

University of Ferrara

E DIPARTIMENTO
DI ECONOMIA
E MANAGEMENT

Stefano Bonnini & Valentina Mini

Probability and Inference concepts for Regression Analysis

Lecture 3

When we apply linear regression method:

- We are using sampling data (not population data)
- We base our conclusion on inference (recalling probabilistic assumption)
- We assure that continuous sampling data are:
 - * normally distributed
 - * linked by a linear relation

Brief Notes on Probability and Inference



Games of chance

Game where a randomizing device (dice, playing cards, roulette wheels, lottery, ...) influences the outcome



Each of the possible outcomes has a given probability of occurrence



Probability distribution:

each event E is given a probability $P(E)$

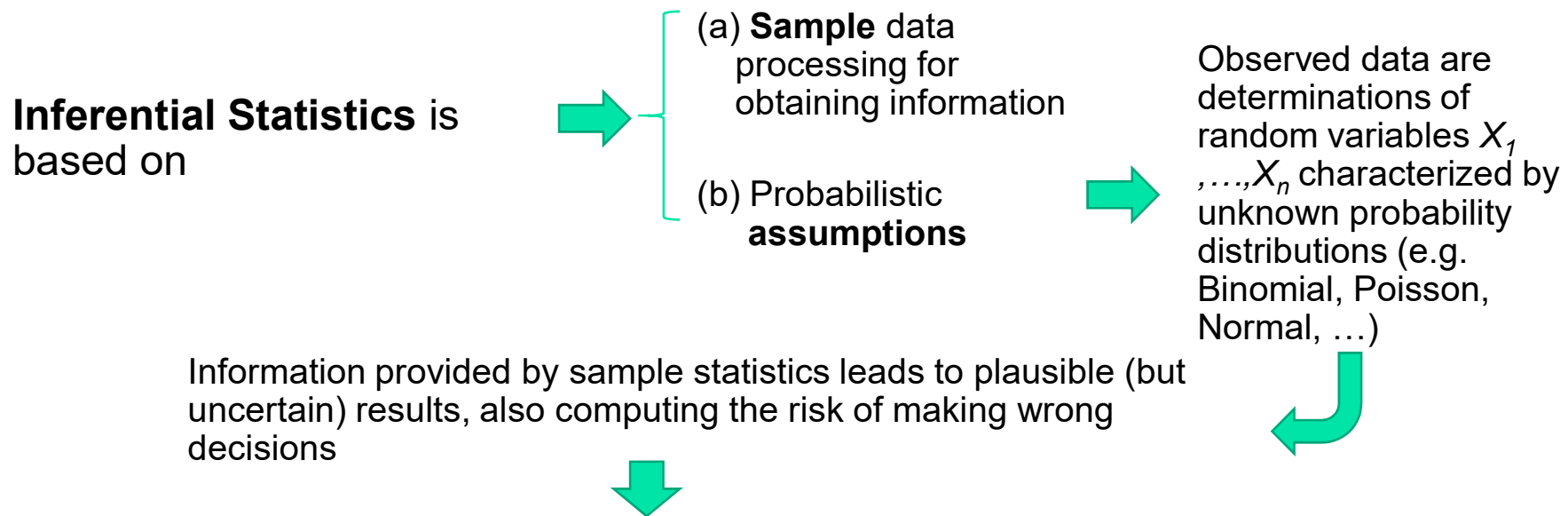


Probability function for discrete numerical variables $\rightarrow P(x)$: probability of number x
 $\sum_{x \in A} P(x)$: probability of the set A

Probability density function for continuous numerical variables $\rightarrow f(x)$: density of x
 $\int_{x \in A} f(x) dx$: probability of A

Brief Notes on Probability and Inference

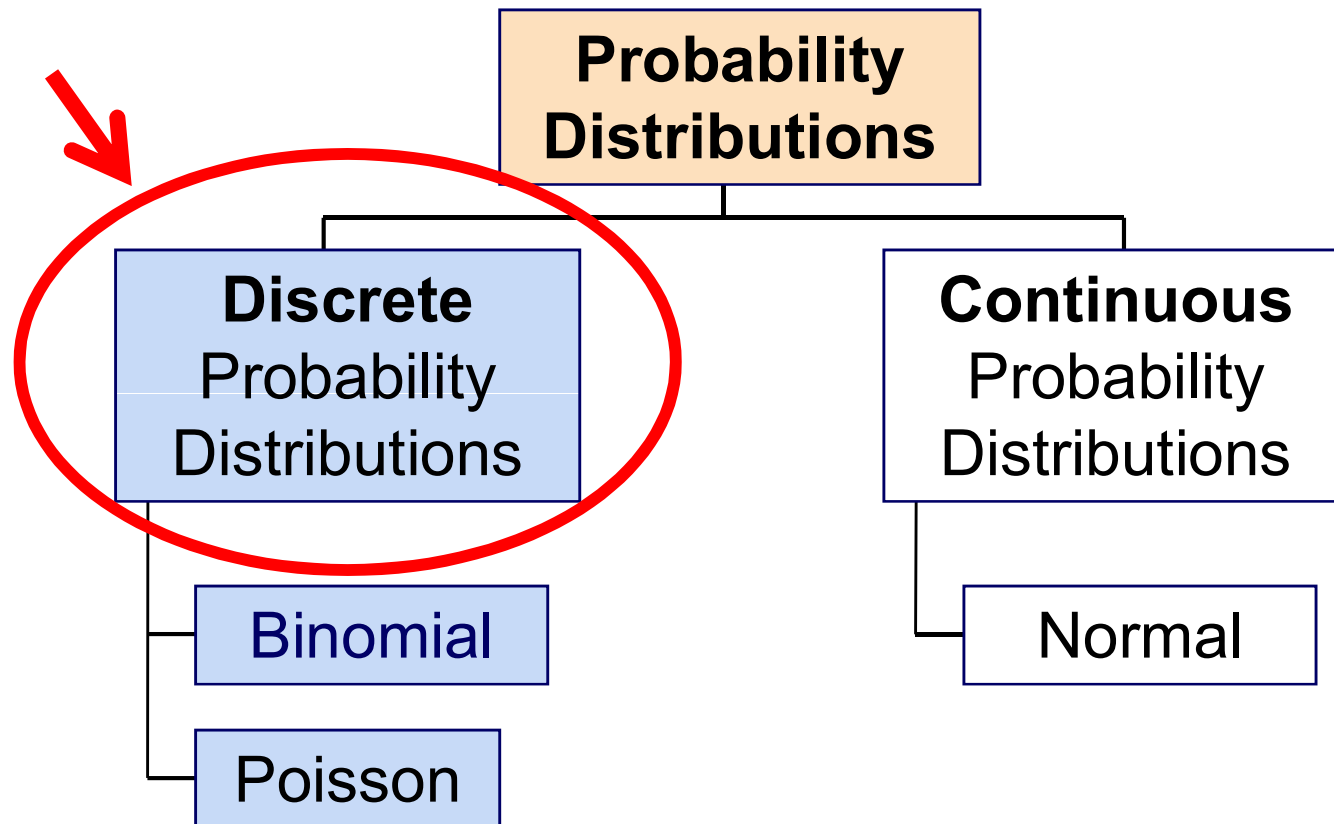
Inferential Statistics and Probability Theory



Parametric methods assume that distribution functions are known except for some unknown parameters

Nonparametric methods are based on less stringent assumptions and give more importance on (a) sample data processing

Probability Distributions



Brief Notes on Probability and Inference

- A **probability distribution for a discrete random variable** is a mutually exclusive listing of all possible numerical outcomes for that variable and a probability of occurrence associated with each outcome.

Number of Classes Taken	Probability
2	0.2
3	0.4
4	0.24
5	0.16

$$\mu = E(X) = \sum_{i=1}^N X_i P(X_i) = 2 \cdot 0.2 + 3 \cdot 0.4 + 4 \cdot 0.24 + 5 \cdot 0.16 = 3.36$$

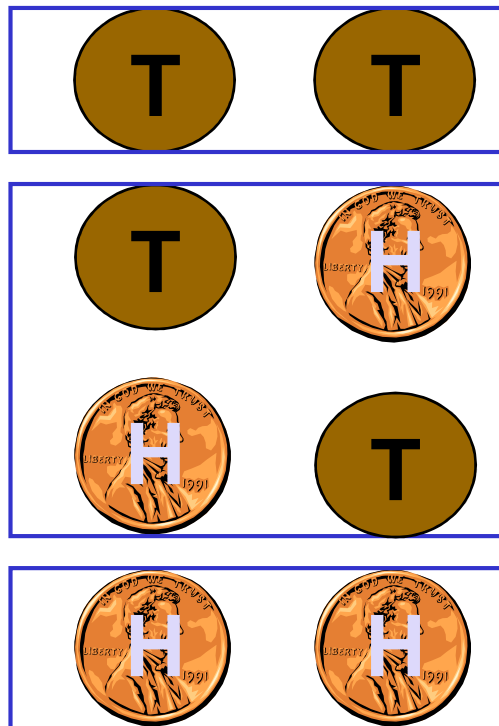
$$\sigma^2 = \sum_{i=1}^N [X_i - E(X)]^2 P(X_i) = (2 - 3.36)^2 \cdot 0.2 + (3 - 3.36)^2 \cdot 0.4 + (4 - 3.36)^2 \cdot 0.24 + (5 - 3.36)^2 \cdot 0.16 = 0.9504$$

$$\sigma = \sqrt{\sigma^2} = 0.9749$$

Brief Notes on Probability and Inference

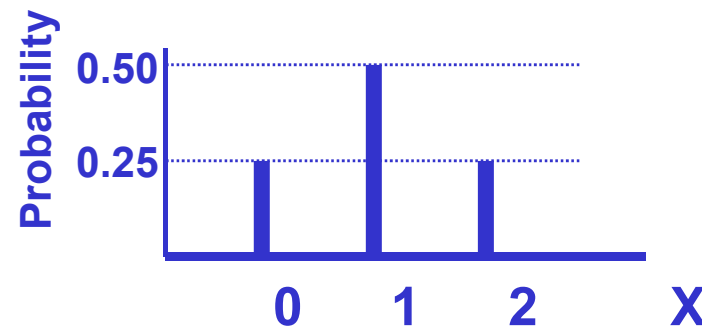
Experiment: Toss 2 Coins. We may obtain head or tails. Let $X = \#$ heads.

4 possible outcomes



Probability Distribution

<u>X Value</u>	<u>Probability</u>
0	$1/4 = 0.25$
1	$2/4 = 0.50$
2	$1/4 = 0.25$



Brief Notes on Probability and Inference

- **Expected Value (or mean)** of a discrete random variable (Weighted Average)

$$\mu = E(X) = \sum_{i=1}^N X_i P(X_i)$$

- **Example:** Toss 2 coins,
 $X = \#$ of heads,
compute expected value of X :

$$E(X) = ((0)(0.25) + (1)(0.50) + (2)(0.25)) \\ = 1.0$$

X	P(X)
0	0.25
1	0.50
2	0.25

Brief Notes on Probability and Inference

- **Variance** of a discrete random variable

$$\sigma^2 = \sum_{i=1}^N [X_i - E(X)]^2 P(X_i)$$

- **Standard Deviation** of a discrete random variable

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^N [X_i - E(X)]^2 P(X_i)}$$

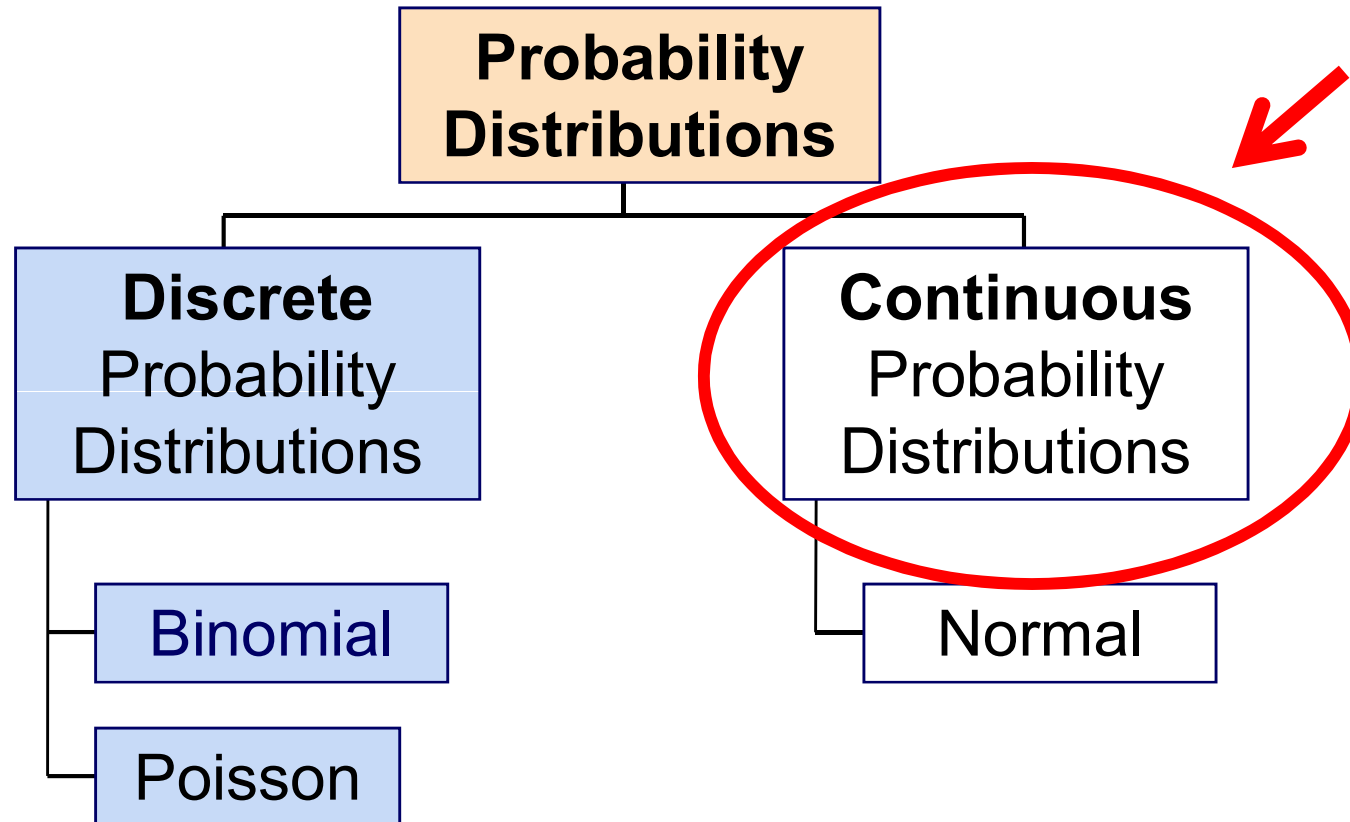
- **Example:** Toss 2 coins, $X = \#$ heads, compute standard deviation (recall $E(X) = 1$)

$$\sigma = \sqrt{\sum [X_i - E(X)]^2 P(X_i)}$$

$$\sigma = \sqrt{(0-1)^2(0.25) + (1-1)^2(0.50) + (2-1)^2(0.25)} = \sqrt{0.50} = 0.707$$

Possible number of heads
= 0, 1, or 2

Probability Distributions



Brief Notes on Probability and Inference

- A **continuous random variable** is a variable that can assume any value on a continuum (can assume an uncountable number of values)
 - thickness of an item
 - time required to complete a task
 - temperature of a solution
 - height, in centimeters

$$\mu = E(X) = \int_{-\infty}^{+\infty} x \cdot f(x) dx$$

$$\sigma^2 = E(X - \mu)^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x) dx$$

$f(x)$: probability density function

Brief Notes on Probability and Inference

A continuous random distribution:

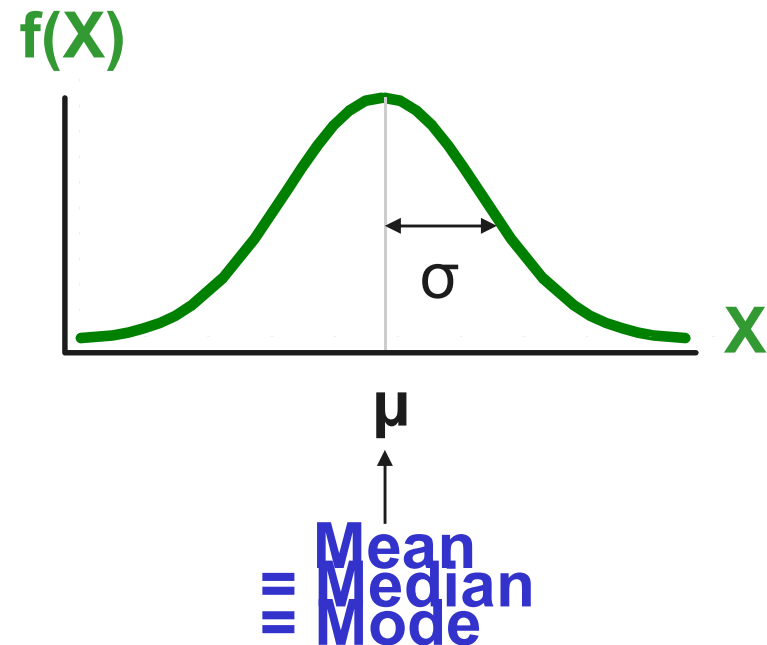
- **Bell Shaped**
- **Symmetrical**
- **Mean, Median and Mode are Equal**

Location (central tendency) is determined by the mean, μ

Spread is determined by the standard deviation, σ

The random variable has an infinite theoretical range:

$+\infty$ to $-\infty$



Brief Notes on Probability and Inference

- The formula for the **normal probability density function** is

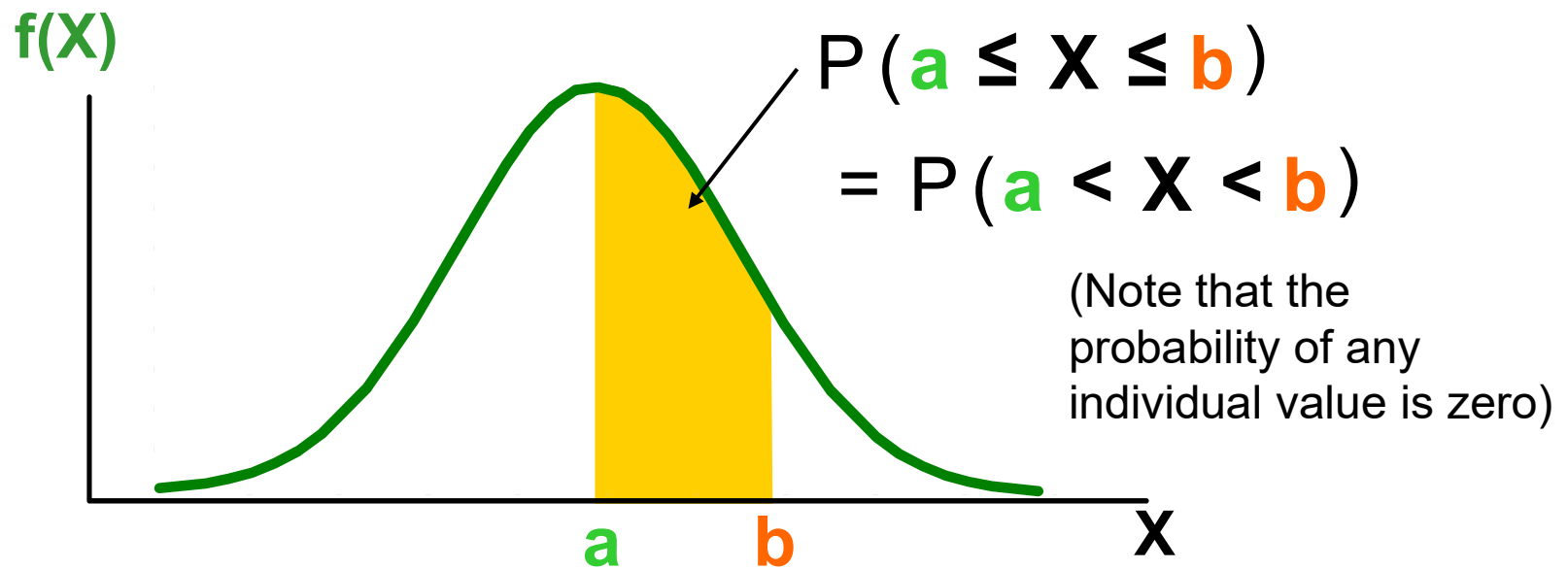
$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{(X-\mu)}{\sigma}\right)^2}$$

Where

- e = the mathematical constant approximated by 2.71828
- π = the mathematical constant approximated by 3.14159
- μ = the population mean
- σ = the population standard deviation
- X = any value of the continuous variable

Brief Notes on Probability and Inference

Probability is measured by the area under the normal curve



Brief Notes on Probability and Inference

- **Not all continuous distributions are normal**
- It is important to evaluate the plausibility of the assumption of normality.
- Normally distributed data should approximate the theoretical normal distribution:
 - The normal distribution is bell shaped (symmetrical) where the mean is equal to the median.
 - The empirical rule applies to the normal distribution.
 - The interquartile range of a normal distribution is 1.33 standard deviations.

We need to compare data characteristics to theoretical properties:

1. Construct **charts or graphs**

- For small- or moderate-sized data sets, construct a boxplot to check for symmetry
- For large data sets, does the histogram or polygon appear bell-shaped?

2. Compute **descriptive summary measures**

- Do the mean, median and mode have similar values?
- Is the interquartile range approximately 1.33σ ?
- Is the range approximately 6σ ?

(continued)

We need to compare data characteristics to theoretical properties:

3. Observe the distribution of the data set

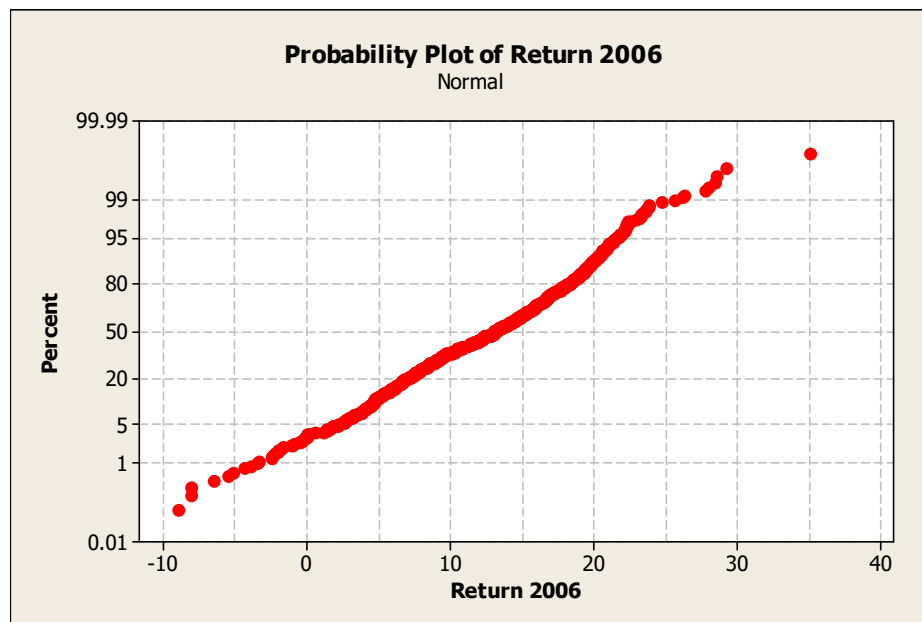
- Do approximately 2/3 of the observations lie within mean ± 1 standard deviation?
- Do approximately 80% of the observations lie within mean ± 1.28 standard deviations?
- Do approximately 95% of the observations lie within mean ± 2 standard deviations?

4. Evaluate normal probability plot

- Is the normal probability plot approximately linear (i.e. a straight line) with positive slope?

Brief Notes on Probability and Inference

(continued)



Scatter plot is approximately a straight line except for a few outliers at the low end and the high end.

Brief Notes on Probability and Inference

Finally, to understand if our model will be able to describe the reality (and not only the sample) we need to apply the

TEST OF HYPOTHESIS

A **test of hypothesis** is an inferential procedure based on sample data to test some assertions related to one or more populations

NULL HYPOTHESIS H_0

The **null hypothesis** usually corresponds to the status quo or the hypothesis of no effect, no difference, etc.

ALTERNATIVE HYPOTHESIS H_1

The **alternative hypothesis** represents the assertion that needs to be proved by the empirical evidence through sample data.