# Chapter 5: Estimation

## (Ott & Longnecker Sections: 4.12, 4.14 and 5.2)

Duzhe Wang

https://dzwang91.github.io/stat371/

Part 3

UNIVERSITY OF WISCONSIN–MADISON

# What do we study?

**Key Concepts:** QQ plot, central limit theorem

# Outline

1. QQ plot
2. Central limit theorem

# QQ plot

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

# QQ plot

- We usually assume that the sample is from Normal distribution, how 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 is from Normal distribution, how 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.

# QQ plot

- We usually assume that the sample is from Normal distribution, how 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

- Recall the $p \times 100\%$ quantile point $q$ for random variable X is a point that satisfies

$$P(X \leq q) = F_X(q) = p$$

where $F_X$ is the cumulative distribution function of X.

# QQ plot

- Recall the $p \times 100\%$ quantile point $q$ for random variable X is a point that satisfies
$$P(X \leq q) = F_X(q) = p$$
where $F_X$ is the cumulative distribution function of X.
- The quantile $q$ is given by $q = F_X^{-1}(p)$.

# QQ plot

- Recall the $p \times 100\%$ quantile point $q$ for random variable X is a point that satisfies

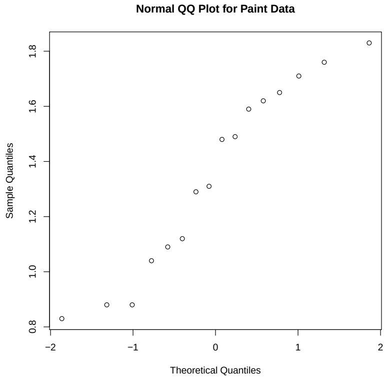
$$P(X \le q) = F_X(q) = p$$

where $F_X$ is the cumulative distribution function of X.

- The quantile $q$ is given by $q = F_X^{-1}(p)$.
- QQ-plot of two random variables $X$ and $Y$ is defined to be a parametric curve $C(p)$ parameterized by $p \in [0, 1]$:

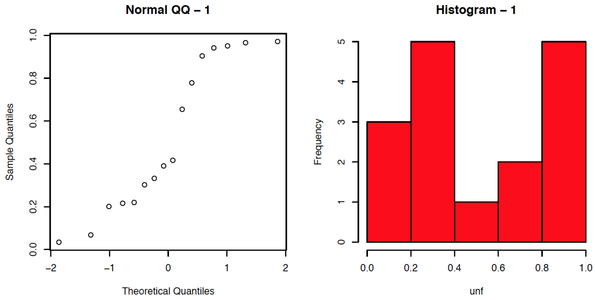
$$C(p) = (F_X^{-1}(p), F_Y^{-1}(p))$$

# QQ plot example

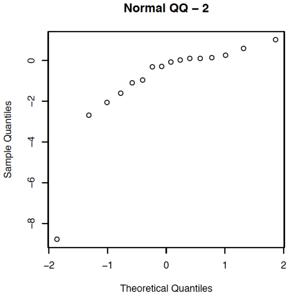


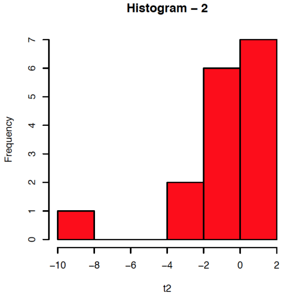

The plot is not perfectly straight, but it is pretty good.

# QQ plot example

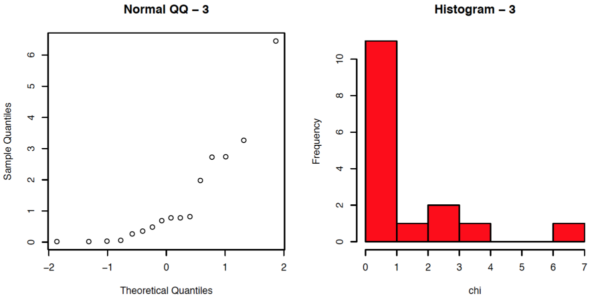


# QQ plot example



# QQ plot example



# Outline

1. QQ plot

2. Central limit theorem

# Central limit theorem

- 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. Does it matter?

# Central limit theorem

- 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. Does it matter?

# Central limit theorem

## Theorem

*(Informal) It does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the distribution of sample means tend to follow the normal distribution as sample size is larger and larger.*

# Central limit theorem

**Theorem**

*(Informal) It does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the distribution of sample means tend to follow the normal distribution as sample size is larger and larger.*

**Theorem**

*(Formal) 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, $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$.*

This theorem it is very important!

# Central limit theorem

- How large is large enough?

# Central limit theorem

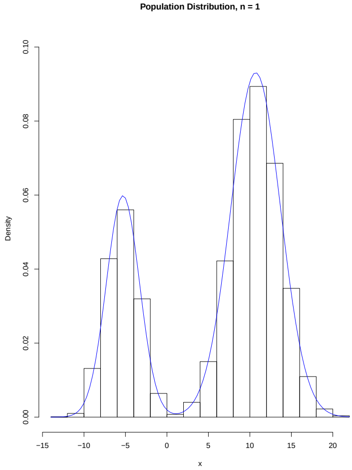
- How large is large enough?
- The required size for $n$ depends on the nature of the population distribution of $X_i$. The closer the distribution of $X_i$ is to normal, the smaller $n$ is required for the approximation to be good.

# Central limit theorem

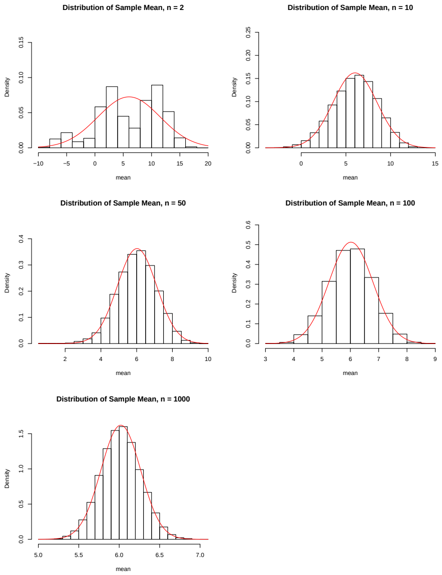
- How large is large enough?
- The required size for $n$ depends on the nature of the population distribution of $X_i$. The closer the distribution of $X_i$ is to normal, the smaller $n$ is required for the approximation to be good.
- For reasonably symmetric distributions with no outliers, $n = 5$ could be sufficient. For distributions with extreme skew or heavy tails/outliers, you may need $n = 100$ 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.