Multivariate Statistical Analysis

Types of Analysis

- In most cases data analysis is done to test the association between two or more variables (bivariate)
- In health research very frequently the aim is to establish the association between risk factor and disease or between therapy and outcome (patient oriented research)

12-Month Prevalence of DSM-IV Alcohol Dependence by Age

	18-29 yrs	30-34 yrs	45-64 yrs	65+ yrs	All
Women	6	3	1	.2	1
Men	13	6	3	1	6

From Grant et al., 1994

Life is Not Simple

- In reality most statistical analysis test complex relationships in which more than two variables are considered
- In health research either multiple risk factors are under consideration or many subject variables must be considered in ascertaining the outcome of a clinical trial
- The ultimate goal of these analysis is either explanation or prediction, i.e., *more than just establishing an association*

Examples of Multivariate Analyses

- Postulating progression of a specific type of lung cancer as a function of baseline stage of diagnostic, prior therapeutic regiments, age, sex among other clinical and /or demographic factors

- Evaluating the likelihood of domestic violence taking into account age of the individuals, whether or not they consume alcohol, ethnic background and level of education.

Multivariate Analyses

- The idea (*model) for the first example (progression of lung cancer)* in this case will be:
  - Progression of cancer = overall effect (irrespective of the effect of other factors) + effect of baseline stage of dx + effect of prior therapeutic regiments + effect of age & gender

Multivariate Analyses: How

- The model underlying the different effects is quantifiably measured for each individual
- Then the different effects are solved for using mathematical techniques (calculus and numerical methods) to quantify their effects and tests for their significance

Advantages of Multivariate Analyses

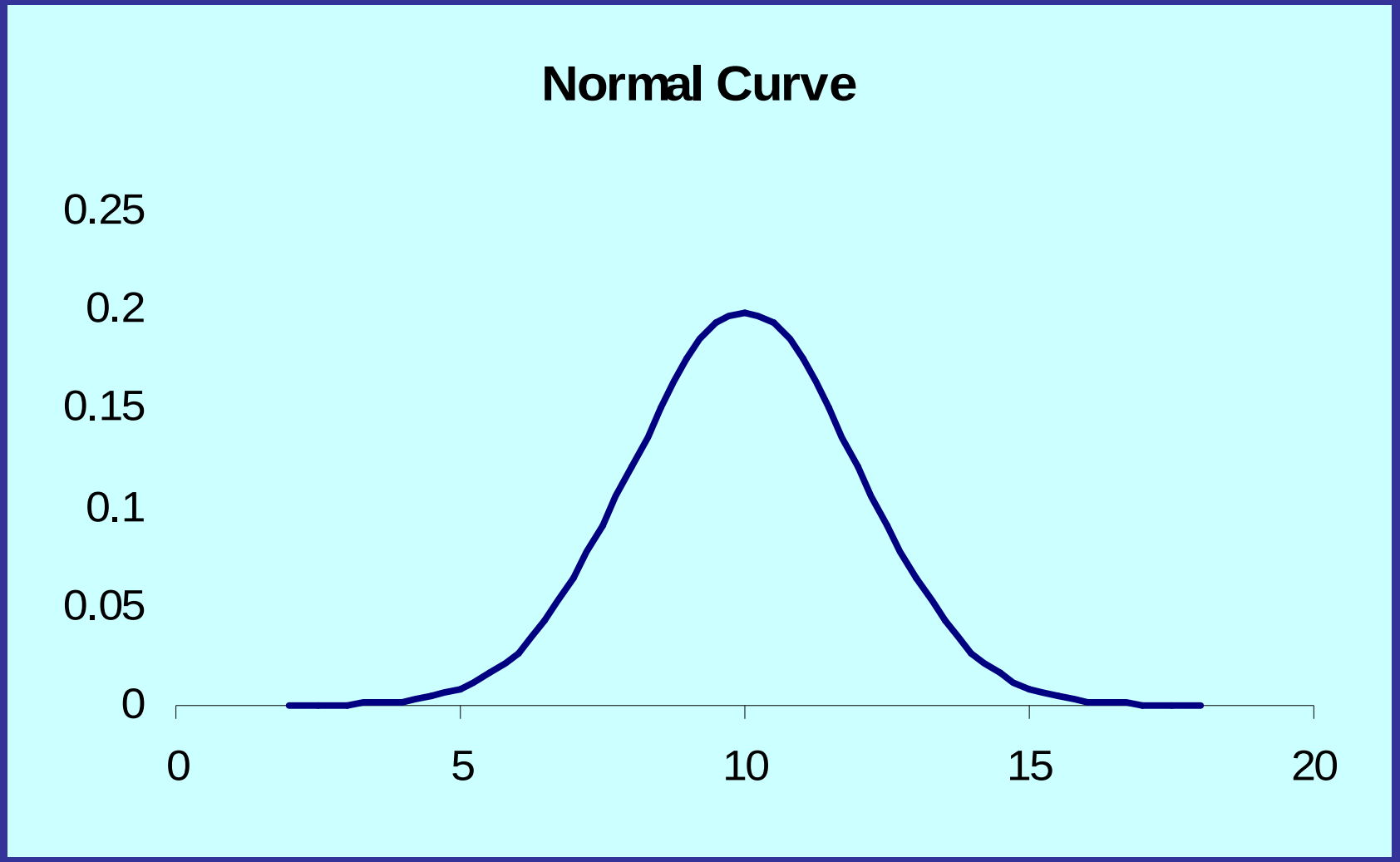
- It assures that the results are not biased and influenced by other factors that are not accounted for.

- Case in point: if difference in cancer progression due to different therapies is not a mere reflection of differences in prior therapies and or stage of diagnostic and or demographic factors.

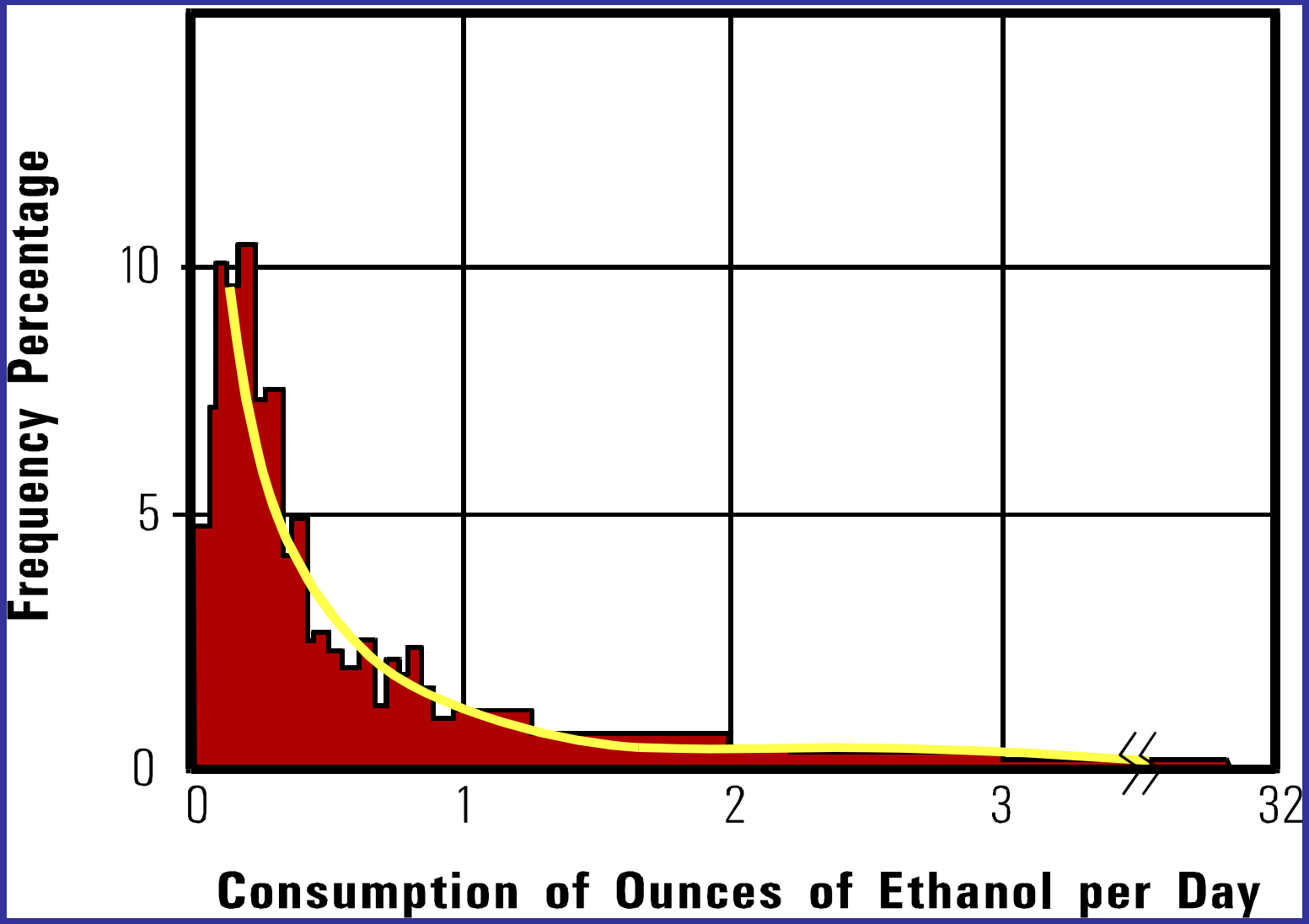
Which Technique?

- Statistical techniques are like tools: the task to be accomplished determines the selection of the tool
- The selection of a statistical technique is greatly based on the type of data under analysis (e.g., measurement level, type of distribution) and the reason (research question) for the analysis

Choice of Technique and Data Distribution



Choice of Technique and Data Distribution: Log Normal Curve



Choice of Technique: Levels (Types) of Data

- Nominal (Categorical) Measures: Are exhaustive and mutually exclusive (e.g., religion**).**
- Ordinal Measures: All of the above plus can be rank-ordered (e.g., social class).
- Interval Measures: All of the above plus equal differences between measurement points (temperature).
- Ratio Measures: All of the above plus a true zero point (weight).

The Link Between Measurement and Statistical Analysis

- A chi square can be used to test the association between two nominal level variables (e.g., gender and liver cirrhosis).
- A correlation can be used to assess the association between two interval or ratio level measures (e.g., height and weight).

What Type of Measure Should I Use?

- Whenever possible use interval or ratio measures
- However, many attributes of individuals (e.g., gender, religion) or health outcomes (cured/not cured) cannot be easily measured by interval or ratio data
- Thus, nominal or ordinal measures are also frequently used

What Level for Multivariate Analysis?

- In the past most techniques required at least interval data, that is, all measures in the analysis (including independent and dependent variables) should have at least equal differences between points on any part of the scale
- Nowadays many techniques (e.g., logistic regression) accept categorical, ordinal, interval or ratio level variables.
- These techniques are very popular in health research because of the difficulty of implementing interval or ratio measurements

Level of Measurement and Multivariate Statistical Technique

Independent Variable	Dependent Variable	Technique
Numerical	Numerical	Multiple Regression
Nominal or Numerical	Nominal	Logistic Regression
Nominal or Numerical	Numerical (censored)	Cox Regression
Nominal or Numerical	Numerical	ANOVA, MANOVA
Nominal or Numerical	Nominal (2 or more values)	Discriminant Analysis
Numerical	No Dependent Variable	Factor and Cluster Analysis

Assumptions in Linear Regression

1) The relationship under analysis is linear
2) The values of the independent variable are fixed (not random variable)
3) The values of the dependent variable are random
4) For each value of the independent variable the values of the dependent variable are normally distributed
5) The variance of the dependent variable is the same for all values of the independent variable

The Basic Regression Model

Simple Linear Regression

$$Y = a + bX$$

Y = Dependent variable

a = intercept

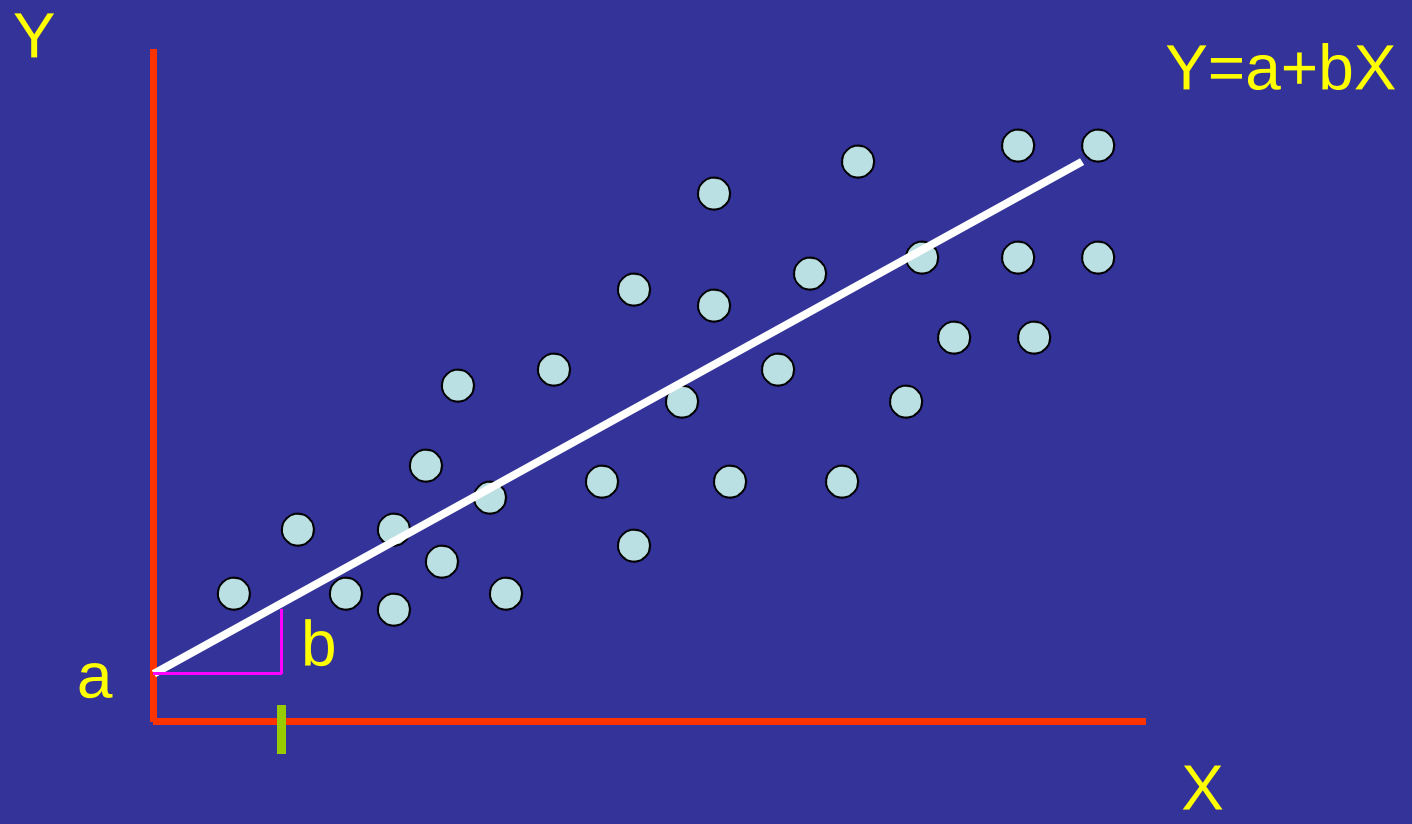
b = slope/regression coefficient (change in Y with a one unit change in X)

X = predictor value

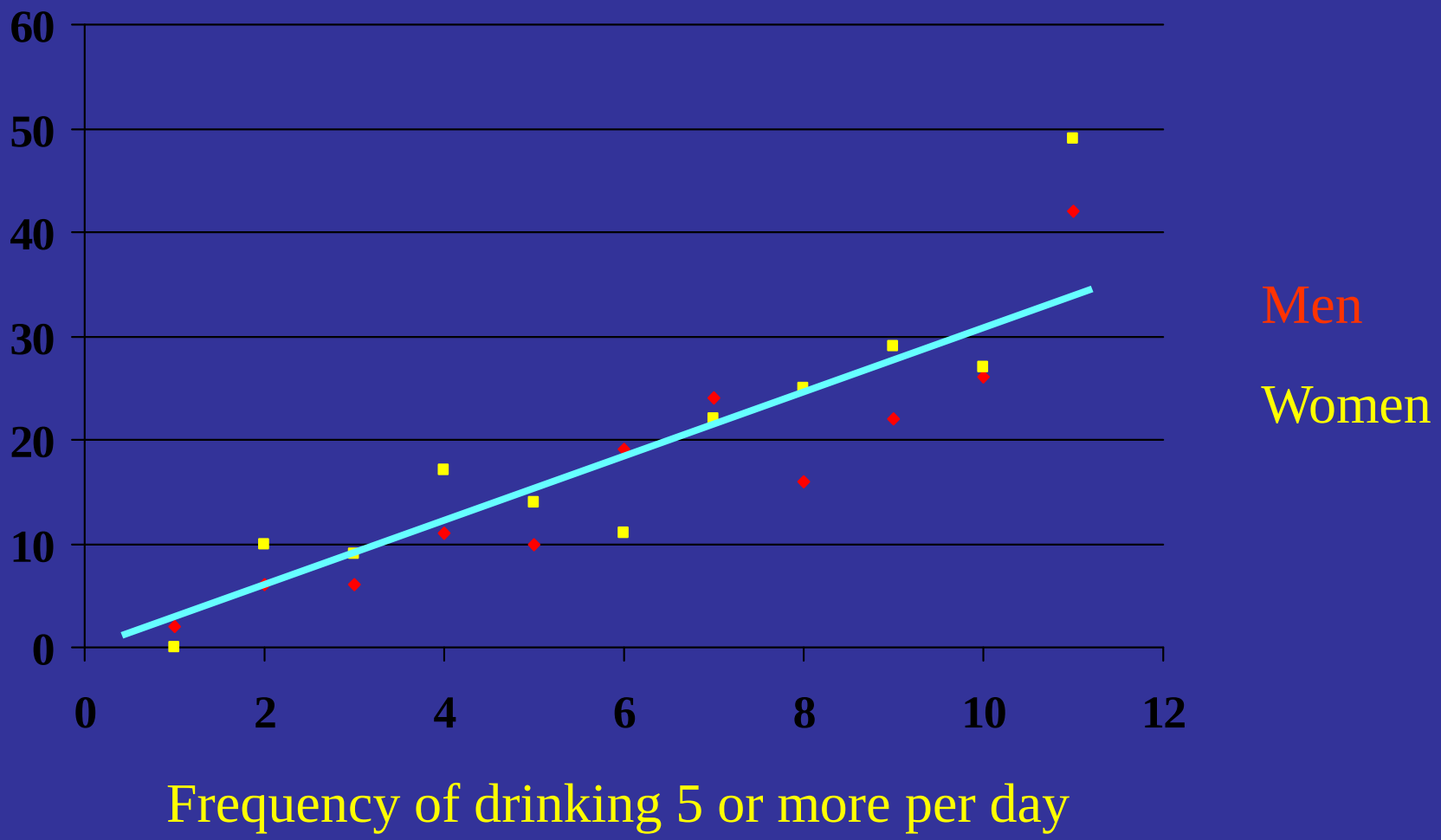
Multiple Linear Regression

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$$

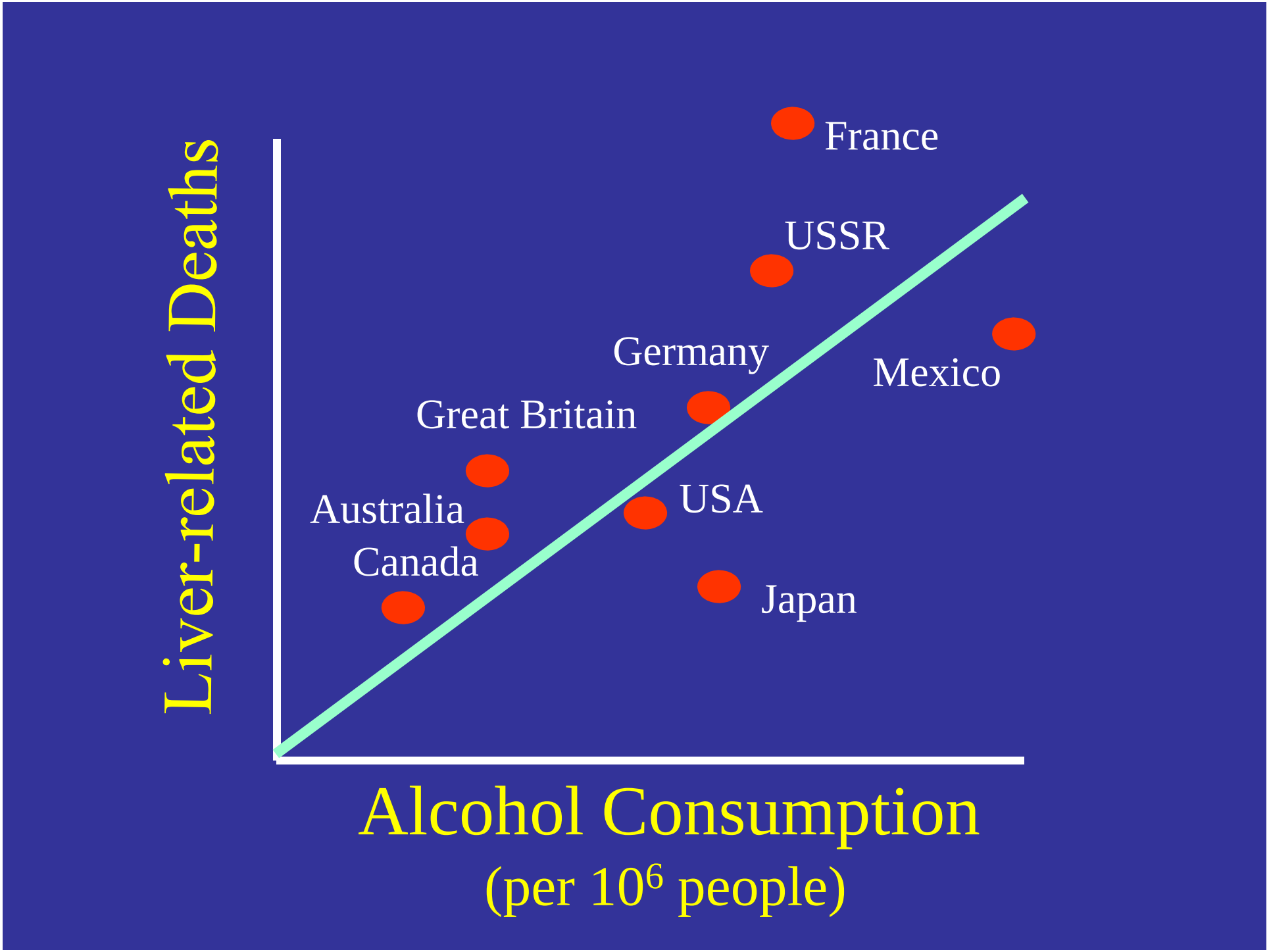
The Regression Line



Prevalence of Alcohol Dependence by Frequency of Drinking 5 or More/Day



From Caetano et al., 1997

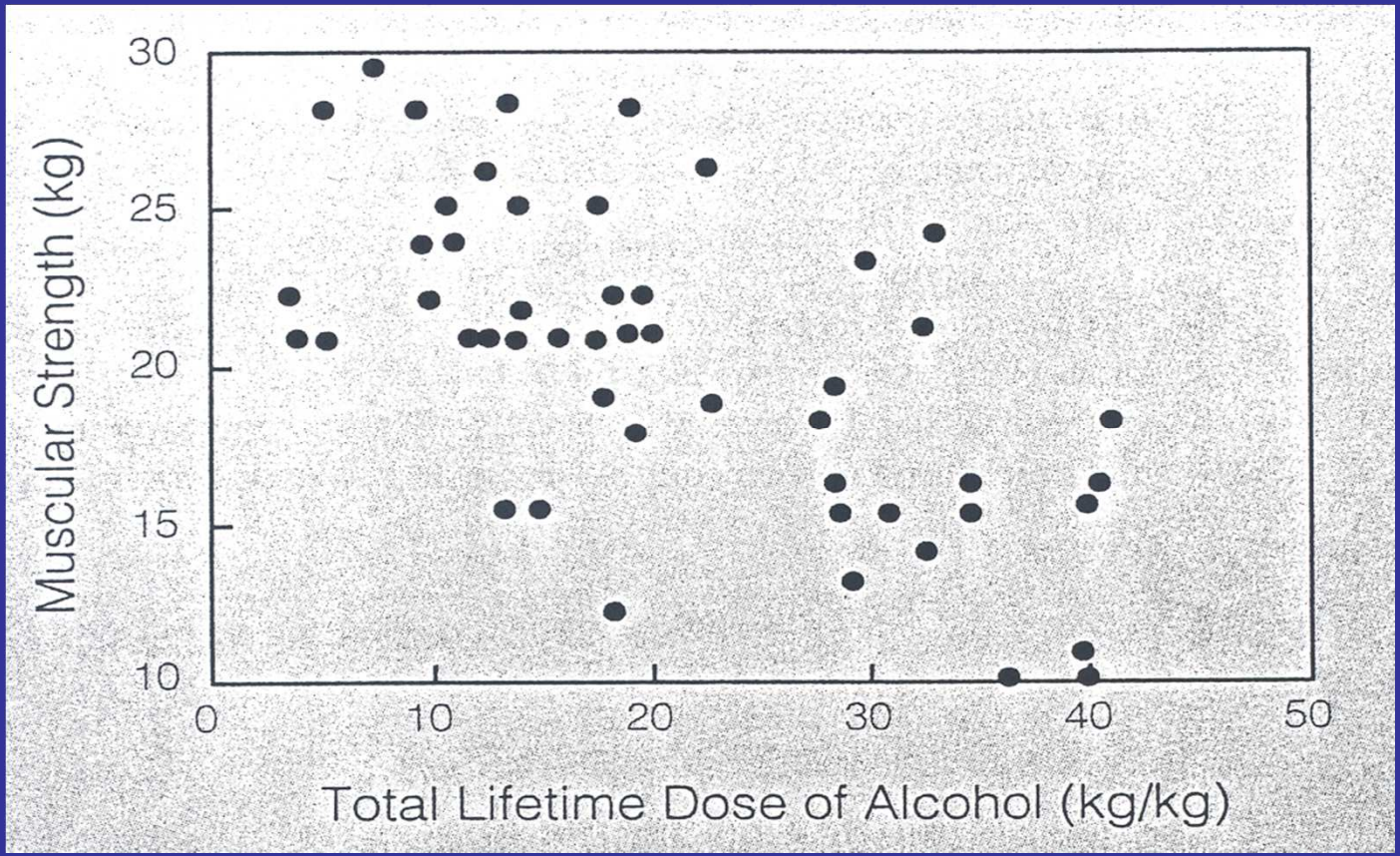

Liver-related Deaths
France
USSR
Germany
Mexico
Great Britain
Australia
USA
Canada
Japan
Alcohol Consumption
(per 10^6 people)

Measurements of Total Lifetime Dose of Alcohol and Muscular Strength for 50 Alcoholic Men

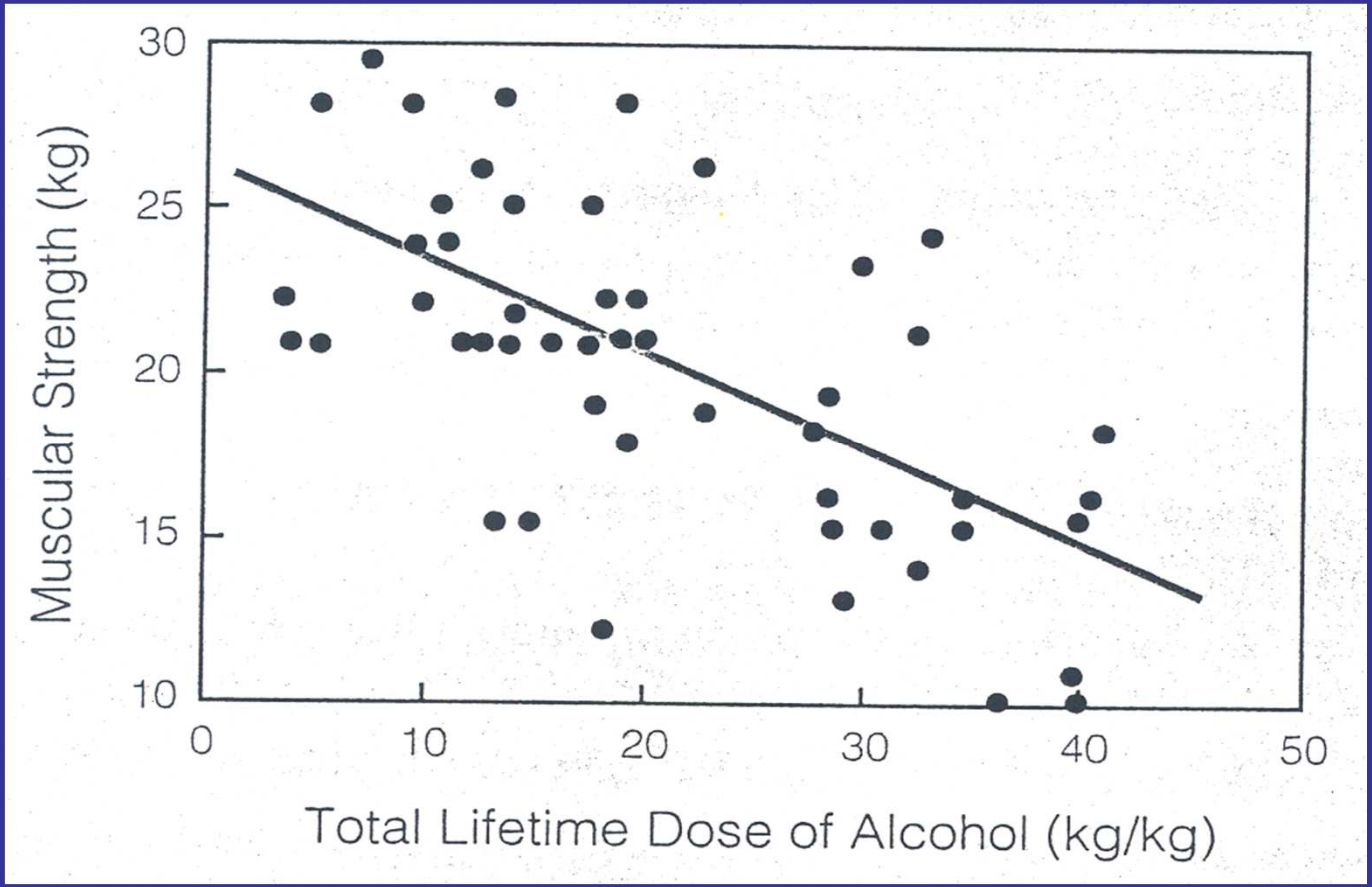
Total Lifetime Dose of Alcohol (kg/kg of body weight)	Muscular Strength (kg)	Total Lifetime Dose of Alcohol (kg/kg of body weight)	Muscular Strength (kg)
36.2	10.0	13.7	20.9
39.7	10.0	15.7	20.9
39.5	10.8	17.4	20.9
18.2	12.2	18.9	21.1
29.2	13.1	20.0	21.1
32.5	14.0	32.3	21.2
13.2	15.5	3.5	22.3
14.8	15.5	9.8	22.1
28.6	15.2	14.0	21.8
30.8	15.2	18.3	22.2
34.5	15.2	19.7	22.2
39.7	15.5	29.8	23.3
28.3	16.2	9.7	23.9
34.5	16.2	11.1	24.0
40.3	16.2	32.9	24.1
19.1	17.9	10.8	25.1
27.7	18.2	14.0	25.1
40.8	18.2	17.5	25.1
17.7	19.1	12.6	26.2
22.8	18.8	22.6	26.3
28.3	19.3	5.2	28.2
4.0	20.9	9.4	28.2
5.2	20.9	13.5	28.4
11.7	20.9	19.1	28.2
12.5	20.9	7.4	29.5

Urbano Marquez et al., 1989

Scatterplot of the Total Lifetime Dose of Alcohol vs. Muscular Strength for the Observations



The Least Squares Regression Line Added to the Scatterplot



Regression Equation: $Y = 26.4 - .296X$

Replicating Life's Complex Relationships in Data Analysis: Developing "The Model"

- A set of theoretically and or evidence based associations between variables is a "model"
- Data analysis tests the extent to which the proposed theoretical associations are observed in the data

Model Development

- Existing knowledge in the health field indicates that most, if not all disease-related outcomes, have multiple and varied risks
- This includes gene-environment interactions
- Thus, in health research the associations being tested usually involve a series of independent variables (“causes”) and a dependent variable (disease)

Model Development

- Model development is based on previous knowledge (previous research evidence about associations) and new and hypothetical ones.
- It is mostly about deciding which independent variables (risk factors) should be in the multivariate analysis.

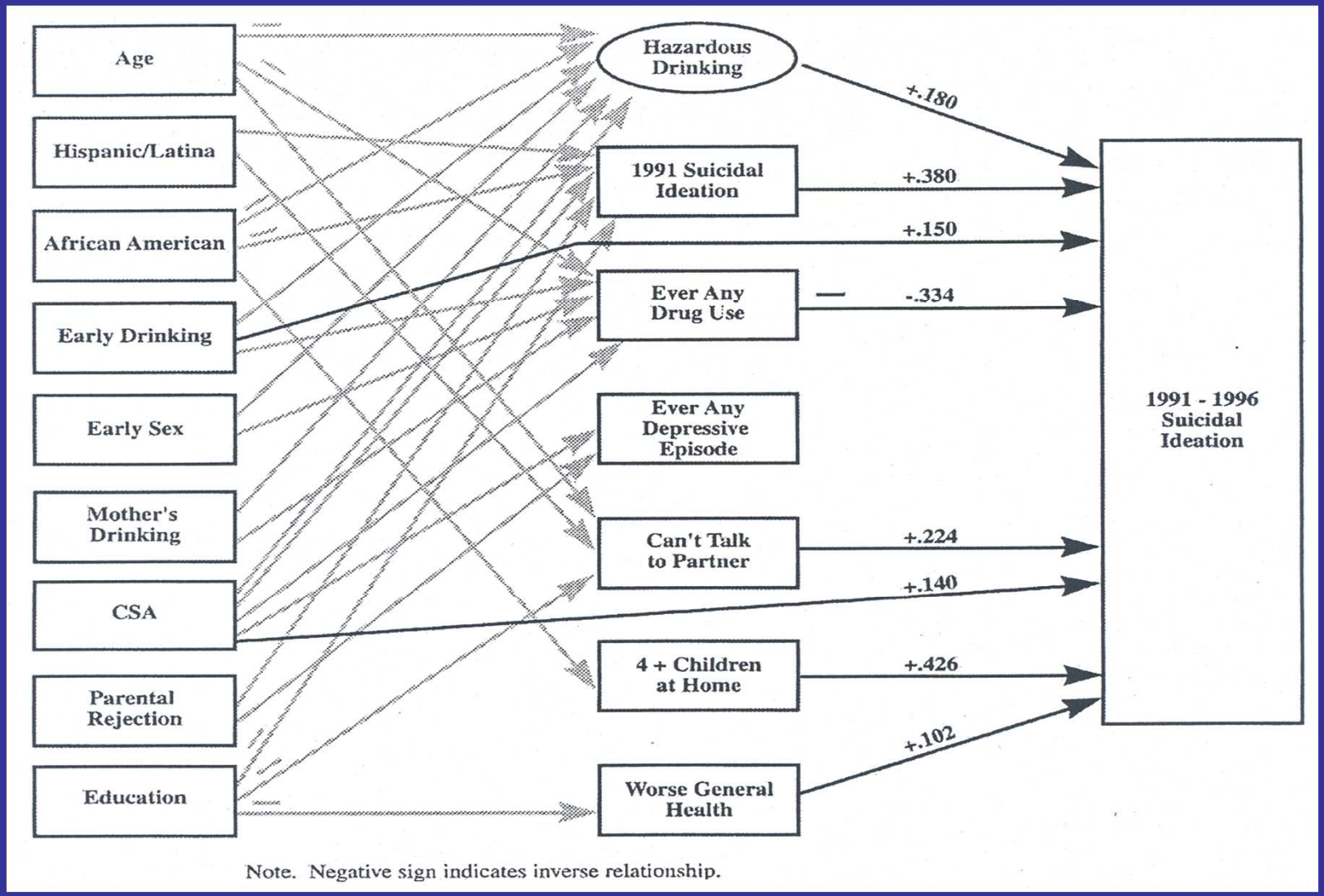
Example of Multiple Risk Factors

A SUMMARY OF RISK FACTORS IN THE CAUSATION OF CANCER

- Smoking
- Dietary Factors
- Obesity
- Exercise
- Occupation
- Genetic Susceptibility
- Infectious Agents
- Reproductive Factors
- Socioeconomic Status
- Environmental Pollution
- Ultraviolet Light
- Radiation
- Prescription Drugs
- Electromagnetic Fields

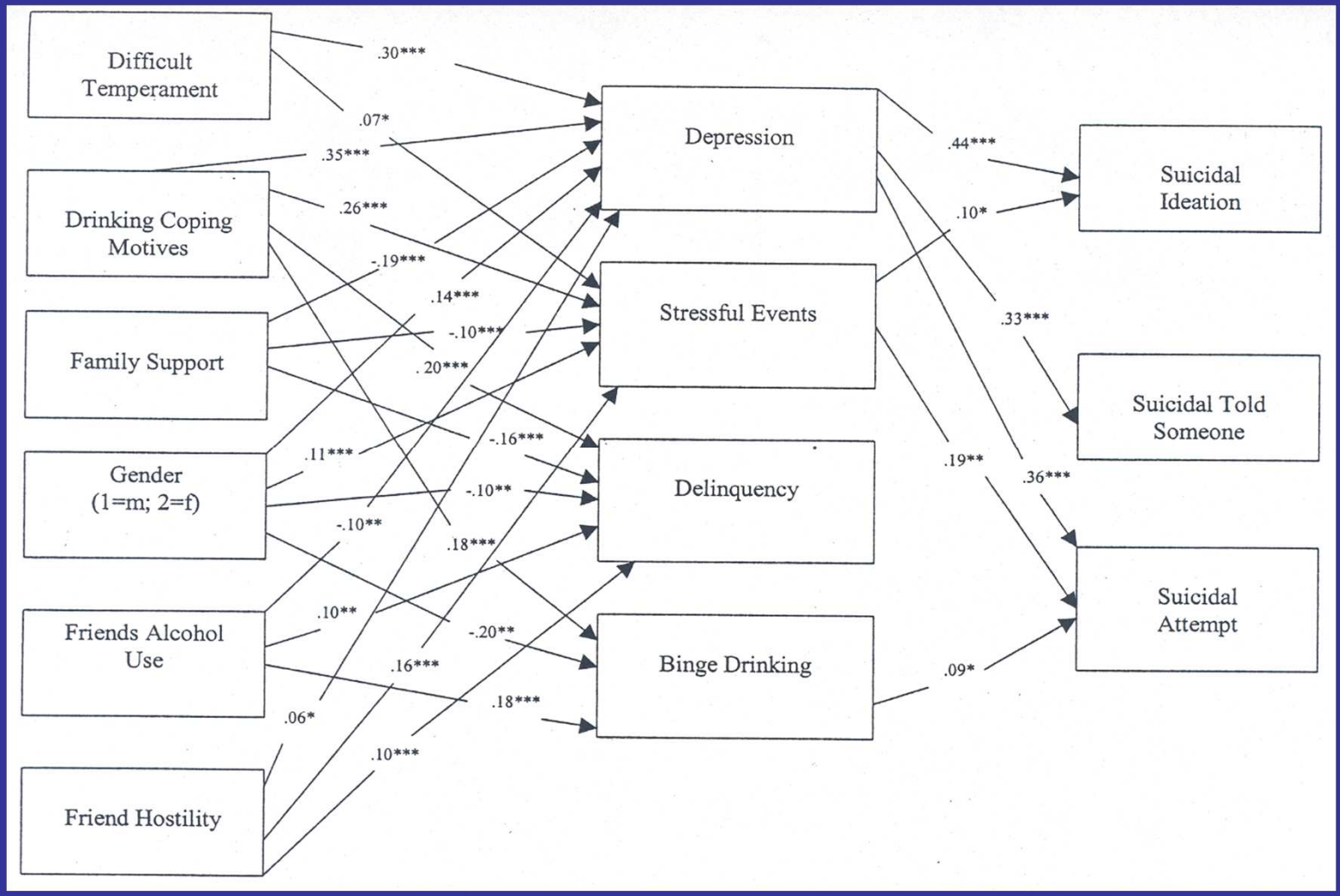
REV

Structural Equation Model of 1991 to 1996 Suicidal Ideation (Women Drinkers only)



From Wilsnack et al., 2004

Path Analytic Model of Risk Factors that Predict Suicidal Behaviors



From Windle, 2004

Points to Remember

- Most data analysis tries to answer complex questions involving more than 2 variables
- These questions can only be adequately addressed by multivariate statistical techniques.
- There are a variety of multivariate techniques all of which are based on assumptions about the nature of the data and the type of association under analysis.

Points to Remember

- Multivariate statistical techniques test theoretical models (research question) about associations against observed data.
- These theoretical models are based on previous knowledge and on new hypotheses about plausible associations between variables.

Bibliography

- R. D. De Veaux, P.F. Velleman “Intro do Stats”, Pearson Education Inc., Boston, 2004.
- B. Dawson, R.G. Trapp “Basic and Clinical Biostatistics”, 3rd edition, McGraw-Hill, New York, 2001.
- S. Glantz “Primer of Biostatistics”, 5th Edition, McGraw-Hill, New York, 2002.