

# NONPARAMETRIC STATISTICS

## LECTURE 1

---

Jyotishka Datta

University of Arkansas, Fayetteville


August 25, 2017



# REVIEW OF UNIVARIATE DISTRIBUTIONS AND R

- Expectation and Variance
- Order Statistics, Quantiles
- Normal & Binomial distribution.
- Standard Continuous distributions.
- Sign test theory
- Introduction to R & generate random variables
- Implement Sign test using R

# RECAP

- $X$  is a random variable if for every real number  $x$ , there exists a probability that the value of the random variable doesn't exceed  $x$ , denoted by  $P(X \leq x)$  or  $F(x)$ .  
  
C.D.F.
- $F_X(x) \equiv F(x)$ : CDF or Cumulative distribution function satisfies:
  - 1  $F(x)$  is non-decreasing.
  - 2  $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow \infty} F(x) = 1$ .
  - 3  $F(x)$  is continuous from right.

# RECAP

- A random variable  $X$  is called **continuous** if its CDF  $F(x)$  is continuous (differentiable almost everywhere). The derivative of CDF is called PDF (probability density function), denoted by  $f(x) = \frac{\partial F(x)}{\partial x}$ .
- For a continuous  $X$ ,

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt \text{ where } \int_{-\infty}^{\infty} f(x)dx = 1$$

- A random variable  $X$  is called **discrete** if it can take only a **finite or countably infinite** number of values. The probability mass function of  $X$  is defined as  $f(x) = P(X = x)$ .
- For a discrete RV,

$$f(x) \geq 0 \forall x, \text{ and } \sum_x f(x) = 1 \text{ and } F(s) = \sum_{x \leq s} P(X = x)$$

## RECAP (CONTINUED) I

- **Expectation:** A random variable takes different values with different probabilities. We might want to know what value it takes on average. Just taking the arithmetic average is a very naive concept since a simple average of just the possible values of the random variable will be misleading, because some values may have so little probability that they are relatively inconsequential. Therefore, we should weigh the observations by their probabilities.
- The average or the mean value, also called the expected value of a random variable is a weighted average of the different values of  $X$ , weighted according to how important the value is.
- Let  $X$  be a discrete / continuous random variable. We say that the expected value of  $X$  is defined as

$$\mu = \mathbb{E}(X) = \begin{cases} \sum_{i=1}^n x_i f(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{continuous} \end{cases}$$

## RECAP (CONTINUED) II

- The expected value of a function  $g(X)$  of a random variable  $X$ , denoted by  $\mathbb{E}[g(X)]$ , given by:

$$\mathbb{E}(g(X)) = \begin{cases} \sum_i g(x_i)f(x_i), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)f(x)dx, & \text{if } X \text{ is continuous.} \end{cases}$$

- Mean of a random variable  $X$ ,  $\mathbb{E}(X)$  measures the center/location of its distribution.
- Variance  $\mathbb{V}(X)$  measures the dispersion/spread of its distribution.

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}[\{X - \mathbb{E}(X)\}^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\ &= \begin{cases} \int_{-\infty}^{\infty} x^2 f(x)dx - \left(\int_{-\infty}^{\infty} xf(x)dx\right)^2, & \text{if } X \text{ is continuous} \\ \sum_x x^2 f(x) - (\sum_x xf(x))^2, & \text{if } X \text{ is discrete} \end{cases} \end{aligned}$$

# PROPERTIES

- Expectation is a linear operator  $\mathbb{E}(a + bX) = a + b \mathbb{E}(X)$ , but variance is not:  
 $\mathbb{V}(a + b X) = b^2 \mathbb{V}(X)$ .

# PROPERTIES

- 1 Expectation is a linear operator  $\mathbb{E}(a + bX) = a + b \mathbb{E}(X)$ , but variance is not:  $\mathbb{V}(a + bX) = b^2 \mathbb{V}(X)$ .
- 2 More generally, if  $X_1, \dots, X_n$  are independent,

$$\mathbb{E}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i \mathbb{E}(X_i), \quad \mathbb{V}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 \mathbb{V}(X_i) \quad (1)$$



# PROPERTIES

- 1 Expectation is a linear operator  $\mathbb{E}(a + bX) = a + b \mathbb{E}(X)$ , but variance is not:  $\mathbb{V}(a + bX) = b^2 \mathbb{V}(X)$ .
- 2 More generally, if  $X_1, \dots, X_n$  are independent,

$$\mathbb{E}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i \mathbb{E}(X_i), \quad \mathbb{V}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 \mathbb{V}(X_i) \quad (1)$$

- 3 Suppose  $X_i$ 's are independent and identically distributed with mean  $\mu$  and variance  $\sigma^2$ ,  $\mu$  and  $\sigma$  unknown parameters. We can estimate them by the sample mean and sample variance:
- 4 Sample mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ .

# PROPERTIES

- 1 Expectation is a linear operator  $\mathbb{E}(a + bX) = a + b \mathbb{E}(X)$ , but variance is not:  $\mathbb{V}(a + bX) = b^2 \mathbb{V}(X)$ .
- 2 More generally, if  $X_1, \dots, X_n$  are independent,

$$\mathbb{E}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i \mathbb{E}(X_i), \quad \mathbb{V}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 \mathbb{V}(X_i) \quad (1)$$

- 3 Suppose  $X_i$ 's are independent and identically distributed with mean  $\mu$  and variance  $\sigma^2$ ,  $\mu$  and  $\sigma$  unknown parameters. We can estimate them by the sample mean and sample variance:
- 4 Sample mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ .
- 5 Sample variance  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ .

# SAMPLE MEAN AND VARIANCE

- Consequence of (1):

$$\mathbb{E}(\bar{X}) = \mu, \mathbb{V}(\bar{X}) = \frac{\sigma^2}{n}, \mathbb{E}(S^2) = \sigma^2, \mathbb{V}(S^2) = \frac{2\sigma^4}{n-1}$$

**PROBLEM** The number of pieces of mail a household receives daily is said to have a mean of 4 with standard deviation 1.75. The number of items is discrete and not normally distributed. What would be the distribution of an average number of mail items per household per vehicle if samples of 50 houses were selected?

**ANSWER:**  $\mathbb{V}(\bar{X}) = \frac{\sigma^2}{n} \Rightarrow \text{s.d.}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = 1.75/\sqrt{50} = 0.25$ . Hence,  $\bar{X} \sim \mathcal{N}(4, 0.25)$ .

- (Alt-) Measures of center: Median, Mode. and Measures of dispersion: Range, MAD (Mean Absolute Deviation), IQR (Inter-quartile range).

# MEDIAN

- The median  $M$  is the mid-point of a distribution. Half the observations are smaller than the median and the other half are larger than the median, i.e.  $P(X \geq M) = P(X < M) = \frac{1}{2}$ .
  - Rule for finding the median:
    - ① Arrange all observations from smallest to largest. If the number of observations  $n$  is odd, the median  $M$  is the center observation in the ordered list. If the number of observation  $n$  is even, the median  $M$  is the mean of the two center observations in the ordered list.
- Data-set 1: 5, 7, 9, 13, 15, Data-set 2: 5, 7, 9, 10, 13, 15

# MEDIAN

- The median  $M$  is the mid-point of a distribution. Half the observations are smaller than the median and the other half are larger than the median, i.e.  $P(X \geq M) = P(X < M) = \frac{1}{2}$ .
- Rule for finding the median:
  - ① Arrange all observations from smallest to largest. If the number of observations  $n$  is odd, the median  $M$  is the center observation in the ordered list. If the number of observation  $n$  is even, the median  $M$  is the mean of the two center observations in the ordered list.  
Data-set 1: 5, 7, 9, 13, 15, Data-set 2: 5, 7, 9, 10, 13, 15
- Notation:  $X_1, \dots, X_n \stackrel{IID}{\sim} F_X$ . Suppose  $X_{(1)}$  denote the smallest observation,  $X_{(2)}$  the second smallest and so on, then  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$  denote the sorted sample. These are called the **Order Statistics**.
- The median is

$$M = \begin{cases} X_{(\frac{n+1}{2})}, & \text{if } n \text{ is odd} \\ \text{any number in } (X_{(n/2)}, X_{(n/2)+1}) & \text{if } n \text{ is even} \end{cases}$$

# QUANTILES

- Suppose a random variable  $Z$  follows a continuous distribution. The  $\alpha$ -th quantile of  $Z$  is defined as

$$\begin{aligned} F^{-1}(\alpha) &= Q_{\alpha} = \min\{z : F(z) \geq \alpha\} \\ &= \min\{z : P(Z \leq z) \geq \alpha\}, \text{ i.e. } P(Z < Q_{\alpha}) = \alpha. \end{aligned}$$

- For example, if  $Z \sim N(0; 1)$ ,  $Q_{0.95} = 1.645$ .
- The definitions are similar for discrete distributions.
- Some popular quantiles of a distribution are known as the quartiles.
- The first quartile is the 0.25th quantile, the second quartile is the 0.5th quantile (median), and the third quartile is the 0.75th quantile.

# NORMAL DISTRIBUTION

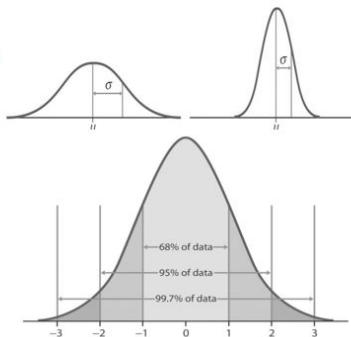
- Normal density curves: Symmetric, Bell-shaped and unimodal. Notation  $\mathcal{N}(\mu, \sigma)$ .
- Two parameters  $\mu$  and  $\sigma$  - knowing them allows us to make various conclusions about Normal distributions.
- Probability density function:  $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2})$ ,  $x \in \mathbb{R}$ .
- $X \sim \mathcal{N}(\mu, \sigma) \Rightarrow \mathbb{E}(X) = \mu, \mathbb{V}(X) = \sigma^2$ . Standard Normal:  $X \sim \mathcal{N}(0, 1)$ .

# NORMAL DISTRIBUTION

- Normal density curves: Symmetric, Bell-shaped and unimodal. Notation  $\mathcal{N}(\mu, \sigma)$ .
- Two parameters  $\mu$  and  $\sigma$  - knowing them allows us to make various conclusions about Normal distributions.
- Probability density function:  $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2})$ ,  $x \in \mathbb{R}$ .
- $X \sim \mathcal{N}(\mu, \sigma) \Rightarrow \mathbb{E}(X) = \mu, \mathbb{V}(X) = \sigma^2$ . Standard Normal:  $X \sim \mathcal{N}(0, 1)$ .

## 68-95-99.7 Rule for any Normal Curve:

- 68% of the observations fall within one standard deviation of the mean.
- 95% of the observations fall within two standard deviations of the mean.
- 99.7% of the observations fall within three standard deviations of the mean.





# NORMAL DISTRIBUTION

- Normal density curves: Symmetric, Bell-shaped and unimodal. Notation  $\mathcal{N}(\mu, \sigma)$ .
- Two parameters  $\mu$  and  $\sigma$  - knowing them allows us to make various conclusions about Normal distributions.
- Probability density function:  $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2})$ ,  $x \in \mathbb{R}$ .
- $X \sim \mathcal{N}(\mu, \sigma) \Rightarrow \mathbb{E}(X) = \mu, \mathbb{V}(X) = \sigma^2$ . Standard Normal:  $X \sim \mathcal{N}(0, 1)$ .
- Standardized score: If  $X \sim \mathcal{N}(\mu, \sigma)$ ,  $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$ .
- The CDF of  $\mathcal{N}(0, 1)$ ,  $F_Z(z) = P_{Z \sim \mathcal{N}(0,1)}(Z \leq z) \doteq \Phi(z)$  - does **NOT** admit a closed form, but available from Normal tables or any stat software.
- CDF of any normal random variable can be written in terms of std. normal CDF:

$$X \sim \mathcal{N}(\mu, \sigma), F_X(x) = P(X \leq x) = P(\underbrace{\frac{X - \mu}{\sigma}}_{Z \sim \mathcal{N}(0,1)} \leq \frac{x - \mu}{\sigma}) = \Phi(\frac{x - \mu}{\sigma})$$

# BINOMIAL DISTRIBUTION

- Represents a sequence of independent coin tossing experiment.
- Suppose a coin with probability  $p$ ;  $0 < p < 1$  for heads in a single trial is tossed independently a pre-specified  $n$  times,  $n \geq 1$ . Let  $X$  be the number of times in the  $n$  tosses that a head is obtained. Then the pmf of  $X$  is:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, x = 0, 1, \dots, n$$

- Coin tossing, of course, is just an artifact.
- Suppose a trial can result in only one of two outcomes, called a success(S) or a failure(F), the probability of obtaining a success being  $p$  in any trial. Such a trial is called a Bernoulli trial.  $X_i \sim \text{Bernoulli}(p)$

## BINOMIAL DISTRIBUTION

Binomial is sum of  $n$  independent Bernoulli trials.

If  $X \sim \text{Bin}(n, p)$ . Mean  $\mathbb{E}(X) = np$ , Variance:  $\mathbb{V}(X) = np(1 - p)$ .

# EMPIRICAL CDF (ECDF) I

- The CDF  $F$  completely determines the distribution. How do we estimate it?
- Empirical CDF: Let  $x_1, \dots, x_n$  be independent and identically distributed samples from distribution  $F(\cdot)$ . Then the empirical CDF is defined as:

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(x_i \leq t) = \text{Proportion of observations } \leq t.$$

- Is  $\widehat{F}_n(t)$  a good estimator of the true CDF  $F(t)$ ?

## EMPIRICAL CDF (ECDF) II

- If  $\delta_i(t) = I\{X_i \leq t\}$ , then  $P(\delta_i(t) = 1) = P(X \leq t) = F(t)$ .
- Hence,  $\delta_1(t), \dots, \delta_n(t) \sim \text{Bernoulli}(F(t))$ .
- If  $X_i \sim \text{Bernoulli}(p)$  (coin-toss),  $X = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$ .
- This means,  $T_n(t) = \sum_{i=1}^n \delta_i(t)$  = Total number of observations below  $t$  is a sum of Bernoulli's.

### THEOREM

*The total number of sample points below  $t$ ,  $T_n(t) \sim \text{Bin}(n, F(t))$ , for a given  $t$ . By properties of Binomial,  $\mathbb{E}(T_n(t)) = nF(t)$  and  $\mathbb{V}(T_n(t)) = nF(t)(1 - F(t))$ .*

# EMPIRICAL CDF (ECDF) III

- Now look at the e-CDF:  $\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(x_i \leq t) = \frac{1}{n} T_n(t)$ .
- Recall  $\mathbb{E}(bX) = b\mathbb{E}(X)$  but  $\mathbb{V}(bX) = b^2\mathbb{V}(X)$ .

## COROLLARY

The e-CDF is a random variable  $F_n(t) = \frac{1}{n} T_n(t)$ , which has mean and variance:

$$E(F_n(t)) = F(t) \text{ and} \tag{2}$$

$$\text{Var}(F_n(t)) = \frac{F(t)(1 - F(t))}{n} \rightarrow 0 \text{ as } n \rightarrow \infty. \tag{3}$$

## IMPLICATION

The first half tells us that the e-CDF is an unbiased estimator of the true CDF, i.e., the estimator is centered around the true CDF  $F(t)$ , and the second half tells us that the estimation accuracy gets better and better, as we keep increasing the sample size – and ultimately, converges at the true value of  $F(t)$  when  $n \rightarrow \infty$ .

[\*\*https://jdatta.shinyapps.io/eCDFdemo/\*\*](https://jdatta.shinyapps.io/eCDFdemo/)

# SIGN TEST

- Data:  $x_1, \dots, x_n \stackrel{\text{iid}}{\sim} F$  with median  $M$ .
- Null hypothesis:  $M = \mu_0$  and Alternative:  $M \neq \mu_0$  or  $M > \mu_0$  or  $M < \mu_0$ .
- Test statistic:  $S$  = the number of observations that exceed  $\mu_0$ .
- Strategy: Reject the null if  $S$  is too big or too small (depending on the alternative).
- If  $H_0: M = \mu_0$  is true we would expect 50% of the observations to be above  $\mu_0$ , and 50% of the observations to be below  $\mu_0$ .
- Observe:  $S = \sum_{i=1}^n \mathbf{1}(X_i > \mu_0)$ , and  $P(X_i > \mu_0 \mid H_0) = \frac{1}{2}$  for each  $i$ .
- Question: What is the distribution of each  $X_i$ , and  $S$ ?

# SIGN TEST

- Binomial is sum of  $n$  independent Bernoulli trials.
- $X_i \sim \text{Bernoulli}(p)$  for  $i = 1, \dots, n$ , then  $X = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$ .

## SIGN TEST

The sign-test statistic  $S$  has a binomial distribution:  $S \sim \text{Bin}(n, p)$ , where  $p$  = the probability that an observation is greater than  $\mu_0$ .

In particular, if  $H_0$  is true then  $S$  will have a binomial distribution:  $S \sim \text{Bin}(n, \frac{1}{2})$ .



# SIGN TEST

- Binomial is sum of  $n$  independent Bernoulli trials.
- $X_i \sim \text{Bernoulli}(p)$  for  $i = 1, \dots, n$ , then  $X = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$ .

## SIGN TEST

The sign-test statistic  $S$  has a binomial distribution:  $S \sim \text{Bin}(n, p)$ , where  $p$  = the probability that an observation is greater than  $\mu_0$ .

In particular, if  $H_0$  is true then  $S$  will have a binomial distribution:  $S \sim \text{Bin}(n, \frac{1}{2})$ .

- If  $X \sim \text{Bin}(n, p)$ . Mean  $\mathbb{E}(X) = np$ , Variance:  $\mathbb{V}(X) = np(1 - p)$ .
- What are the mean and variance of  $S$  under  $H_0$ ?

# SIGN TEST

- Binomial is sum of  $n$  independent Bernoulli trials.
- $X_i \sim \text{Bernoulli}(p)$  for  $i = 1, \dots, n$ , then  $X = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$ .

## SIGN TEST

The sign-test statistic  $S$  has a binomial distribution:  $S \sim \text{Bin}(n, p)$ , where  $p$  = the probability that an observation is greater than  $\mu_0$ .

In particular, if  $H_0$  is true then  $S$  will have a binomial distribution:  $S \sim \text{Bin}(n, \frac{1}{2})$ .

- If  $X \sim \text{Bin}(n, p)$ . Mean  $\mathbb{E}(X) = np$ , Variance:  $\mathbb{V}(X) = np(1 - p)$ .
- What are the mean and variance of  $S$  under  $H_0$ ?
- Under  $H_0 : p = \frac{1}{2}$ ,  $E(S) = \frac{n}{2}$  and  $\mathbb{V}(S) = \frac{n}{4}$ .
- Intuitively, if  $S$  is too far away from  $n/2$ ,  $H_0$  must be rejected.
- Next: How to do this formally?