(i)}$ is

$$f_{X_{(i)}}(u) = \frac{n!}{(i-1)!\ (n-i)!} f_X(u)[F_X(u)]^{i-1}[1-F_X(u)]^{n-i}\ .$$

and the joint pdf pf $X_{(i)}$ and $X_{(j)}$ for $1 \leq i < j \leq n$ is

$$f_{X_{(i)},X_{(j)}}(u,v)$$
$$= \frac{n!}{(i-1)!\ (j-i-1)!\,(n-j)!} f_X(u) f_X(v)[F_X(u)]^{i-1}[F_X(v)-F_X(u)]^{j-i-1}[1-F_X(v)]^{n-j}\ ,$$

for $u < v$, and $f_{X_{(i)},X_{(j)}}(u,v) = 0$ otherwise.

Proof of Theorem 7:

Application of Order statistics ---- Distribution Fitting

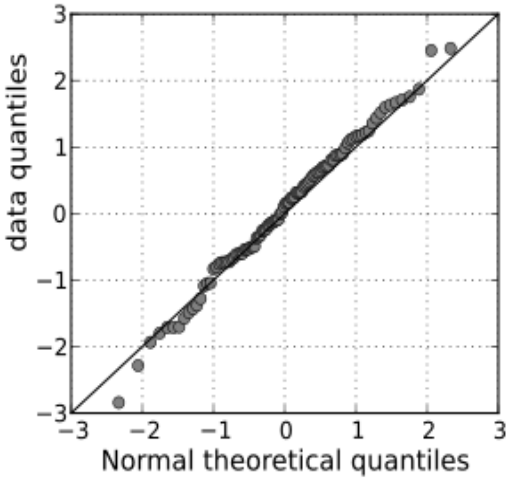
A Quantile-Quantile plot (in short, Q-Q plot) is a statistical graphical approach used to **check whether or not the collected data are drawn from a specified distribution**. The idea of the Q-Q plot is to compare the actual values of the order statistics (i.e. the sorted data) with the population quantiles of the specified distribution. If ***the normal distribution*** is used, then it is also called ***Normal Probability plot*** (or simply ***Normal plot***).

How to use the Q-Q plot?

If the points in the Q-Q plot approximately lie on the straight line, then we can say that the data are from the specified distribution.

So, **the greater the departure from the line, the greater the evidence for the conclusion that the data are not from the specified distribution**.

For ***Normal plot***, we may studentize the data first. If the points approximately lie on the straight line with slope 1 and passing through the origin, then we can say that the data are from the normal distribution.

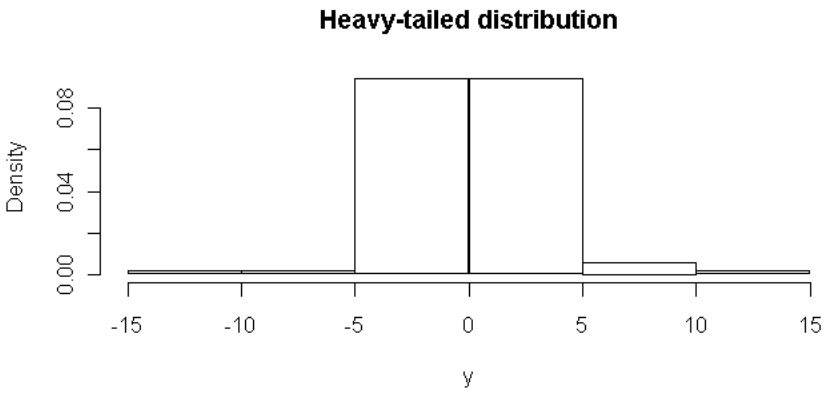


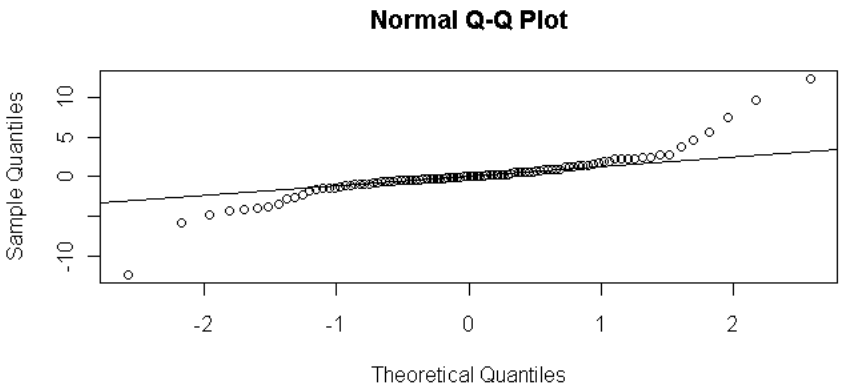
Theory behind the Q-Q plot

The theoretical framework of the Q-Q plot comes from the distribution result of the order statistics. For simplicity, here we consider the continuous case only.

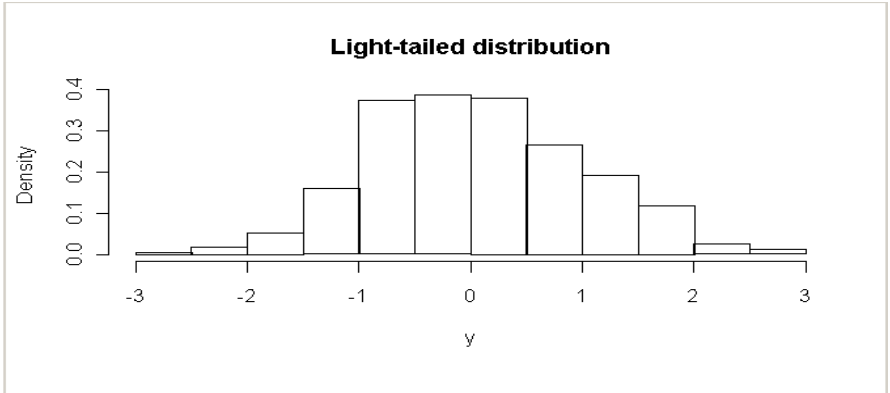
Case 1: If the r.s. is from $U[0,1]$, then find the pdf of $X_{(i)}$.

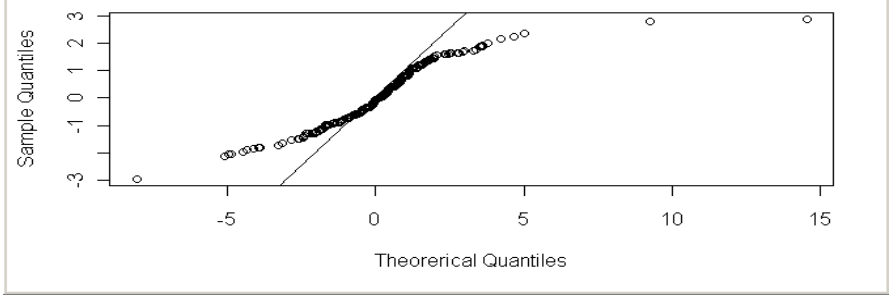
Case 2: In general, for the r.s. from a continuous distribution with cdf F_X and pdf f_X, we know that $F_X(X) \sim U[0,1]$. Then, find the expectation and variance of $F_X[X_{(i)}]$.



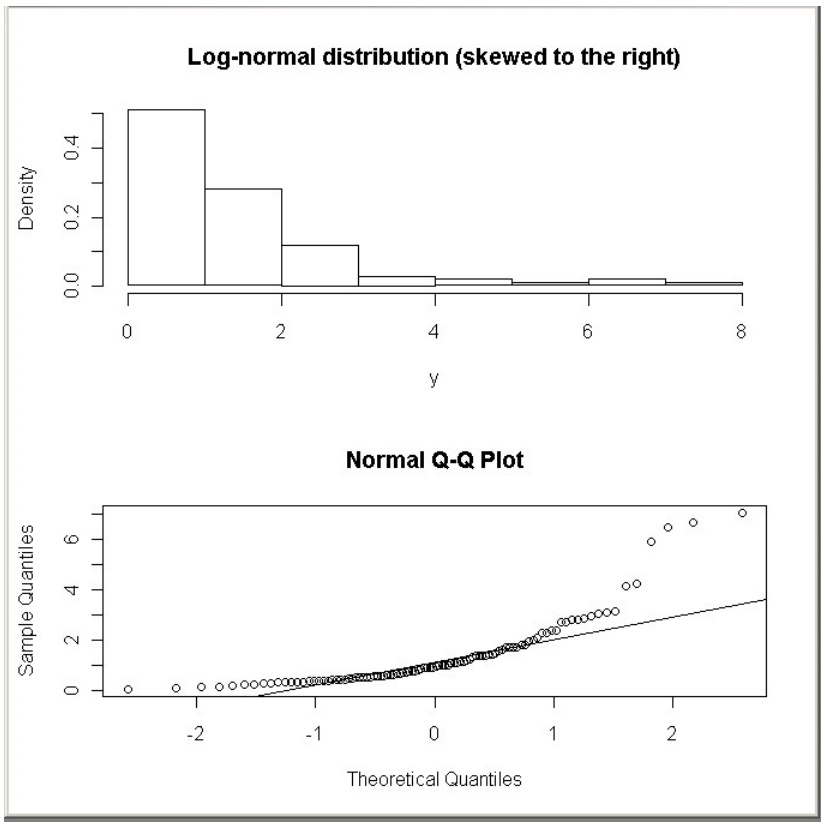


If the data are from a distribution with heavier tails than the specified distribution, then one would expect to see the upper tails of the Q-Q plot turning upwards and the lower tails bending downwards.

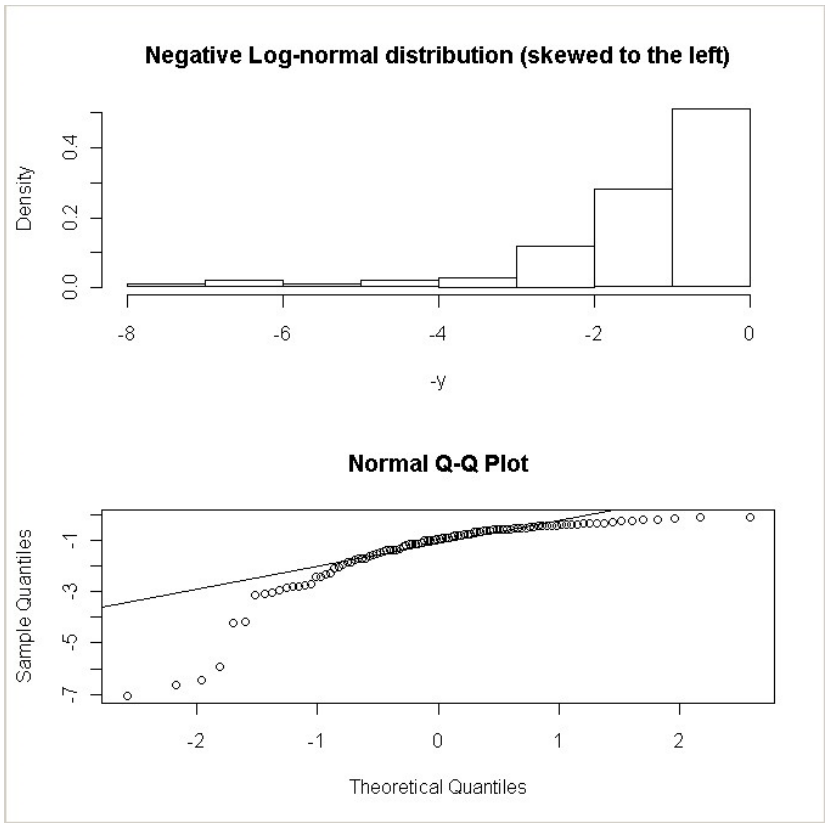




For a light-tailed distribution it would be expected to observe an S-shape with the lower tails bending upwards and the upper tail curving downwards.



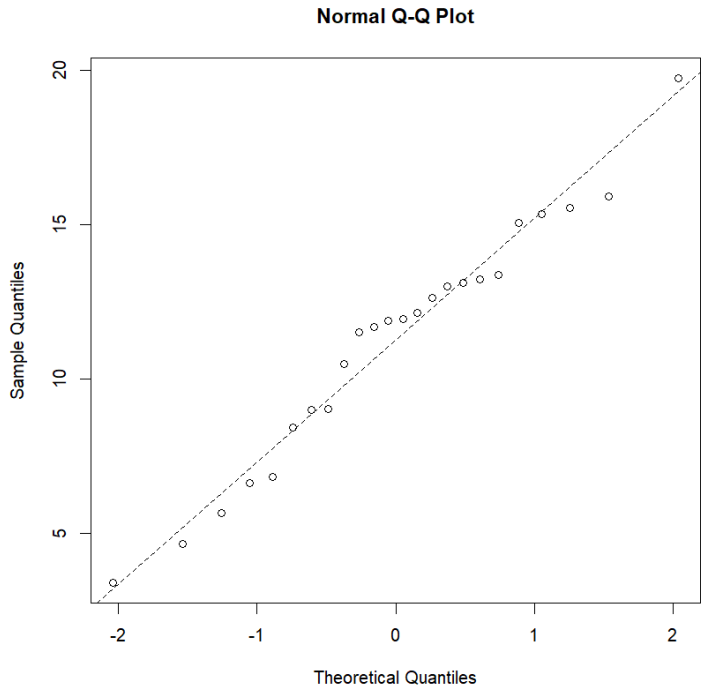
If the data are from a distribution more skewed to the right than the specified distribution, then one would expect to see both of the upper and lower tails of the Q-Q plot turning upwards.



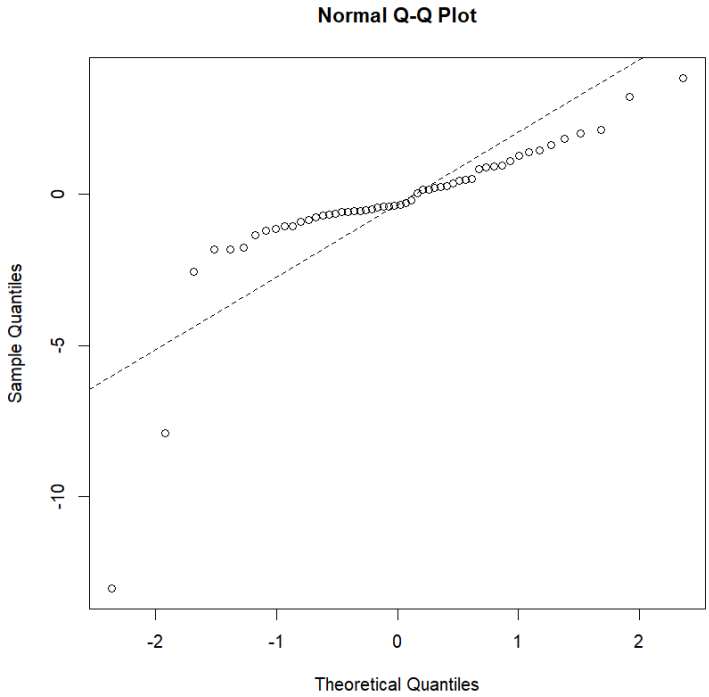
For a more negatively skewed distribution, it would be expected to observe both the upper and lower tails curving downwards.

R corner

```
x = rnorm(24, mean = 10, sd = 4)
qqnorm(x)
abline(a = mean(x), b = sd(x), lty = 2)
```




```
x = rt(54, df = 1.5)
qqnorm(x)
abline(a = mean(x), b = sd(x), lty = 2)
```



```
> x = rnorm(10, mean = 10, sd = 4)
> par(mfrow= c(1, 2))
> qqnorm(x)
> abline(a = mean(x), b = sd(x), lty = 2)
> qqline(x, col = "red")
>
> qqnorm((x-mean(x))/sd(x))
> abline(a = 0, b = 1, lty = 2)
> qqline((x-mean(x))/sd(x), col = "red")
>
```