

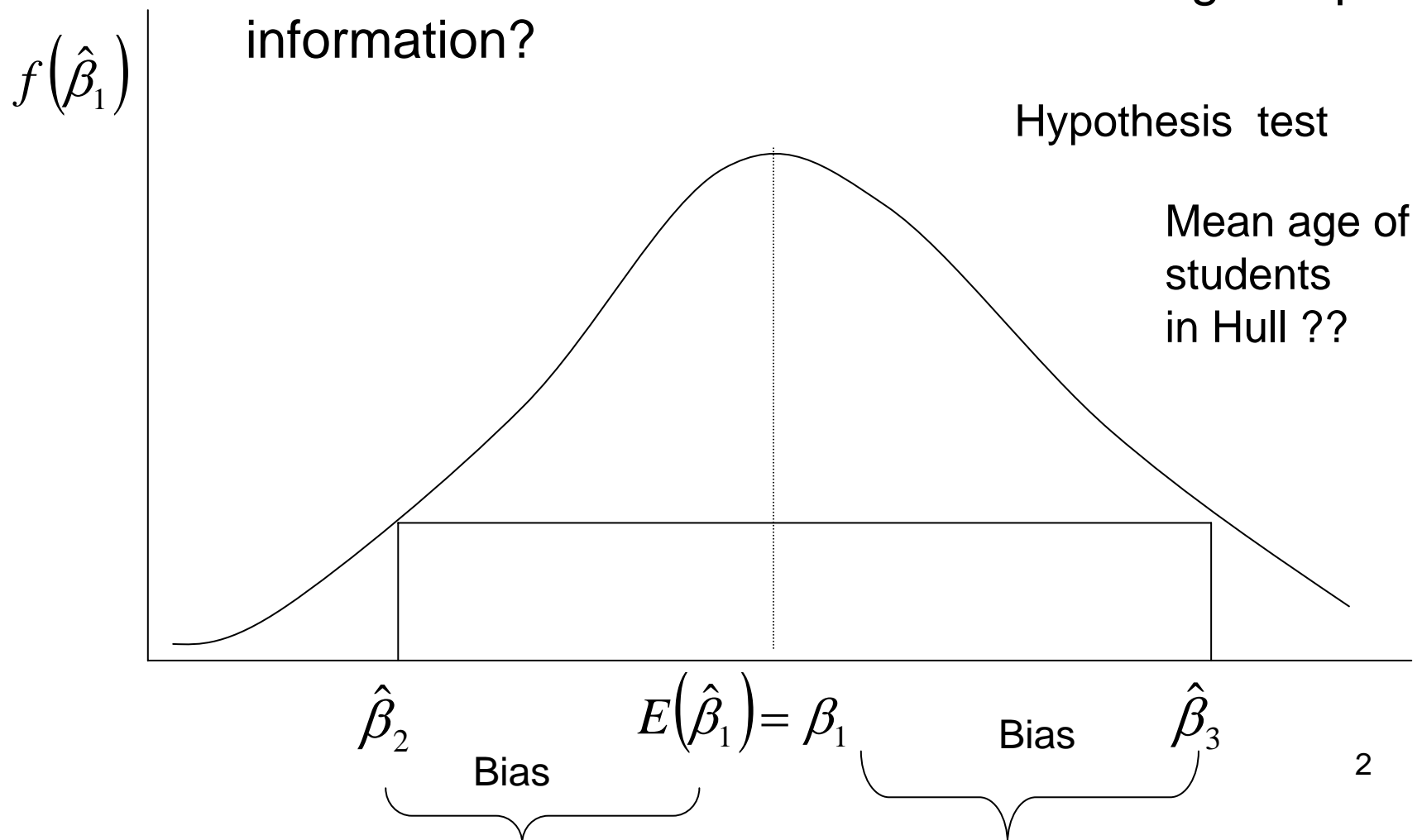
Research Methods For Economists

Lecture 2

Sampling and Frequency Distributions
Mean, Variance, Covariance, Correlation and
Statistical Tests: An Introduction

What is the best way to use sample information to infer about the characteristic of population?

How to minimise the bias or error in using sample information?



Two types of random variables

- Discrete rv: takes only countable number of values
 - Number of workers in an office or individuals in a house.
 - Number of voters in a ward.
 - Number of students in a lecture.
 - Number of vehicles in a parking lot or in a road.
- Continuous: can take any value in an interval
 - Income of individuals; profits or revenue of firms, consumer surplus.
 - Age, weight, height, size of individuals.
 - Output, employment, exchange rate, interest rate, credits, loans.
 - Amount of water, drinks sold/consumed.
 - Oil used by vehicles.
 - Hours of study, sleep and work.

Types of Data: Cross Section, Time series and Panel data

Sample and Population of Random Variables

- Sample
- A small proportion of population
 - Strategic
 - random
 - Stratified
 - With replacement
 - Without replacement
- Monte Carlo Simulation
- Population (Sample Space)
 - Contain all possible elements, universe.
 - Described in terms of mean and variance and other moments.
 - They are assumed to be known.
- Theoretical distributions
 - Binomial, Poisson
 - Normal, Chi-Square, T, F.

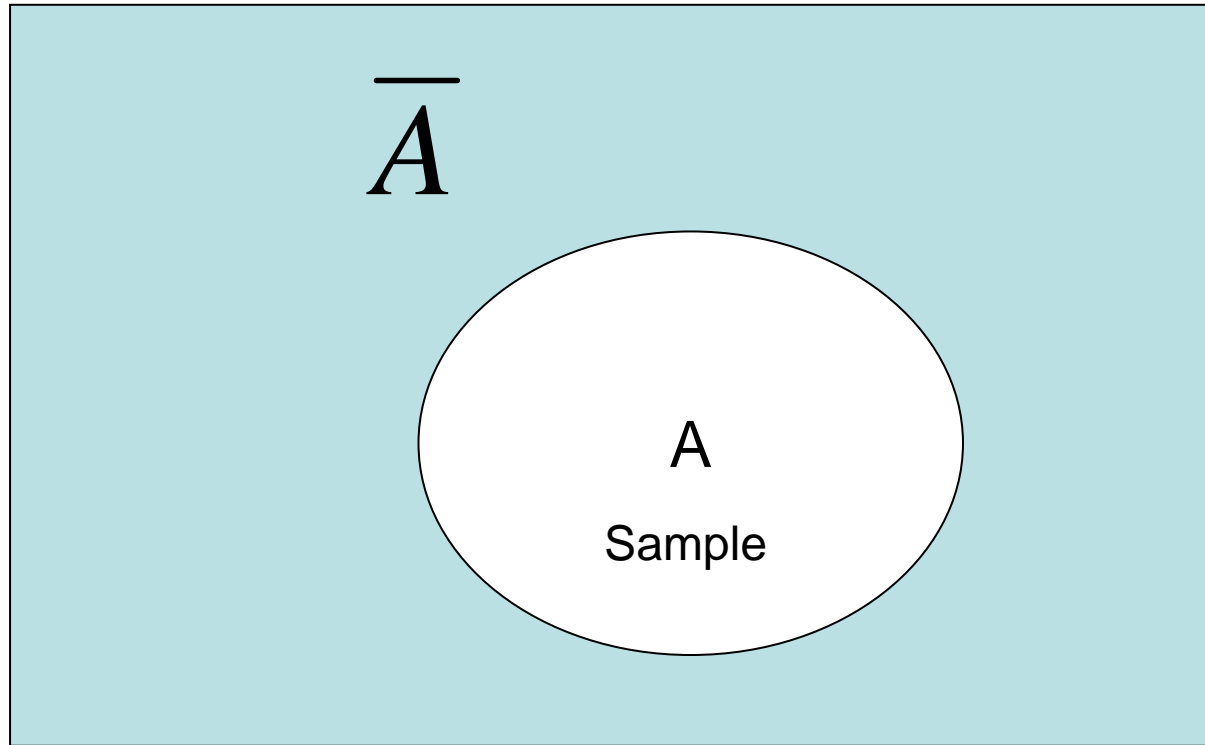
Population: eligible voters 11 regions in the UK

Six wards in Hull.

Sample: Take any five regions or four wards at random.

How many samples can you make?

Sample and Sample Space: using a Venn Diagram



Permutations

$$Pr = \frac{n!}{(n-r)!}$$

Combinations

$$Cn = \frac{n!}{(n-r)!r!}$$

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

Probability : a. Relative Frequency b. Subjective belief

Probability of Mutually Exclusive Events

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

Probability of two overlapping events

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

Probability of independent events

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

See Newbold et. al (2003) Statistics for Business and Economics, Prentice Hall.

Sample and Population: Some Sample Design from UK regions

What is the mean household income in the UK?

Regions: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Taking five regions at a time

Sample (5): 1, 2, 3, 4, 5;

1, 2, 3, 4, 6;

1, 2, 3, 4, 7;

1, 2, 3, 4, 8;

1, 2, 3, 4, 9;

1, 2, 3, 4, 10;

1, 2, 3, 4, 11;

.....

7, 8, 9, 10, 11

$$C_n = \frac{n!}{(n-r)!r!} = \frac{11!}{(11-5)!5!} = 462$$



<http://www.statistics.gov.uk>

England

MaP Source: <http://www.GraphicMaps.Com>

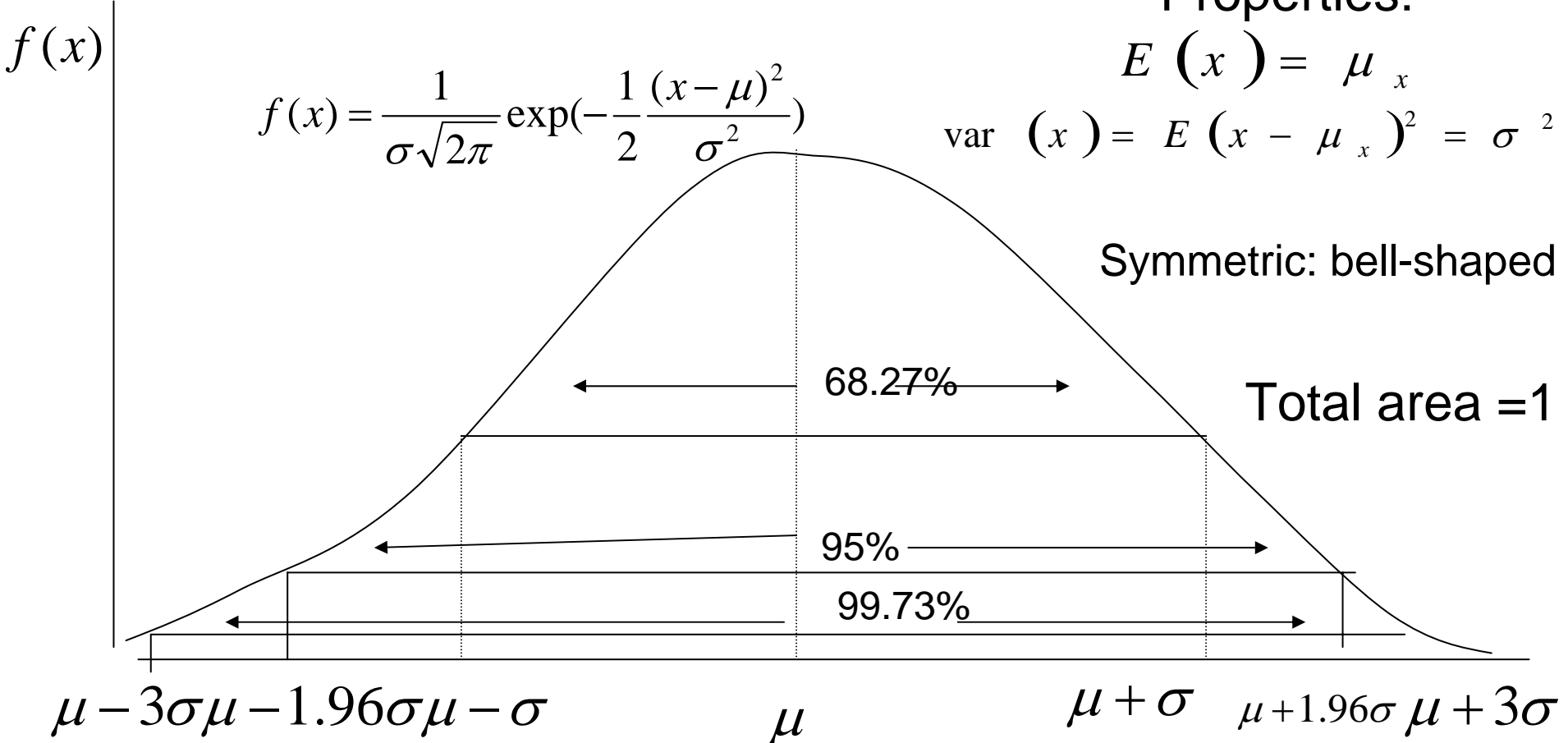
BHPS: British Household Panel Studies is based on samples.

Normal Distribution of X: Bell Shaped Distribution

Properties:

$$E(x) = \mu_x$$

$$\text{var}(x) = E(x - \mu_x)^2 = \sigma^2$$



Symmetric: bell-shaped

Total area = 1

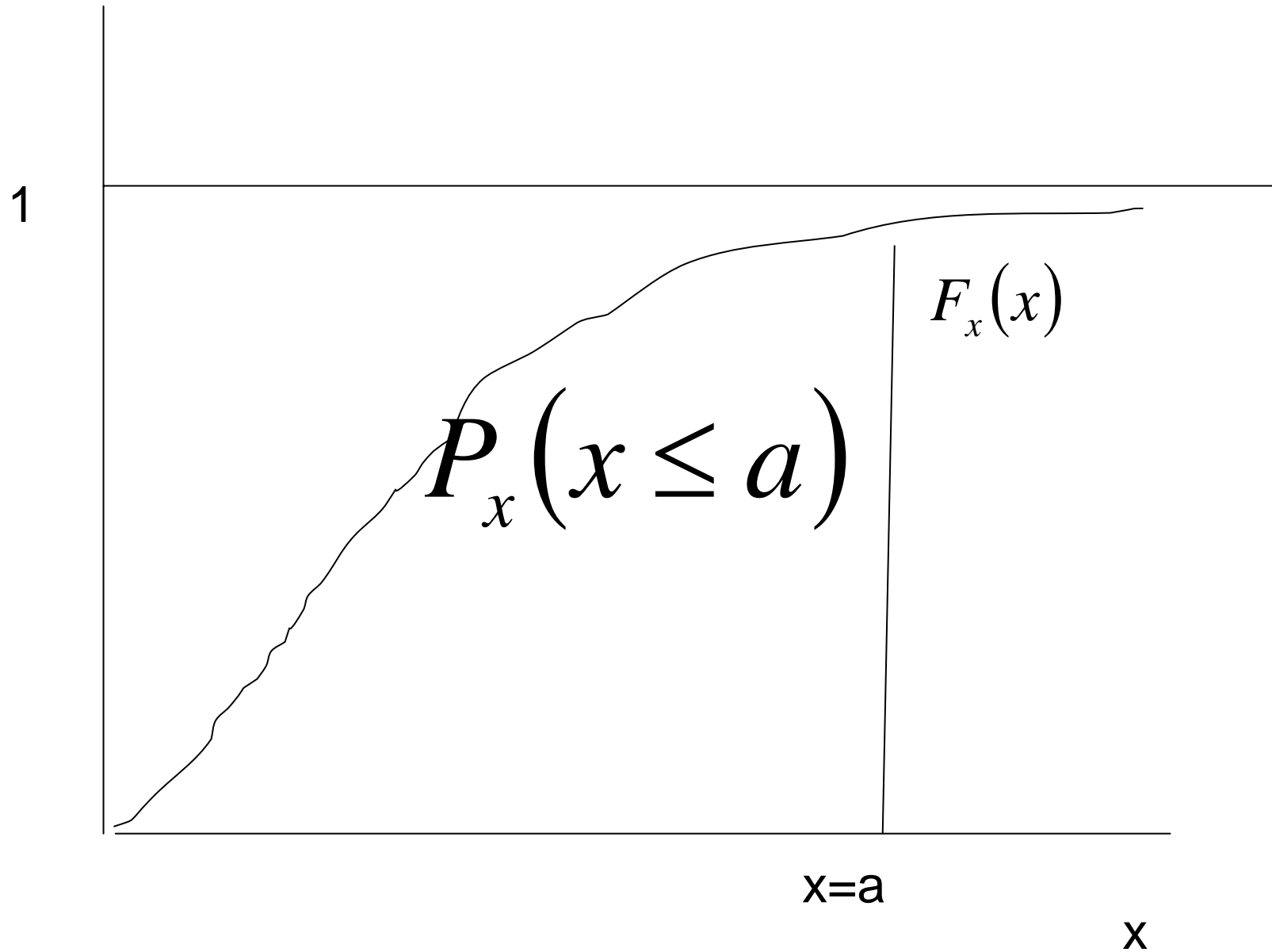


Standard Normal

$$z = N(0,1)$$

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

Cumulative Density Function



Central Limit Theorem

From any variable x , subtracting its mean and dividing the result by its standard deviation yields a random variable with mean zero and variance 1.

$$X = X_1 + X_2 + X_3 + \dots + X_n$$

$$E(X_i) = \mu \quad \text{var}(X_i) = \sigma^2$$

$$E(X) = n\mu \quad \text{var}(X) = n\sigma^2$$

$$Z = \frac{X - n\mu}{\sqrt{n\sigma^2}} \quad Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

All distributions converge to the Normal distribution in the Limit

Transformation to Standard Normal Distribution: Normalization

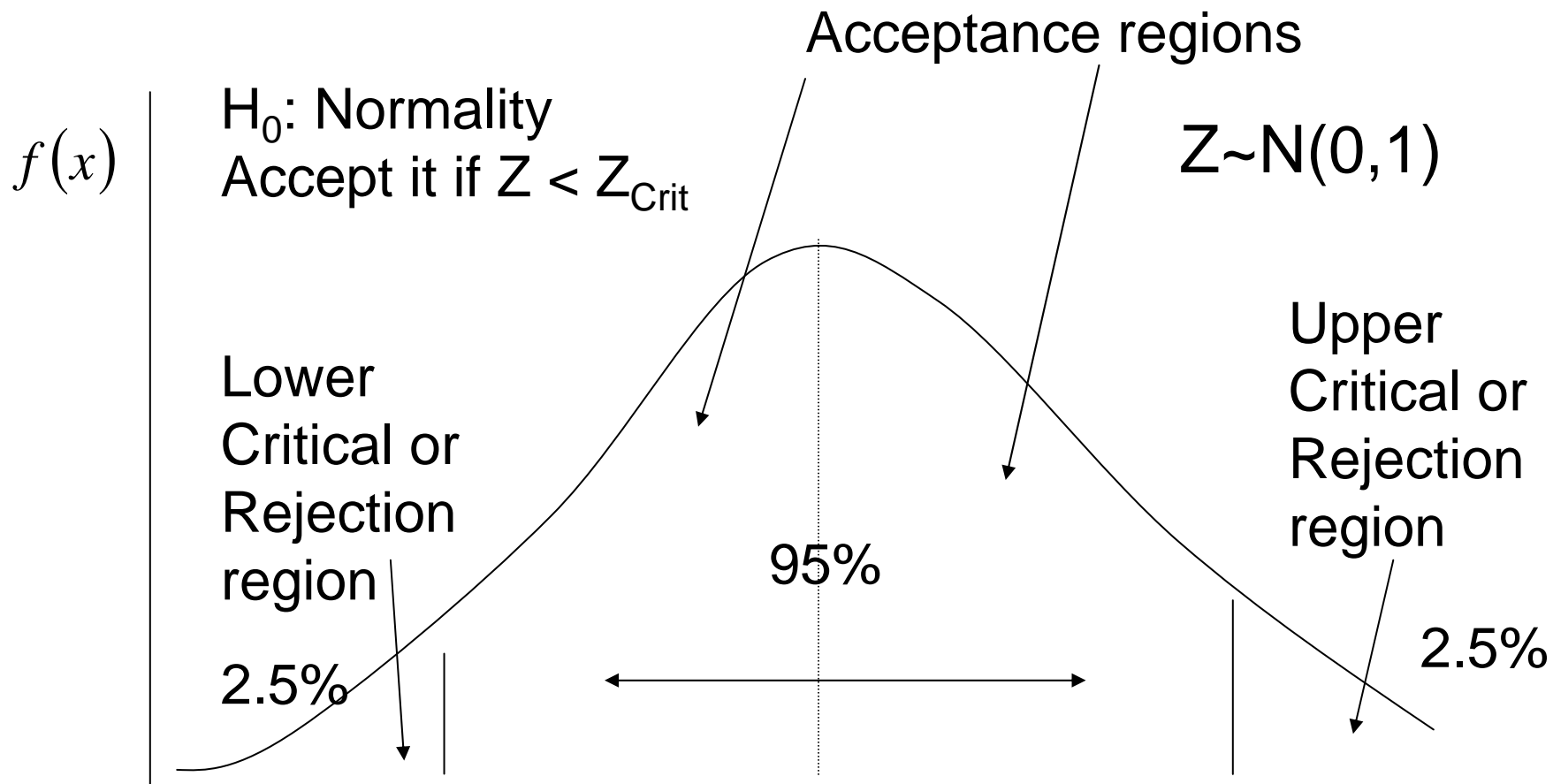
$$z = \frac{X - \mu}{\sigma} = \frac{X - \bar{X}}{SE(X)} \sim N(0,1)$$

Z-statistics calculated from the sample information is compared to theoretical distribution to decide whether a certain sample is drawn from Normal distribution.

Example: score in the exam.

Test of Normality and Level of Significance: Two Tail Test

$$P(1.96 \leq z \leq 1.96) = (1 - \alpha) = 0.95$$



$$\alpha/2$$

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

$$\alpha/2_{12}$$

Confidence Interval of Mean of a Normally Distributed Variable

Standard normal:

$$P(1.645 \leq z \leq 1.645) = (1 - \alpha) = 0.9$$

$$P(1.96 \leq z \leq 1.96) = (1 - \alpha) = 0.95$$

$$P(2.58 \leq z \leq 2.58) = (1 - \alpha) = 0.99$$

For a sample of size n

$$P(\bar{X} - 1.645 \sigma/n \leq \mu \leq 1.645 \sigma/n + \bar{X}) = (1 - \alpha) = 0.9$$

$$P(\bar{X} - 1.96 \sigma/n \leq \mu \leq 1.96 \sigma/n + \bar{X}) = (1 - \alpha) = 0.95$$

$$P(\bar{X} - 2.56 \sigma/n \leq \mu \leq 2.56 \sigma/n + \bar{X}) = (1 - \alpha) = 0.99$$

Ideal Properties and Problems of a good estimate

1. Unbiased-ness

$$E(x) = \mu$$

2. Efficiency

$$\text{var}(x) \leq \text{var}(\tilde{x})$$

3. Consistency

$$\frac{\text{Lim}}{T \rightarrow \infty} \text{bias} = [E(x) - \mu] = 0$$

1. Problems

- Measurement errors
- Outliers
- Biased samples
- Incorrect formula
- Bugs in calculations

2. Solutions

- Statistical tests
- Take many samples before a conclusion

Sample Estimators of Population Parameters

<http://europa.eu.int/comm/trade/issues/bilateral/dataxls.htm>

Mean

$$\bar{X} = \frac{\sum X_i}{N}$$

First Moment:

μ_1

Variance

$$\text{var}(X) = \frac{\sum (X_i - \bar{X})^2}{N - 1}$$

Second Moment:

μ_2

Standard Dev

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N - 1}}$$

Skewness (Pearson)=
(Mean-Mode)/stand.dev.

Skewness:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\mu_3 = \frac{\sum (X_i - \bar{X})^3}{N}$$

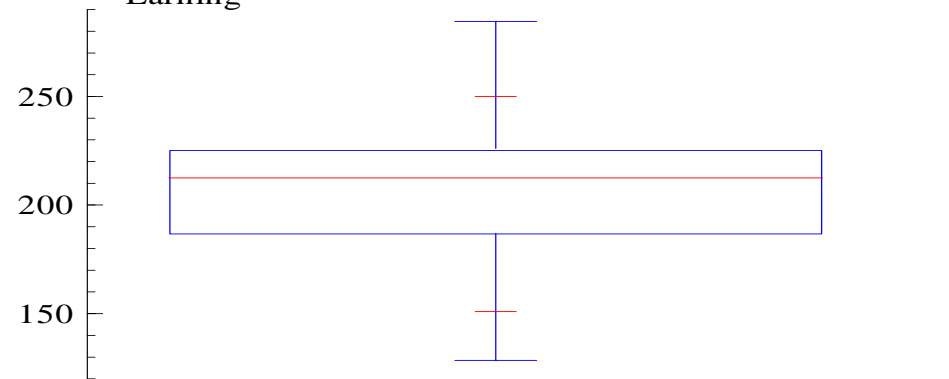
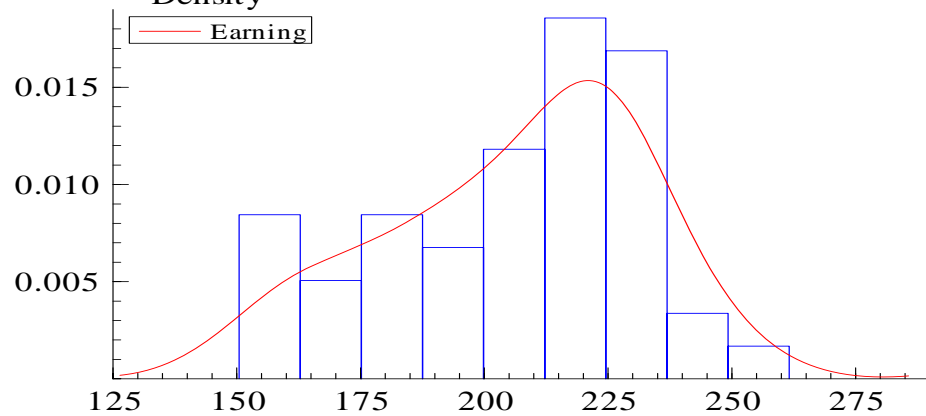
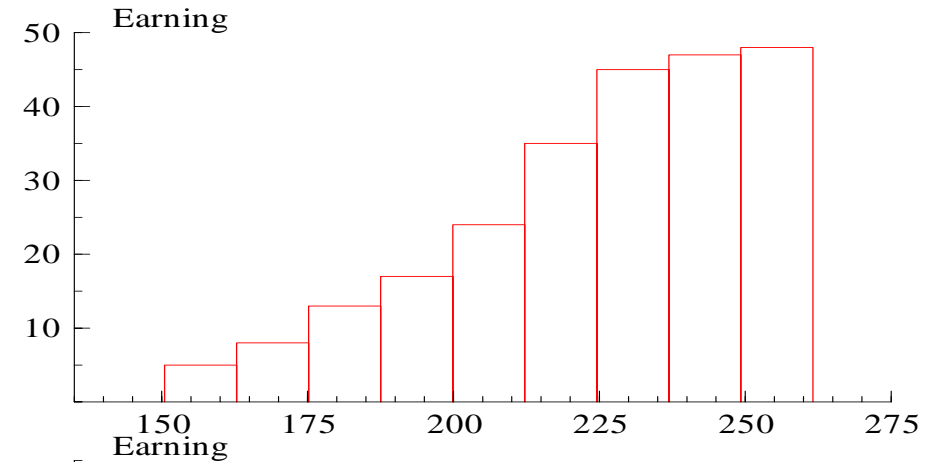
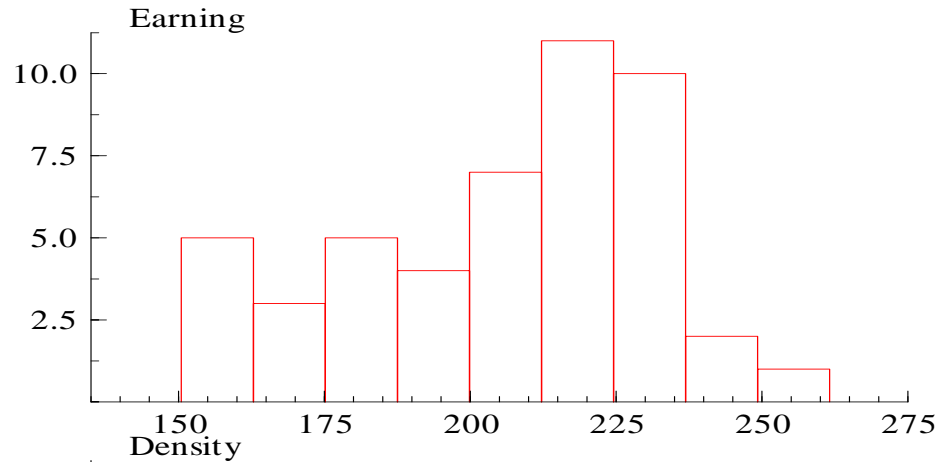
Kurtosis:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$\mu_4 = \frac{\sum (X_i - \bar{X})^4}{N}$$

Normality implies: $\sqrt{\beta_1} = 0$ $\beta_2 = 3$

A Random Variable Can be Represented in a Frequency Diagrams



The Frequency distribution can be expressed in terms of its statistics such as mean, median, model variance, other moments.

Tests for Normality of Score 1 and Earnings with transformed Skewness and Kurtosis

- Normality test for Exam1
- Observations 48
- Mean 57.000
- Std.Devn. 11.301
- Skewness -2.5663
- Excess Kurtosis 8.0476
- Minimum 5.0000
- Maximum 71.000
- Asymptotic test: Chi²(2) = 182.21 [0.0000]**
- Normality test: Chi²(2) = 61.172 [0.0000]**

Exam 1 is not Normal.

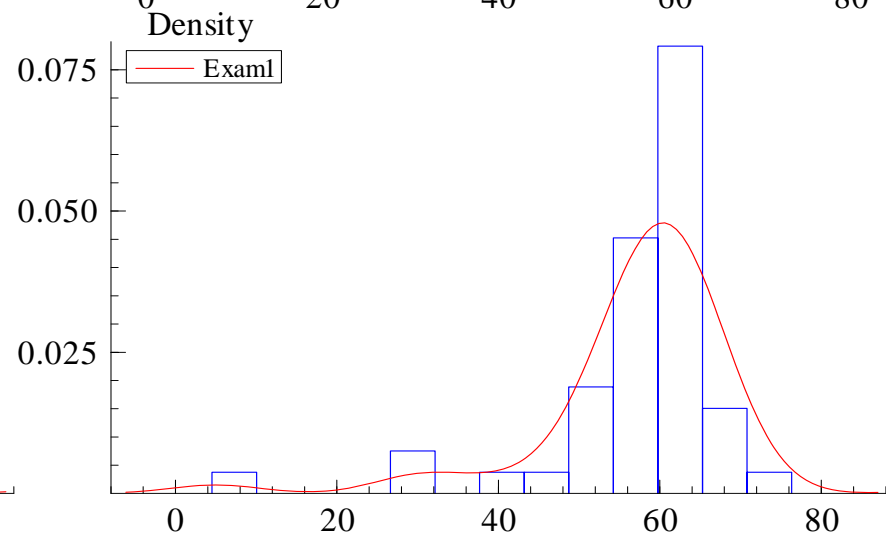
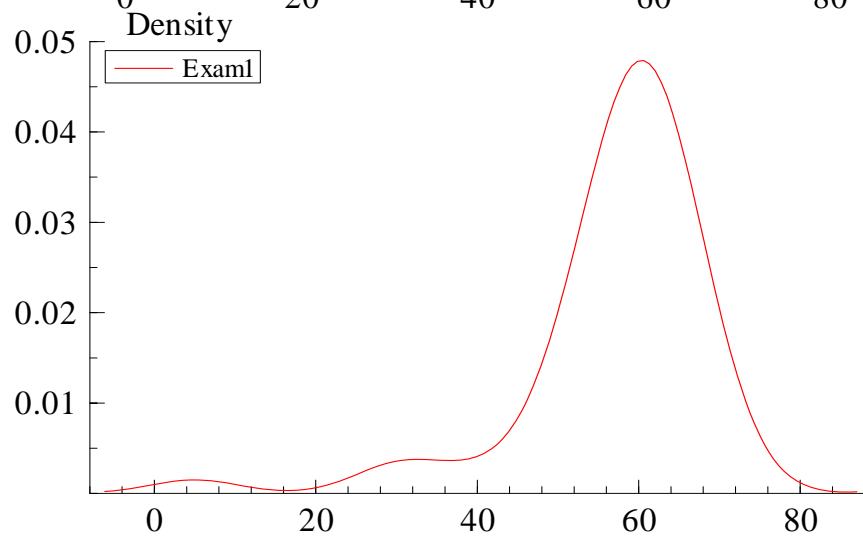
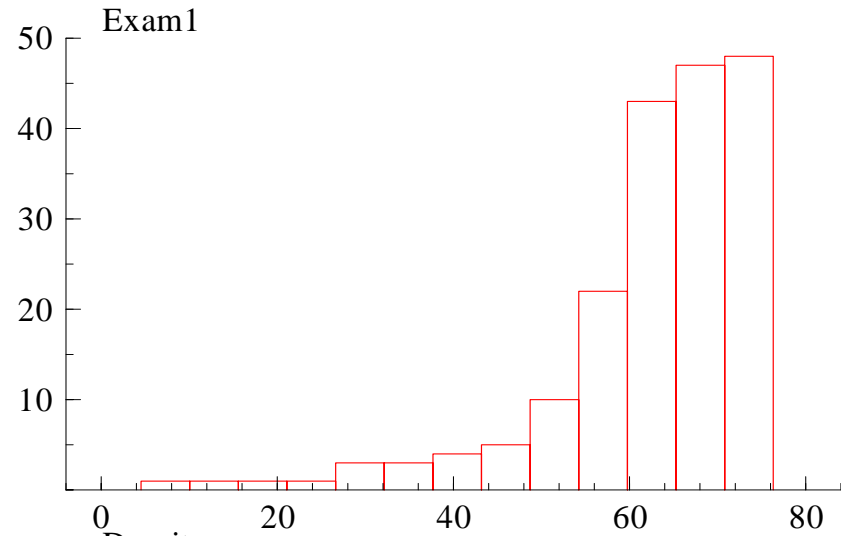
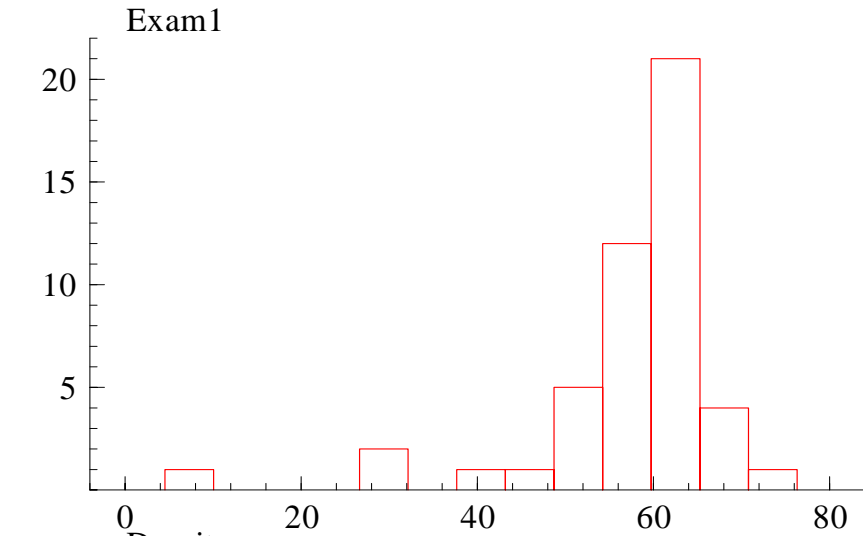
- Normality test for Earning
- Observations 48
- Mean 195.02
- Std.Devn. 28.165
- Skewness 0.25171
- Excess Kurtosis -0.89966
- Minimum 153.00
- Maximum 250.00
- Asymptotic test: Chi²(2) = 2.1257 [0.3455]
- Normality test: Chi²(2) = 2.9870 [0.2246]

Earning is Normal.

Formula Underlying the Normality Test refined form of following (PCgive p.261)

$$e_1 = \frac{T(\sqrt{b_1})}{6} + \frac{T(b_2 - 3)^2}{24} \sim \chi^2(2)$$

Frequency Diagrams prepared with PcGive



Means, standard deviations and correlations (using scores.xls)

The sample is 1 - 48		
Means		
Earning	Exam1	Exam2
206.04	57.000	5.5313

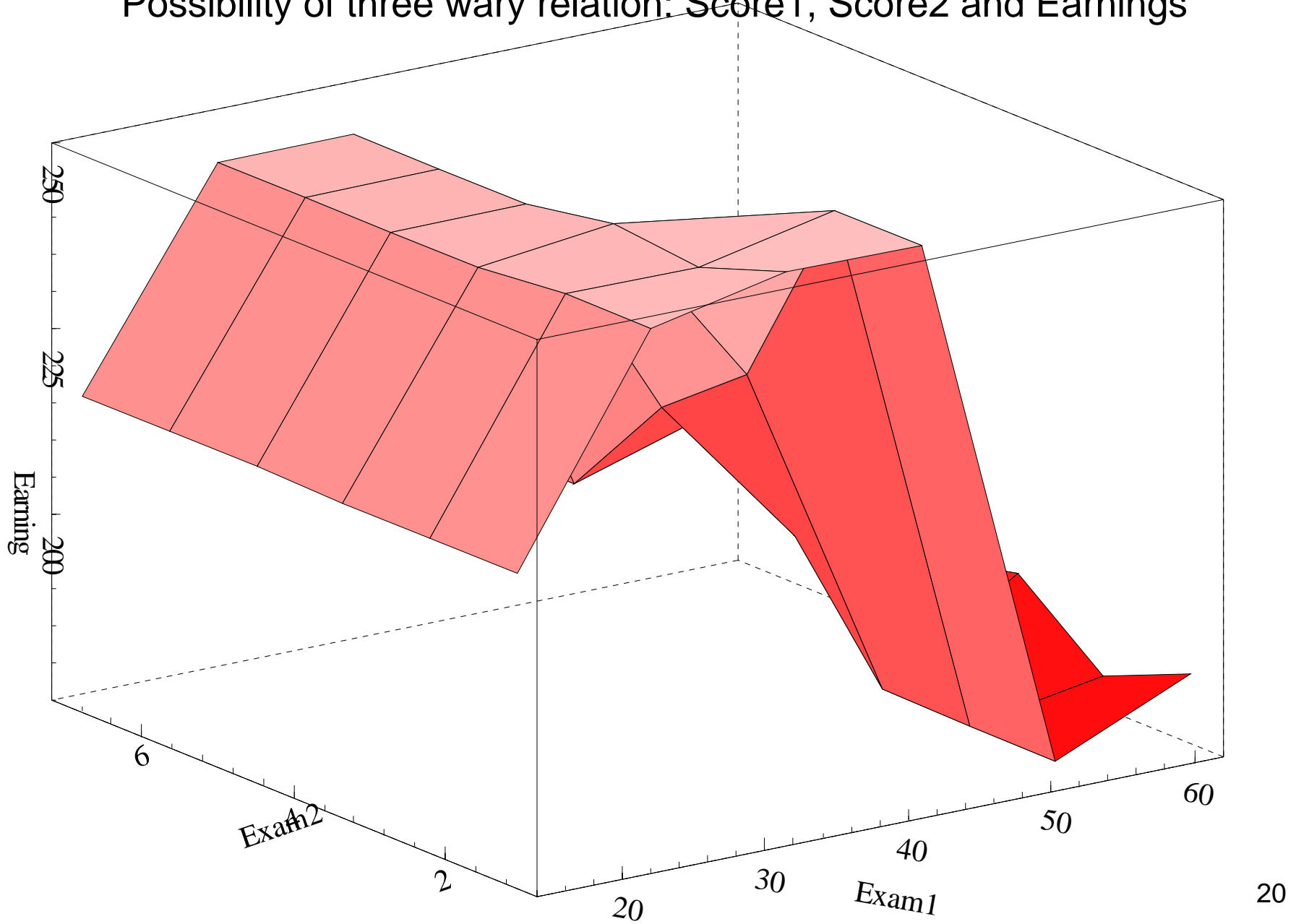
Standard deviations (using T-1)		
Earning	Exam1	Exam2
25.529	11.420	1.3968

	Correlation matrix:		
	Earning	Exam1	Exam2
Earning	1.0000	-0.13800	-0.26191
Exam1	-0.13800	1.0000	0.75784
Exam2	-0.26191	0.75784	1.0000

$$\rho_{x,y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\text{var}(X)}\sqrt{\text{Var}(Y)}}$$

Task: Using above means and standard deviation, find the standard normal distribution for above variables

Possibility of three way relation: Score1, Score2 and Earnings



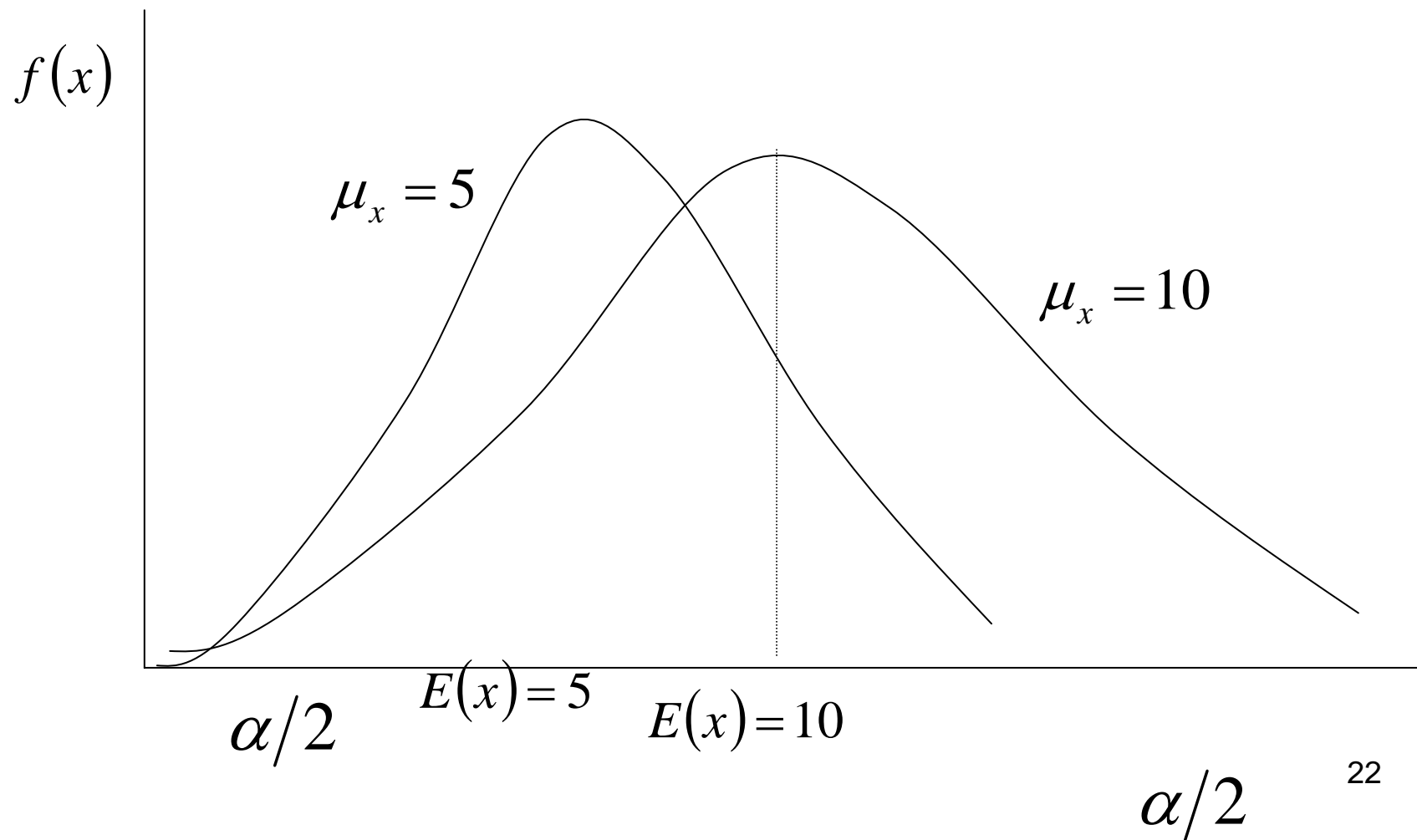
Tests Using Excel (Tools/data analysis)

- ANOVA
- Chitest
- Correlation
- Covariance
- Descriptive statistics
- F-test for sample variance
- Histogram
- Moving average
- Rank and percentile
- Regression
- Sampling
- T-test: mean test
- Z-test: mean test

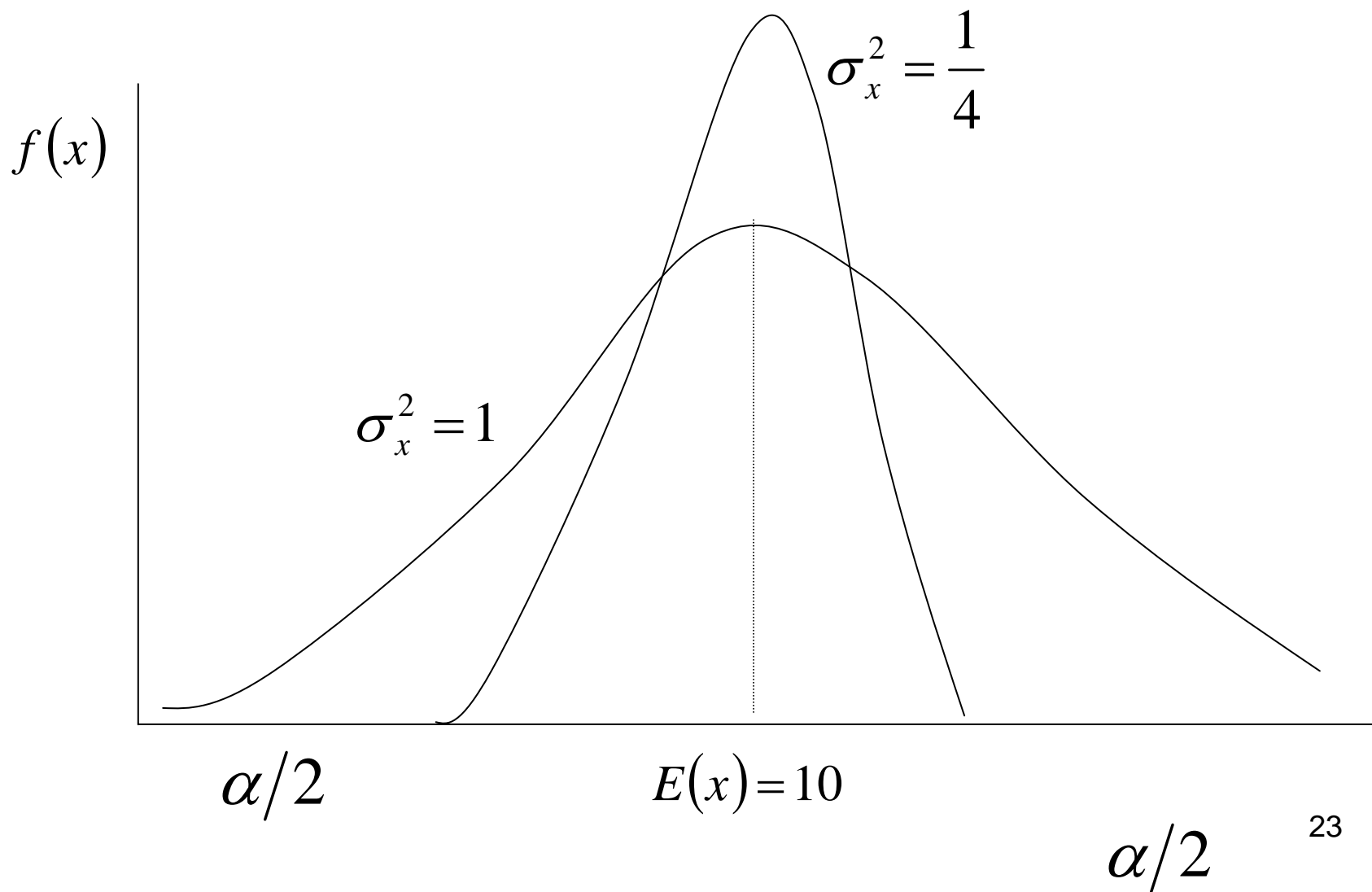
Mean	195.021
Standard Error	4.108
Median	192.500
Mode	189.000
Standard Deviation	28.463
Sample Variance	810.148
Kurtosis	-0.865
Skewness	0.260
Range	97.000
Minimum	153.000
Maximum	250.000
Sum	9361.000
Count	48.000
Largest(1)	250.000
Smallest(1)	153.000
Confidence Level(95.0%)	8.265

@NORMDISR(x, mean, steev, Cum)

Normal Distribution with different means



Normal Distribution with different Variances



t-Distribution

When variance of a normal distribution is unknown it is approximated by *t*-distribution with a standard deviation from the sample

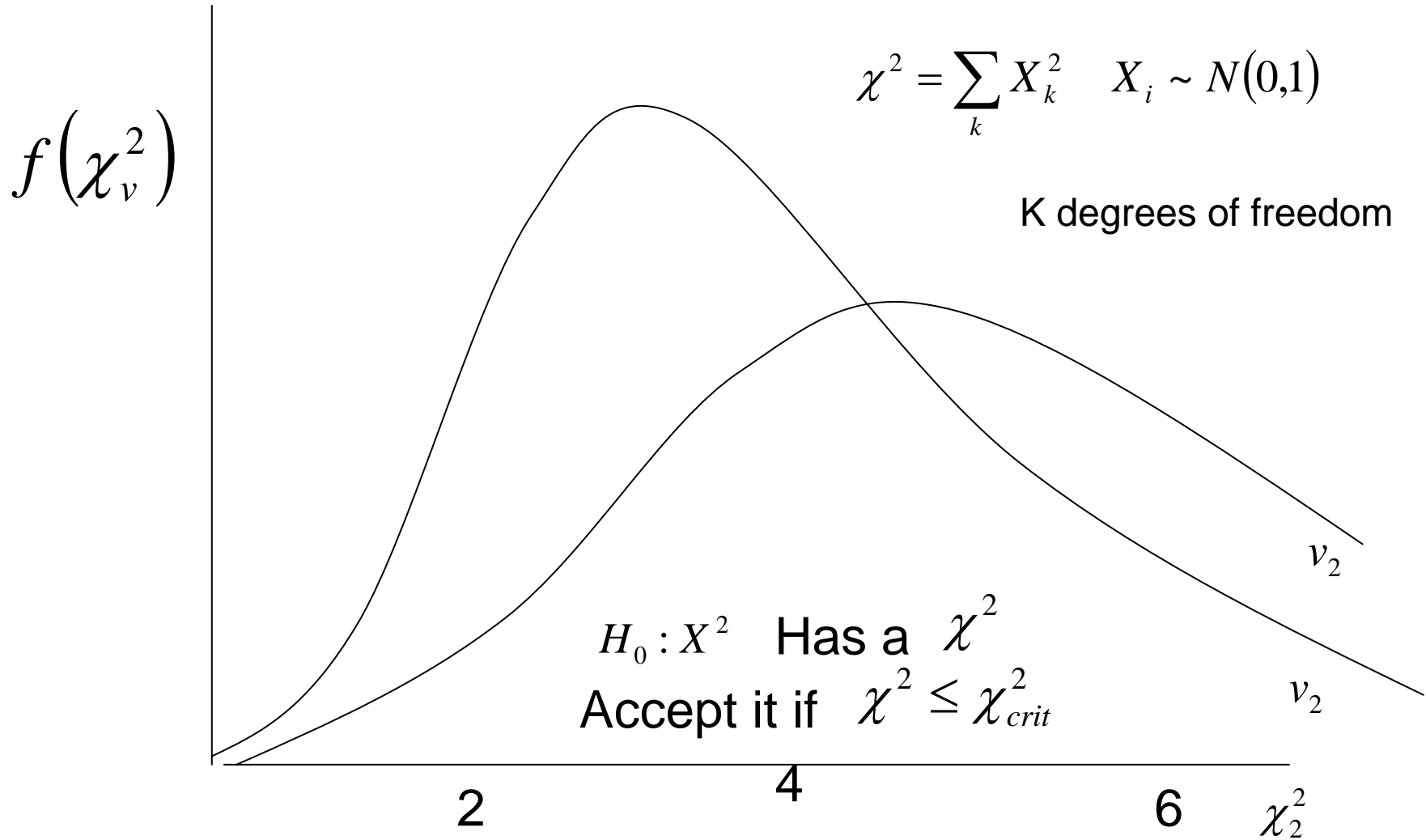
$$t_{n-k-1} = \frac{\bar{X} - \mu}{s_x / \sqrt{n}}$$

t-distribution is used to test the hypothesis particularly for the sample size less than 30.

$$H_0 : \bar{X} = \mu \quad \text{Accept it if } t < t_{\text{crit}}$$

χ^2 Distribution

If X is normally distributed sum of its square has Chi-square with K degrees of freedom



Critical value of Chi-square depends on degrees of freedom and the level of significance

F- Distribution: Variance Test

$$F = \frac{V_1/m_1}{V_2/m_2} \sim F(m_1, m_2)$$

H0 Variance are the same.
 Accept it if $F < F_{crit}$.

V_1 Variance of numerator

V_2 Variance of denominator

m_1 Degree of freedom of numerator

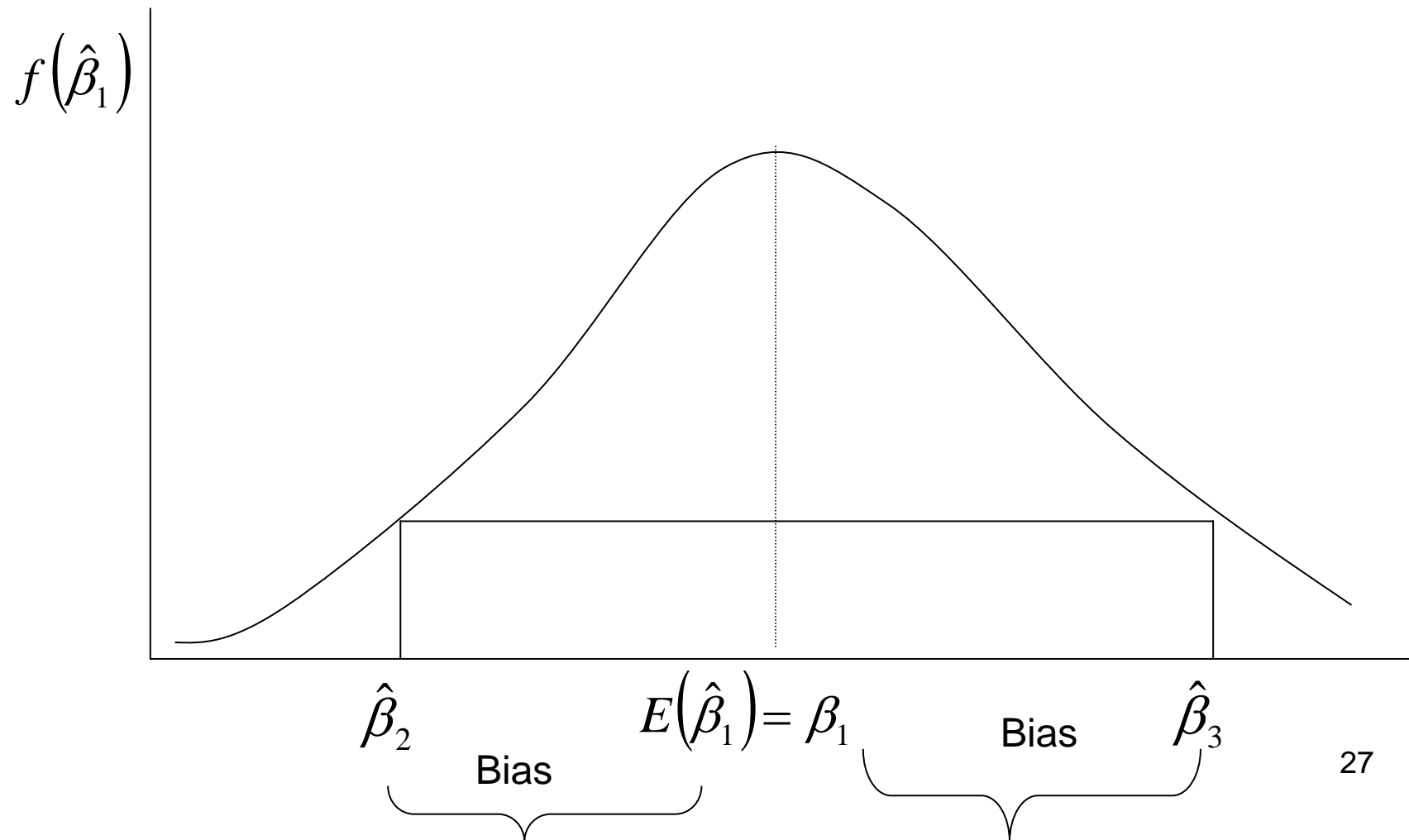
ANOVA

m_2 Degree of freedom of denominator

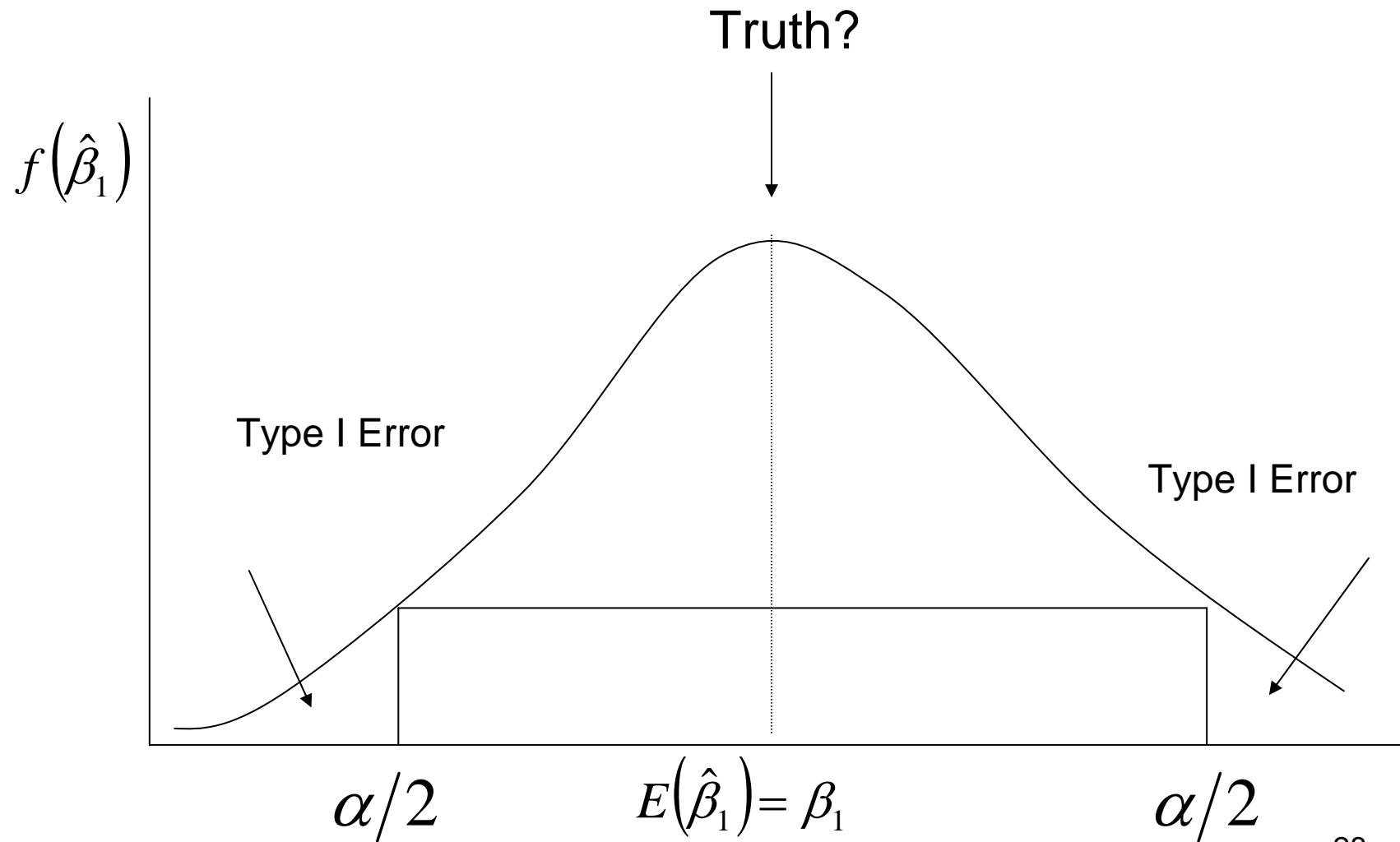
Source of Variation	SS	df	MS	F	P-value	F crit
Rows	105.05	9.00	11.67	1.61	0.25	3.18
Columns	4.05	1.00	4.05	0.56	0.47	5.12
Error	65.45	9.00	7.27			
Total	174.55	19.00				

X	Y
2	3
3	4
1	1
7	7
8	9
9	4
5	10
4	5
9	2
10	4

Unbiasedness of an Estimator



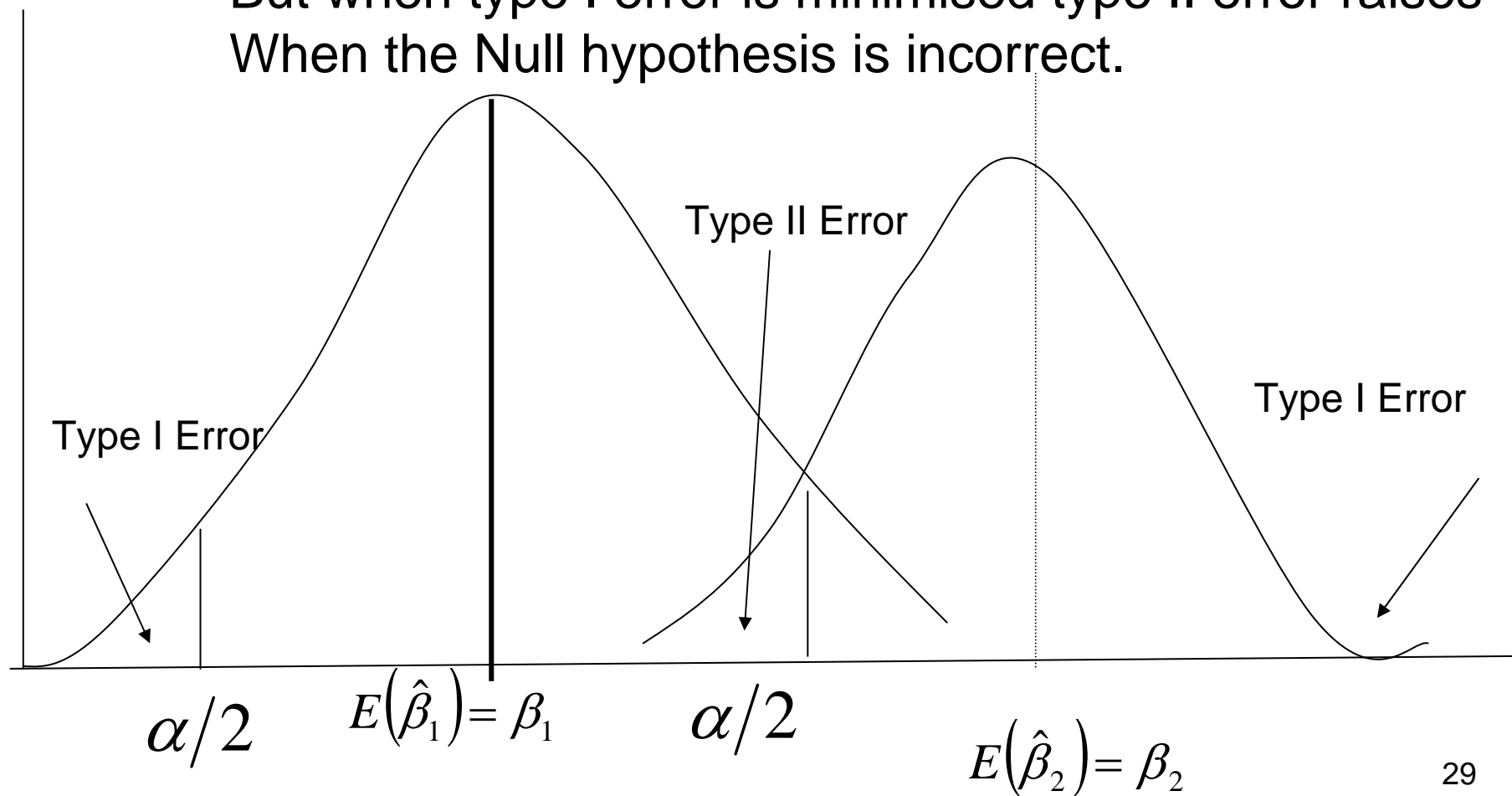
Type I Error: Rejecting a Null Hypothesis when it is true



Type II Error: Accepting a Null Hypothesis When it is False

$f(\hat{\beta}_1)$

Type I error is more serious than type II error.
But when type I error is minimised type II error raises
When the Null hypothesis is incorrect.



Type I and Type II Error in Hypothesis Testing

	True	False
Accept	Correct Decision	Type II Error β
Reject	Type I Error	Correct Decision

P-value:
Probability of
Text statistics
Exceeding that
Of the sample.

α Size of test.

Type II occurs when the
Null hypothesis is wrong.

Power of test: probability of rejecting the null while it is false.

$$\text{Power} = 1 - \beta = 1 - \text{Prob (type II error)}$$

Covariance Correlation and Regression

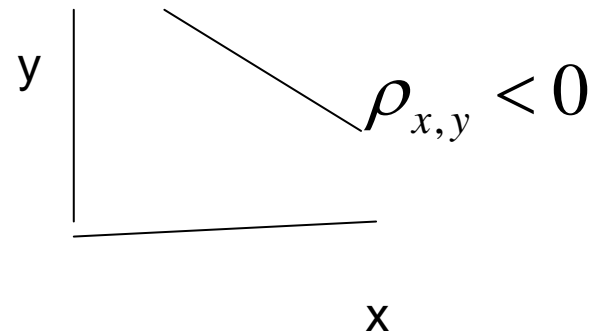
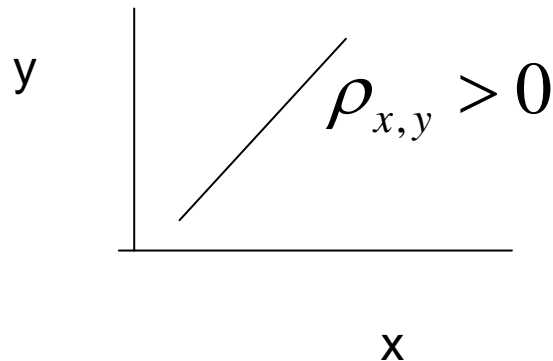
Regression
$$Y_i = \beta_1 + \beta_2 K_i + \beta_3 L_i + e_i$$

Covariance:
$$Cov(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

Correlation:
$$\rho_{x,y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\text{var}(X) \text{Var}(Y)}}$$

Spearman's rank $r_s = 1 - \frac{6 \sum d^2}{N^3 - N}$

Correlation:
$$-1 \leq \rho_{x,y} \leq 1$$



Using Stata for Analysis

```
spearman saleprice lotsize bedroom bath bath stories driveway recroom basement gas  
aircond g  
> arage desireloc  
(obs=546)
```

```
-----+-----  
| salepr~e lotsize bedroom  bath  bath stories driveway recroom  
-----+-----  
saleprice | 1.0000  
lotsize | 0.6030 1.0000  
bedroom | 0.3903 0.1763 1.0000  
  bath | 0.4787 0.2136 0.3769 1.0000  
  bath | 0.4787 0.2136 0.3769 1.0000 1.0000  
stories | 0.3647 0.0588 0.5033 0.3066 0.3066 1.0000  
driveway | 0.3404 0.3348 -0.0042 0.0489 0.0489 0.0982 1.0000  
recroom | 0.2989 0.2097 0.0900 0.1348 0.1348 0.0303 0.0920 1.0000  
basement | 0.2284 0.0629 0.1019 0.1100 0.1100 -0.1339 0.0434 0.3724  
  gas | 0.0852 -0.0138 0.0363 0.0687 0.0687 0.0456 -0.0119 -0.0101  
  aircond | 0.4583 0.2851 0.1815 0.2029 0.2029 0.2386 0.1063 0.1366  
  garage | 0.3636 0.3526 0.1465 0.1748 0.1748 0.0162 0.1964 0.0495  
  desireloc | 0.3465 0.2463 0.1002 0.0774 0.0774 0.0409 0.1994 0.1613
```

Key Commands

Regress varlist
Correlate varlist
Spearman varlist

Format data in excel Save in *.csv format

Import csv data file in Stata9 See list of variables

Use menu or on line commands for statistical analysis

Look at help menu in STATA if you need further help on how to do things

Why can it be wrong to use raw time series data for regression? How can it be made right?

- A times series has a trend, cycle, season and random component.
- Usually time series data are non-stationary.
- Using non-stationary data results in spurious regression- may result in false statement.
- Dicky-Fuller and Augmented Fuller Tests for existence of unit roots.
- In PcGive: Descriptive Statistics/Unit root.
- For instance area and yield of banana are all non-stationary (clear from the unit root test).
- They are non-stationary even
 - In the first differences
 - In the log
- Only the first difference of logs was stationary

Unit Root Test of Banana in the first difference in logs

$$Y_t = \rho Y_{t-1} + e_t \quad H_0: \rho=1$$

prodcam_L1: ADF tests (T=39, Constant; 5%=-2.94 1%=-3.61) $H_A: \rho \leq 1$

D-lag	t-adf	beta Y_1	sigma	t-DY_lag	t-prob	AIC	F-prob
2	-3.181*	0.099137	0.04863	-1.807	0.0794	-5.950	
1	-6.235**	-0.26798	0.05013	2.123	0.0407	-5.912	0.0794
0	-6.060**	0.0077992	0.05246			-5.846	0.0266

prodch_L1: ADF tests (T=39, Constant; 5%=-2.94 1%=-3.61)

D-lag	t-adf	beta Y_1	sigma	t-DY_lag	t-prob	AIC	F-prob
2	-3.564*	0.33520	0.08494	0.2420	0.8102	-4.835	
1	-4.104**	0.35912	0.08382	0.7727	0.4447	-4.884	0.8102
0	-4.426**	0.42529	0.08336			-4.919	0.7282

Prodcam_L1: production in Cambodia in the first difference

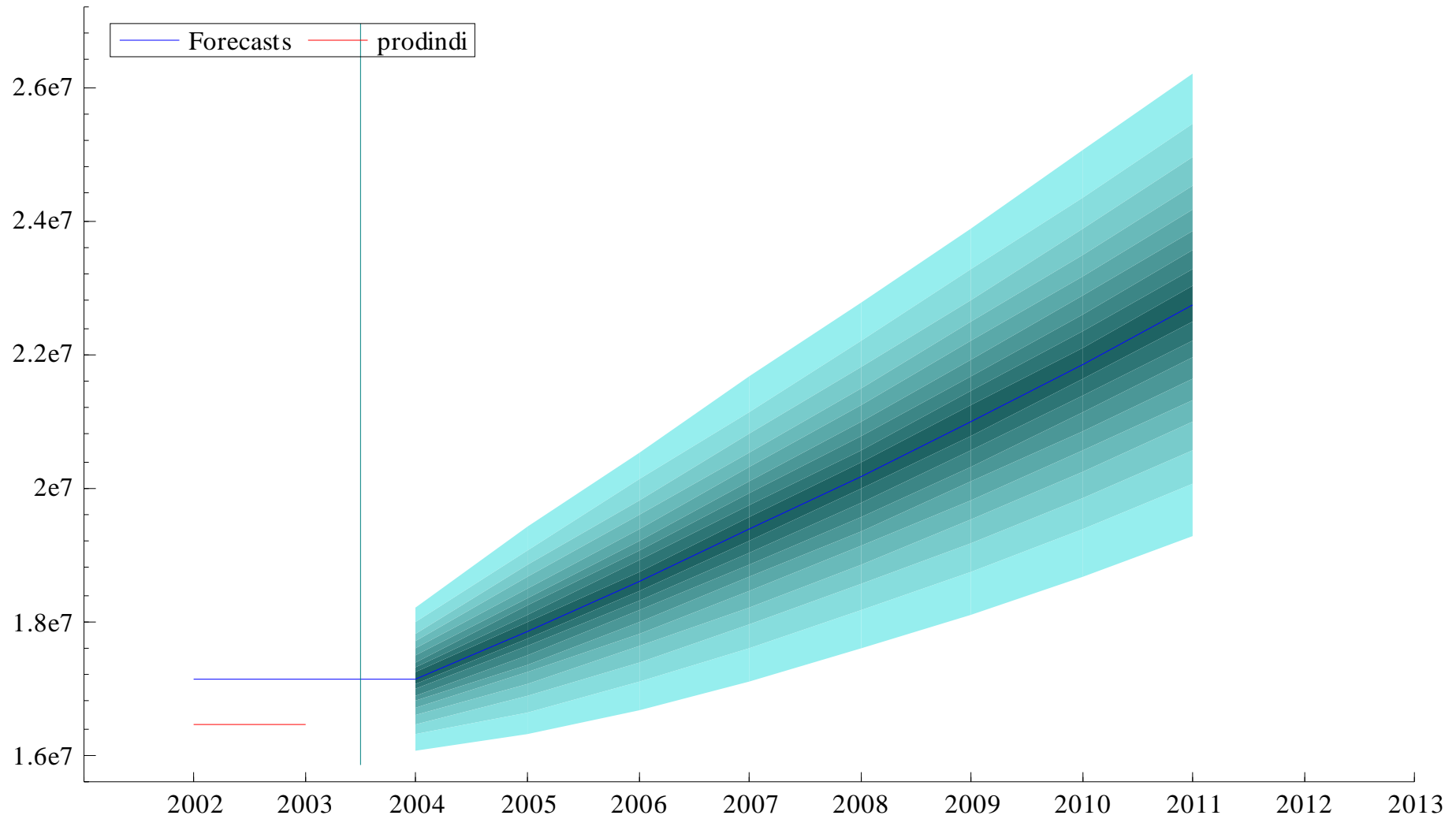
Prodch_L1: production in China in the first difference

Autoregressive Model of Banana Production in India

- Coefficient Std.Error t-value t-prob Part.R²
- prodindi_1 1.03581 0.02332 44.4 0.000 0.9806
- Constant 104960. 1.852e+005 0.567 0.574 0.0082
- sigma 654757 RSS 1.67195594e+013
- R² 0.980617 F(1,39) = 1973 [0.000]**
- log-likelihood -606.224 DW 1.46
- no. of observations 41 no. of parameters 2
- mean(prodindi) 6.96343e+006 var(prodindi) 2.10385e+013

- AR 1-2 test: F(2,37) = 2.0045 [0.1491]
- ARCH 1-1 test: F(1,37) = 1.1206 [0.2967]
- Normality test: Chi²(2) = 14.071 [0.0009]**
- hetero test: F(2,36) = 4.3042 [0.0211]*
- hetero-X test: F(2,36) = 4.3042 [0.0211]*
- RESET test: F(1,38) = 8.4096 [0.0062]**
- prodindi = + 1.036*prodindi_1 + 1.05e+005
- (SE) (0.0233) (1.85e+005)

AR(1) Forecasts of Production of Banana in India



Theoretical Probability Distributions and their Economic Applications

Bornouli and Binomial Distributions

Properties of a Discrete Random Variable

$$P(x_i) \geq 0$$

$$\sum_{i=1}^n P(x_i) = 1$$

$$E[X_i] = \sum_i X_i P(x_i)$$

$$\text{var}[X_i] = \sum_i (X_i - E(X_i))^2 P(x_i)$$

Sum of numbers appearing in a throw of two dice: Discrete Variat

1st

1	2	3	4	5	6
---	---	---	---	---	---

2nd

1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Number of Possible events: 36

Frequency, probability, expected value and variance and a game based upon a discrete random variable

X	Frequency (f)	Probability p(X)	Expected value: x.p	Variance	Even (win)	Odd (lose)
2	1	0.028	0.056	1.389	0.027778	
3	2	0.056	0.167	2.667		0.055556
4	3	0.083	0.333	3.000	0.083333	
5	4	0.111	0.556	2.222		0.111111
6	5	0.139	0.833	0.833	0.138889	
7	6	0.167	1.167	0.000		0.166667
8	5	0.139	1.111	1.111	0.138889	
9	4	0.111	1.000	4.000		0.111111
10	3	0.083	0.833	7.500	0.083333	
11	2	0.056	0.611	9.778		0.055556
12	1	0.028	0.333	8.333	0.027778	
Total	N= 36	1.000	7.000	40.833	0.500	0.500

$$\text{var}[X_i] = \sum_i (X_i - E(X_i))^2 P(x_i)$$

Bernouli Distribution

$$P_x(x = 0) = (1 - p) \quad P_x(x = 1) = p$$

Mean $\mu_x = E(x) = \sum_x xP(x) = 0(1 - p) + (1)p = p$

Variance: $\sigma_x^2 = E(x - \mu_x)^2 = (0 - p)^2(1 - p) + (1 - p)^2 p = p(1 - p)$

Application: a salesman's probability of selling is 0.4

$$\mu_x = E(x) = p = 0.4$$

$$\sigma_x^2 = p(1 - p) = 0.4 \times 0.6 = 0.24$$

Binomial Distribution

In an n number of trials, x cases succeed

$$P_x(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

In five attempts what is the probability that it succeeds five times

$$P_x(x=5) = \frac{5!}{5!(5-5)!} 0.4^5 (1-0.6)^{5-5} = 0.031$$

Poisson Distribution

$$P_x(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

It applies to rare event such as probability of strike by railway staff

Probability of 0 strike: $P_x(x = 0) = \frac{e^{-0.5} 0.5^0}{0!} = 0.606$

Probability of one strike: $P_x(x = 1) = \frac{e^{-0.5} 0.5^1}{1!} = 0.303$

Probability of 5 strikes: $P_x(x = 5) = \frac{e^{-0.5} 0.5^5}{5!} = 0.002$

Continuous Random variable: IQ scores in a certain Exam

Profits of firms

Age of individuals in a country

Income of individuals and households

Consumer's surplus for a given product

Distance travelled by cars in between two destinations

Amount of blood in ones body

Amount of water in a bath tub

$$f(x) = P[X = x_1] \geq 0$$

$$\int_{-\infty}^{\infty} f(x).dx = 1$$

$$\int_a^b f(x).dx = P(a \leq x \leq b)$$

Conditional Distributions

Mutually Exclusive and Exhaustive events vs. Dependent Events

$$f(x, y)$$

$$f(x) = \sum_y f(x, y)$$

$$f(y) = \sum_x f(x, y)$$

$$f(y/x) = \frac{f(x, y)}{f(x)} \quad f(x/y) = \frac{f(x, y)}{f(y)}$$

References

- Excel (2002), Microsoft Corporation; Tools/data analysis.
- Doornik J A and D.F. Hendry ((2003) PC-Give Volume I-III, GiveWin Timberlake Consultants Limited, London.
- Dougherty C. (2002) Introduction of Econometrics by, Second Edition, Oxford University Press.
- Hill, Griffiths and Judge (2001) Undergraduate Econometrics, Second Edition, John Willey and Sons, 2001.
- Koop G. (2000) Analysis of Economic Data, Wiley, UK.
- Newbold P, W.L. Carlson and B. Thorn (2003) Statistics for Business and Economics, 5th edition, Prentice Hall
- Studenmund A.H. (2001) Using Econometrics: Practical Guide, Pearson Education.
Data Source
Data-archive.ac.uk, www.statistics.gov.uk
[ESDS Access and Preservation datasets \(top 10\)](#)
[ESDS Government datasets](#)
[ESDS International datasets](#)
[ESDS Longitudinal datasets](#)
[ESDS Qualidata datasets \(top 10\)](#)