# Nonparametric Statistics

## Ch.1 Introduction

# Parametric vs Nonparametric

**1. Parametric models**

A parametric model is a model where the underlying distribution of the observations is characterized by a few (finite) number of parameters.

Ex1] Let $X_1, \ldots, X_{100}$ be the heights of 100 men. We may assume that they are independently and identically distributed as $N(\mu, \sigma^2)$.

Ex2] The number of car accidents in some city can be modelled as Poisson($\lambda$) under suitable assumptions.

**2. Nonparametric models**

A nonparametric model is the one where only few assumptions are made so that it cannot be characterized a finite number of parameters.

Ex] Let $X_1, \ldots, X_{100}$ be the heights of 100 men. We may assume that they are independent and identical observations from a ***symmetric*** distribution.

# Parametric vs Nonparametric (continued)

- Why nonparametric?

1. Robustness : It requires few assumptions, which means that we do not have to worry much about the misspecification of the model.
2. Insensitive to outliers
3. Useful when parametric assumptions are not adequate or one has no prior information about the underlying models.

# Parametric vs Nonparametric (continued)

Ex] (**one-sample t-test**) We collect 25 observations to see if the weight of a beverage product is controlled at 300g as it appears on the label. The sample mean and the sample standard deviations are:

$$\bar{X} = 295,\ \mathrm{S} = 10.$$

What is your conclusion?

- Answer by one-sample t-test procedure]

Hypothesis : $H_0 : \mu = 300 \quad vs \quad \mu \neq 300$

Observed test statistics : $\frac{295 - 300}{\frac{10}{25^{1/2}}} = -2.5$

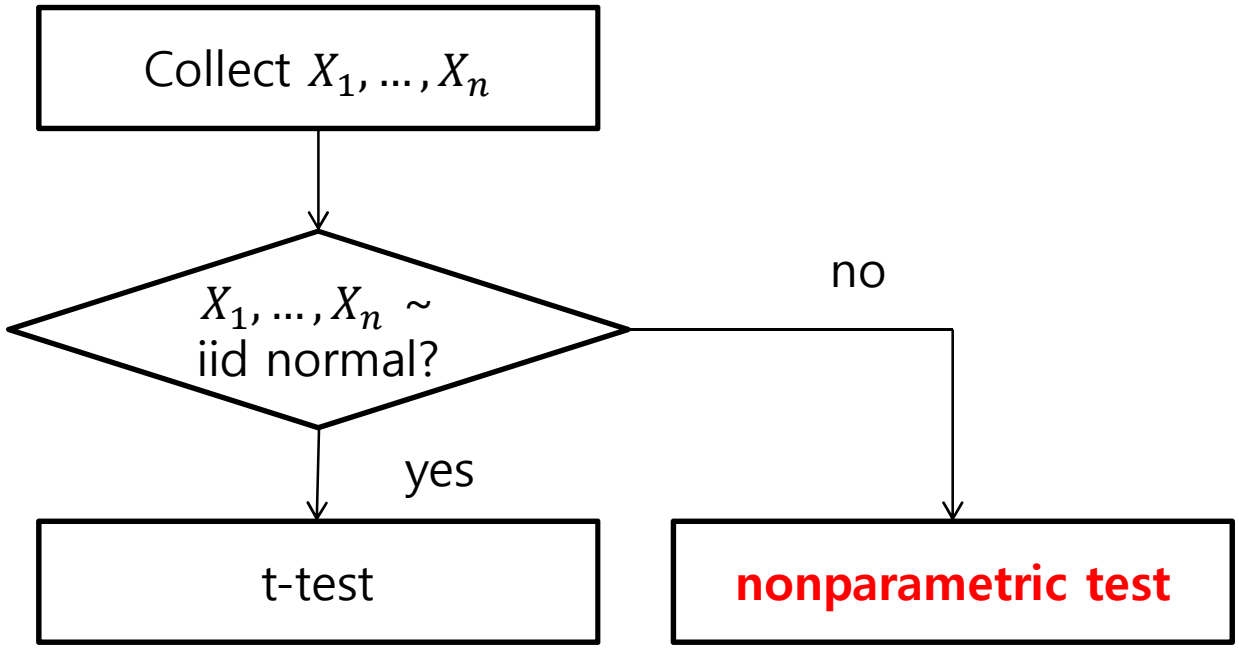
P-value : less than 0.05

We reject the null hypothesis at the significance level 0.05. Therefore, we conclude that the weight of the product is not properly controlled.

# Parametric vs Nonparametric (continued)

**However!** The test is done under the assumption that the observations come from $N(\mu, \sigma^2)$. Therefore, the test may lead us to a wrong conclusion if

- The observations follow some other distribution.
- There exist outliers that affect the whole testing procedure.



# Parametric vs Nonparametric (continued)

Ex] (**independence test**) We want to know if blood pressure is associated with BMI. 27 people are examined to see this. The Pearson correlation coefficient is 0.5. Are they independent or not?

- Answer under the normality assumption]

Hypothesis : $H_0 : \rho = 0 \quad vs \ \rho \neq 0$

Observed test statistics : $(27-2)^{1/2} \times \frac{0.5}{(1-0.5^2)^{1/2}} = 2.89$

P-value : less than 0.05

We reject the null hypothesis at the significance level 0.05. Therefore, we conclude that they are not independent.

**However!** The test is done under the normality assumption. Therefore, the test may lead us to a wrong conclusion if

- The observations follow some other distribution.
- There exist outliers that affect the whole testing procedure.

# Parametric vs Nonparametric (continued)

- Parametric models are ***efficient*** when employed assumptions are reasonable.
- Nonparametric models are ***safe*** because we include as few assumptions as possible.
- Many nonparametric approaches utilize the ranks of observations rather than their magnitudes. It means that there is loss of information which makes nonparametric approaches less efficient than parametric ones when the parametric assumptions are true. However, it is known that they are only slightly less efficient, and they are much more efficient when the parametric assumptions fail.
- Nonparametric approaches can be used to check if the parametric assumptions are reasonable or not.
- In most cases, there exist the nonparametric counterparts of parametric estimation or testing procedures.
- When the number of observations is large so that the central limit theorem is applicable, one may not need nonparametric approaches in some cases.
- Assumptions on the underlying distribution affect the whole statistical inference such as point estimation, interval estimation and testing.

# Sign and Rank

1. Sign

Every real number has a sign. When it is larger than 0, the sign is + (or +1). When it is smaller than 0, the sign is – (or -1). The number zero has no sign.

2. Rank

Real numbers are naturally ordered. The rank of a number in a set is the function from real numbers to positive integers. Consider the following set.

$$\{2\,,-1\,,5\,,0\,,7\}$$

Then, rank{2}=3, rank{-1}=1.

Note] Recall the sample 'mean' and the sample 'median' are the estimators of the location parameter of a distribution. Which one is more related to rank?