Chapter 5: Estimation

Ott & Longnecker Sections: 4.12, 4.14 and 5.2

Duzhe Wang

the Department of Statistics, UW-Madison

Part 3

What do we study?

Key Concepts: QQ plot, central limit theorem

Outline

1. QQ plot
2. CLT

QQ plot

We usually assume that the sample are from Normal distribution, can we test this assumption?

QQ plot

We usually assume that the sample are from Normal distribution, can we test this assumption?

One easy way is using a **normal quantile-quantile plot** or **normal QQ plot**.

QQ plot

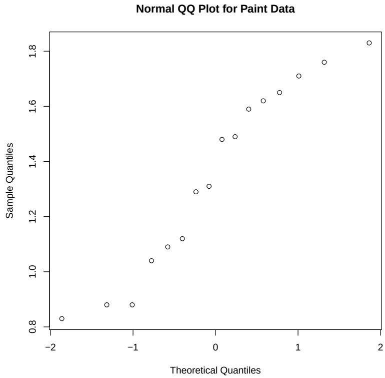
We usually assume that the sample are from Normal distribution, can we test this assumption?

One easy way is using a **normal quantile-quantile plot** or **normal QQ plot**.

If a set of observations is approximately normally distributed, a QQ plot will result in an approximately straight line.

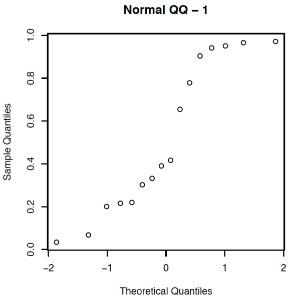
R function: qqnorm(data)

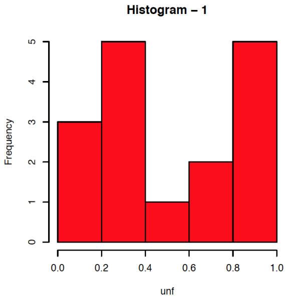
QQ plot example



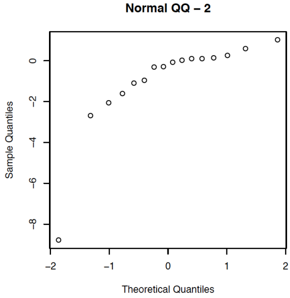
The plot is not perfectly straight, but it is pretty good.

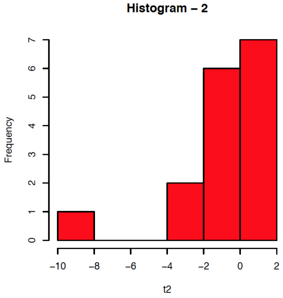
QQ plot example



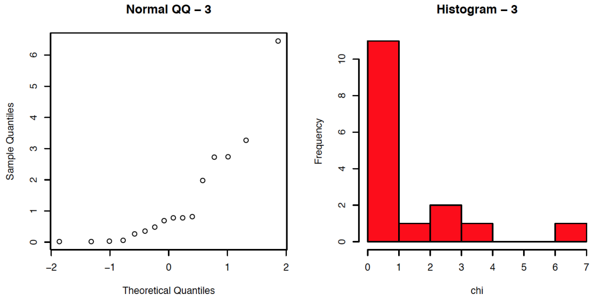


QQ plot example





QQ plot example



QQ plot reference

If you are interested in why QQ plot is a good test for normality. You can watch a great Youtube video from https://www.youtube.com/watch?v=X9$_{I}SJ0YpGw$.

Outline

CLT

In many common situations, it is reasonable to assume that our sample is from a normal distribution. But as you can see from the QQ plot, some samples are not from Normal distribution.

In many common situations, it is reasonable to assume that our sample is from a normal distribution. But as you can see from the QQ plot, some samples are not from Normal distribution.

Here comes a strong statement and it is very **important**!

CLT

Here comes a strong statement and it is very **important**!

Theorem

Let X_1, X_2, ..., X_n be a collection of iid RVs with $E(X_i) = \mu$ and $VAR(X_i) = \sigma^2$. For large enough n, the distribution of $\bar{X}$ will be approximately normal with $E(\bar{X}) = \mu$ and $VAR(\bar{X}) = \frac{\sigma^2}{n}$. That is, $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$.

How large is "large enough"?

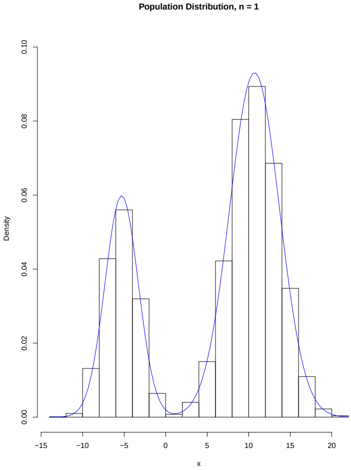
CLT

How large is "large enough"?

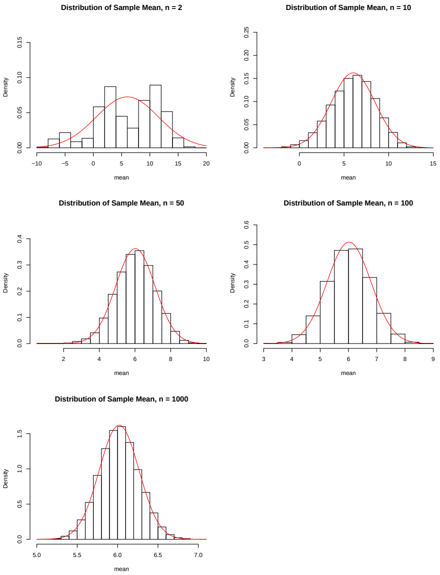
The required size for n depends on the nature of the population distribution of X_i. The closer the distribution of X_i is to normal, the smaller n is required for the approximation to be good. For reasonably symmetric distributions with no outliers, $n = 5$ could be sufficient. For distributions with extreme skew or heavy tails/outliers, you may need upwards of $n = 100$ or more. But for much real-world data, $n = 30$ is a relatively safe cut-off, and this sample size is what is typically prescribed to use the CLT.

CLT simulation in R

We consider the population distribution which is a mixture of two normal distributions.



CLT simulation in R



What's the next?

In the next lecture, we'll discuss confidence intervals.