Introduction to

Statistical Data Analysis II

JULY 2011

Afsaneh Yazdani

Preface

Major branches of Statistics:

- Descriptive Statistics

- Inferential Statistics

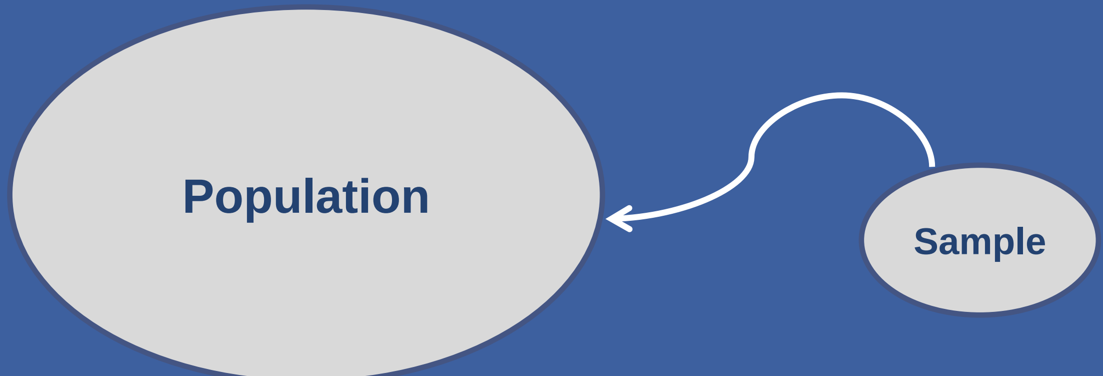
Preface

What is Inferential Statistics?

Preface

What is Inferential Statistics?

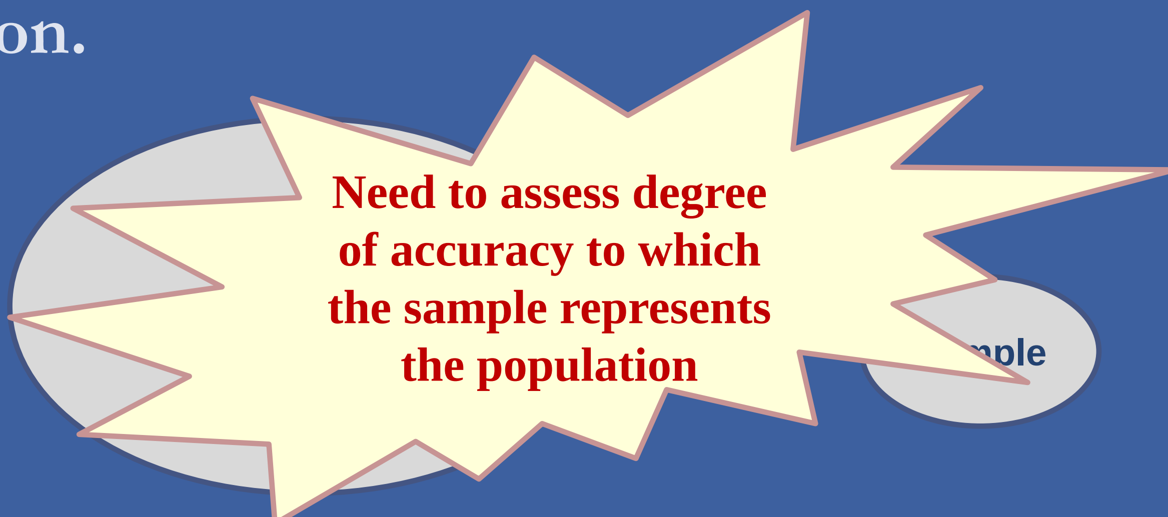
Making statements about population based on information contained in the sample of that population.



Preface

What is Inferential Statistics?

Making statements about population based on information contained in the sample of that population.



Preface

What is Inferential Statistics?

Making statements about population based on information contained in the sample of that population.

Preface

Probability is the:

- Language of uncertainty
- Tool for making inferences

Probability

Probability Definitions:

Classical Interpretation:

Each possible distinct result is called an **outcome**;
An **event** is identified as a collection of outcomes.
Then probability of an event 'E' is:

$$Pr(event\ E) = \frac{Number\ of\ outcomes\ favorable\ to\ event\ E\ (N_e)}{Total\ number\ of\ possible\ outcomes\ (N)}$$

Probability

Probability Definitions:

Relative frequency Interpretation:

Is an empirical approach to probability; if an experiment is conducted ‘n’ different times and if event ‘E’ occurs on n_e of these trials, then the probability of event ‘E’ is approximately:

$$Pr(event\ E) \cong \frac{n_e}{n}$$

Probability

Probability Definitions:

Relative frequency Interpretation:

Is an empirical approach to probability; if an experiment is conducted 'n' different times and if event 'E' occurs on n_e of these trials, then the probability of event 'E' is approximately:

$$Pr(event\ E) \cong \frac{n_e}{n}$$

very large number of observations or repetitions

Probability

Probability Definitions:

Subjective Interpretation:

Subjective or personal probability, the problem is that they can vary from person to person and they cannot be checked.

Probability

Basic Event Relations and Probability Laws:

The probability of an event, say event ‘A’, will always satisfy the property:

$$0 \leq P(A) \leq 1$$

Probability

Basic Event Relations and Probability Laws:

The probability of an event, say event ‘A’, will always satisfy the property:

$$0 \leq P(A) \leq 1$$

Impossible Event

Sure Event

Probability

Basic Event Relations and Probability Laws:

Two events ‘A’ and ‘B’ are said to be **mutually exclusive** if the occurrence of one of the events excludes the possibility of the occurrence of the other event:

$$P(\text{either A or B}) = P(A) + P(B)$$

Probability

Basic Event Relations and Probability Laws:

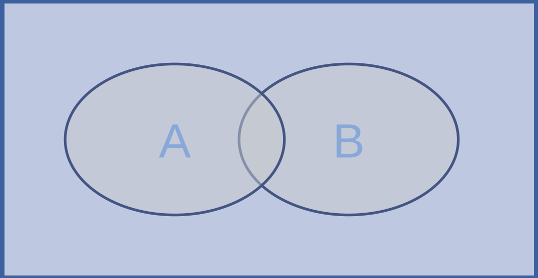
The complement of an event ‘A’ is the event that ‘A’ does not occur. The complement of ‘A’ is denoted by the symbol $\bar{A}$:

$$P(\bar{A}) + P(A) = 1$$

Probability

Basic Event Relations and Probability Laws:

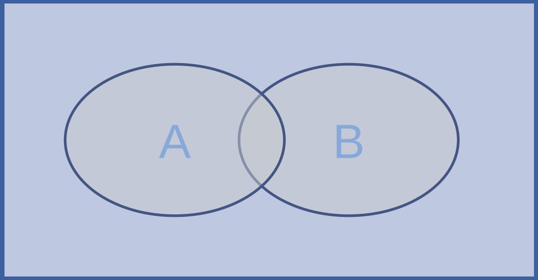
The **union** of two events 'A' and 'B' is the set of all outcomes that are included in either 'A' or 'B' (or both). The union is denoted as $A \cup B$.



Probability

Basic Event Relations and Probability Laws:

The intersection of two events ‘A’ and ‘B’ is the set of all outcomes that are included in both ‘A’ and ‘B’. The intersection is denoted as $A \cap B$.



Probability

Basic Event Relations and Probability Laws:

Consider two events ‘A’ and ‘B’; the **probability of the union** of ‘A’ and ‘B’ is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Conditional Probability and Independence:

Consider two events 'A' and 'B' with nonzero probabilities, P(A) and P(B). The **conditional probability** of event 'A' given event 'B'

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Probability

Conditional Probability and Independence:

Multiplication Law implies that the probability of the intersection of two events 'A' and 'B' is:

$$P(A \cap B) = \mathrm{P}(\mathrm{A}|B)P(B)$$

Probability

Conditional Probability and Independence:

Two events ‘A’ and ‘B’ are independent if:

$$P(A|B) = P(A)$$

$$P(A \cap B) = P(A)P(B)$$

Probability

Random variable:

The quantitative variable 'Y' is called a random variable when the value that 'Y' assumes in a given experiment is a chance or random outcome.

Random Variable

Discrete Random Variable:

When observations on a quantitative random variable can assume only a **countable** number of values.

Random Variable

Continuous Random Variable:

When observations on a quantitative random variable can assume any one of the uncountable number of values in a line interval.

We have drawn a sample from a population

We need to make an inference about the population

We need to know the probability of observing a particular sample outcome

We need to know the probability associated with each value of the variable ‘Y’

We need to know the probability distribution of the variable ‘Y’

The Binomial

A **binomial experiment** has the following properties:

- The experiment consists of 'n' identical trials.
- Each trial results in one of two outcomes (a success/a failure).
- The probability of success on a single trial is equal to π and π remains the same from trial to trial.

The Binomial

A binomial experiment has the following properties:

- The trials are independent; that is, the outcome of one trial does not influence the outcome of any other trial.
- The random variable 'Y' is the number of successes observed during the 'n' trials.

Probability Distributions

Discrete Random Variables

The Binomial

The probability of observing 'y' successes in 'n' trials of a binomial experiment is:

$$Pr(Y = y) = \frac{n!}{y!\,(n-y)!}\pi^{y}(1-\pi)^{n-y}$$

Where π is the probability of success.

Probability Distributions

Discrete Random Variables

The Binomial

The probability of observing ‘y’ successes in ‘n’ trials of a binomial experiment is:

$$Pr(Y = y) = \frac{n!}{y!\,(n-y)!}\pi^{y}(1-\pi)^{n-y}$$

Where π is the probability of success.

$$\mu = n\pi$$
$$\sigma = \sqrt{n\pi(1-\pi)}$$

The Poisson

Applicable for modeling of events of a particular time over a unit of time or space.

Probability Distributions

Discrete Random Variables

The Poisson

Let 'Y' be the number of events occurring during a fixed time interval of length 't'. Then the probability distribution of 'Y' is Poisson, provided following conditions:

- Events occur one at a time; two or more events do not occur precisely at the same time

The Poisson

Let ‘Y’ be the number of events occurring during a fixed time interval of length ‘t’. Then the probability distribution of ‘Y’ is Poisson, provided following conditions:

- Occurrence (or nonoccurrence) of an event during one period does not affect the probability of an event occurring at some other time.

Probability Distributions

Discrete Random Variables

The Poisson

Let 'Y' be the number of events occurring during a fixed time interval of length 't'. Then the probability distribution of 'Y' is Poisson, provided following conditions:

- The expected number of events during one period is the same as the expected number of events in any other period.

Probability Distributions

Discrete Random Variables

The Poisson

Let 'Y' be the number of events occurring during a fixed time interval of length 't'. Then:

$$Pr(Y = y) = \frac{\lambda^{y} e^{-\lambda}}{y!}$$

Probability Distributions

Discrete Random Variables

The Poisson

Let 'Y' be the number of events occurring during a fixed time interval of length 't'. Then:

$$Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

$\mu = \sigma = \lambda$

The Binomial & The Poisson

When 'n' is large and 'π' is small
in a binomial experiment,
the Poisson distribution (with $\lambda = n\pi$) provides a
good approximation to the
binomial distribution.

Probability Distributions
Continuous Random Variables

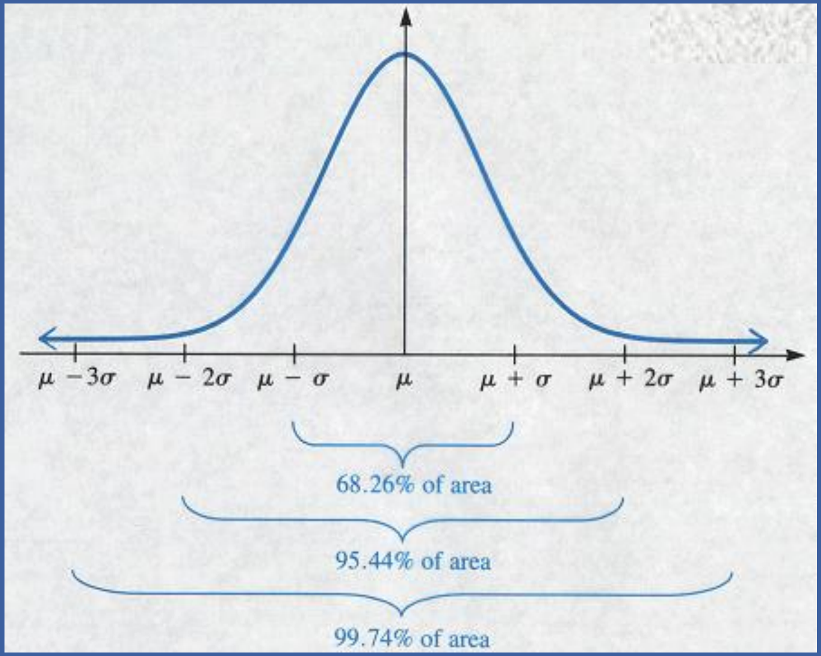
The Normal

Normal distribution (that has a smooth bell-shaped curve, symmetrical about the mean, 'μ') plays an important role in statistical inference.

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\mu)^2}{2\sigma^2}}$$

Probability Distributions
Continuous Random Variables

The Normal

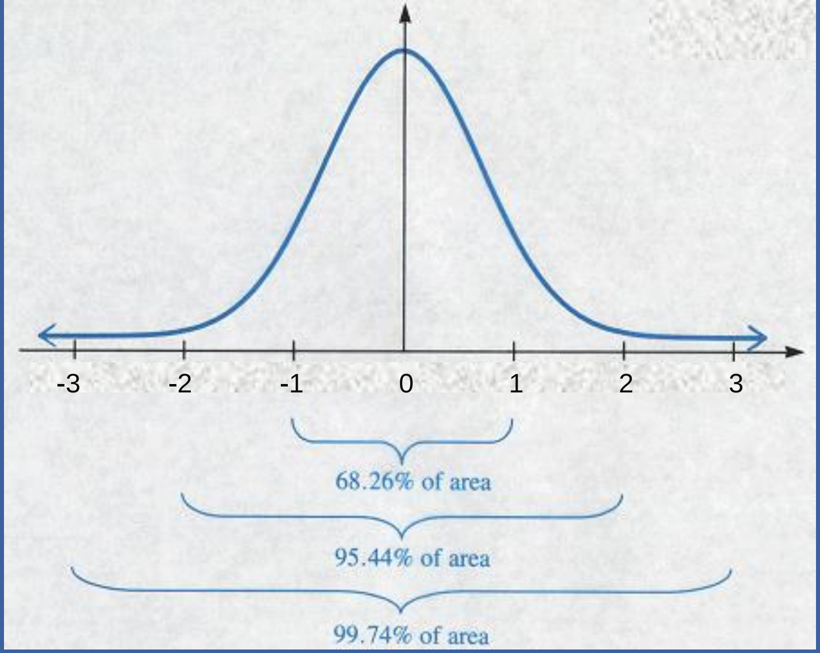


Empirical Law

Probability Distributions

Continuous Random Variables

The Normal



$$z = \frac{y - \mu}{\sigma}$$

Random Sampling

A sample of 'n' measurements selected from a population is said to be a random sample if every different sample of size 'n' from the population has a non-zero probability of being selected.

Random Sampling

A sample of 'n' measurements selected from a population is said to be a random sample if every different sample of size 'n' from the population has a non-zero probability of being selected.

Sample data selected in a nonrandom fashion are frequently distorted by a *selection bias*.
A selection bias exists whenever there is a systematic tendency to over-represent or under-represent some part of the population.

Random Sampling

Sample Statistic:

- Is a function of sample values
- Is a random variable
- It is subject to random variation because it is based on a random sample of measurements selected from the population of interest.
- Like any other random variable, has a probability distribution.

Random Sampling

Sample Statistic:

- Is a function of sample values
- Is a random variable
- It is subject to random variation because it is based on a random sample of measurements selected from the population of interest.
- Like any other random variable, has a probability distribution.

Sampling Distribution

Sampling Distribution

Central Limit Theorem (for $\bar{y}$):

Let:

- $\bar{y}$ be sample mean computed from a random sample of 'n' measurements from a population having a mean, μ and finite standard deviation σ
- $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$ be the mean and standard deviation of the sampling distribution of $\bar{y}$, respectively.

Based on repeated random samples of size 'n' from the population, we can conclude the following:

Sampling Distribution

Central Limit Theorem (for $\bar{y}$):

- $\mu_{\bar{y}} = \mu$
- $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$
- When 'n' is large the sampling distribution of $\bar{y}$ will be approximately normal.
- When the population distribution is normal, sampling distribution of $\bar{y}$ is exactly normal for any sample size 'n'.

Sampling Distribution

The Shape of Sampling Distribution is affected by

- Sample Size ‘n’
- Shape of distribution of population measurements

Sampling Distribution

The Shape of Sampling Distribution is affected by

- Sample Size 'n'
- Shape of distribution of population measurements

$$\begin{cases} \text{if symmetric, CLT hold for } n \geq 30 \\ \text{if heavily skewed, 'n' should be larger} \end{cases}$$

Sampling Distribution

Central Limit Theorem (for $\hat{y} = \sum y$):

Let:

- $\hat{y}$ be the sum of a random sample of 'n' measurements from a population having a mean, μ and finite standard deviation σ
- $\mu_{\hat{y}}$ and $\sigma_{\hat{y}}$ be the mean and standard deviation of the sampling distribution of $\hat{y}$ respectively.

Based on **repeated random samples of size 'n'** from the population, we can conclude the following:

Sampling Distribution

Central Limit Theorem (for $\bar{y}$):

- $\mu_{\hat{y}} = n\mu$
- $\sigma_{\hat{y}} = \sqrt{n\sigma}$
- When 'n' is large the sampling distribution of $\hat{y}$ will be approximately normal.
- When the population distribution is normal, sampling distribution of $\hat{y}$ is exactly normal for any sample size 'n'.

Sampling Distribution

Central Limit Theorem (for $\bar{y}$):

- $\mu_{\hat{y}} = n\mu$
- $\sigma_{\hat{y}} = \sqrt{n\ldots}$
- When 'n' i... be app... W... sampling dist... any sample size

Similar theorems exist for the sample median, sample standard deviation, and the sample proportion.

- We have drawn a sample from a population
- We need to make an inference about the population
- We use sample statistic to estimate a population parameter
- We need to know how accurate the estimate is.
- We need to know the sampling distribution
- We seldom know the sampling distribution
- We use normal approximation from CLT

Be aware of the unfortunate similarity between two phrases:

"Sampling Distribution"

(the theoretically derived probability distribution of a statistic)

"Sample Distribution"

(the histogram of individual values actually observed in a particular sample)

Sampling Distribution

Normal Approximation to the Binomial Probability Distribution

For large ‘n’ and ‘π’ not too near 0 or 1, the distribution of a binomial random variable ‘Y’ may be approximated by a normal distribution with $\mu = n\pi$ and $\sigma = \sqrt{n\pi(1-\pi)}$

Normal Approximation to the Binomial Probability Distribution

This approximation should be used only if

$$n\pi \geq 5 \text{ and } n(1-\pi) \geq 5$$

Sampling Distribution

Normal Approximation to the Binomial Probability Distribution

This approximation should be used only if

$$n\pi \geq 5 \text{ and } n(1-\pi) \geq 5$$

Actual binomial distribution is seriously skewed to right

Actual binomial distribution is seriously skewed to left

Sampling Distribution

Why normality is important:

- Helps to draw inferences about population based on the sample

- Most statistical procedures require that population distribution be normal or can adequately be approximated by a normal distribution

Sampling Distribution

Tools for Evaluating Whether or Not a Population Distribution Is Normal

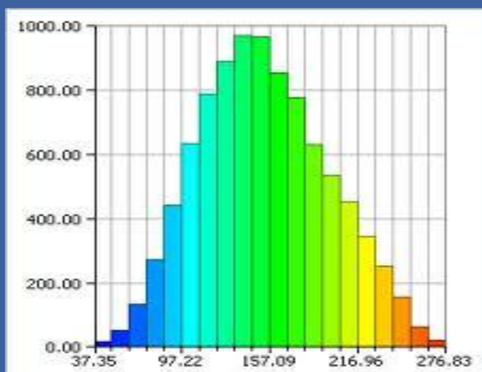
- Graphical Procedure, &
- Quantitative Assessment

Of how well a normal distribution models the population distribution

Checking Normality

Graphical Procedures

Histogram



Stem-and-leaf plot

8.	0	0					
9.	0						
10.	0	0					
11.	0	0	5				
12.	0	0	0	2			
13.	2	5	8	8			
14.	0	0	0	0	4	6	8
15.	0	0	5				
16.	0	2	6	8			
17.	0	0	5				
18.	0	2	5				
19.	0	5					
20.	0	5					

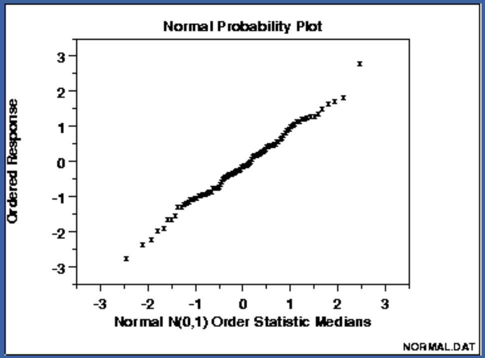
Checking Normality

Graphical Procedures

Normal Probability Plot

Compares the quantiles from the data observed from the population to the corresponding quantiles from the standard normal distribution.

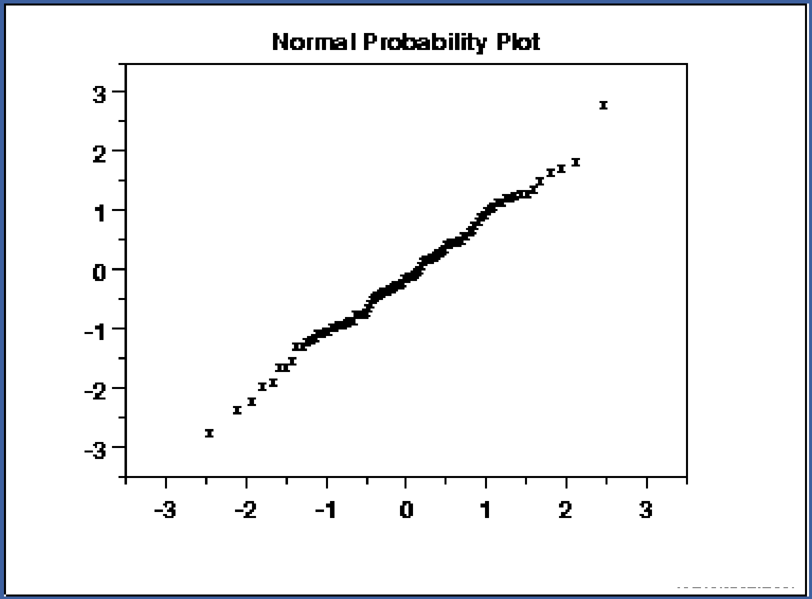
- Sort the data: $y_{(1)}, y_{(2)}, \dots, y_{(n)}$
- $y_{(i)} = Q\left(\frac{i-0.5}{n}\right)$
- Plot $Q\left(\frac{i-0.5}{n}\right)$ versus $z_{\left(\frac{i-0.5}{n}\right)}$



Checking Normality

Quantitative Assessment

Correlation Coefficient of $Q\left(\frac{i-0.5}{n}\right)$ versus $z_{\left(\frac{i-0.5}{n}\right)}$



Checking Normality

Quantitative Assessment

- Kolmogorov-Smirnov
- Shapiro Wilk
- Shapiro Francia
- Cramer-von Mises
- Anderson-Darling