Assumptions in Regression Analysis

Sixth session

The Assumptions

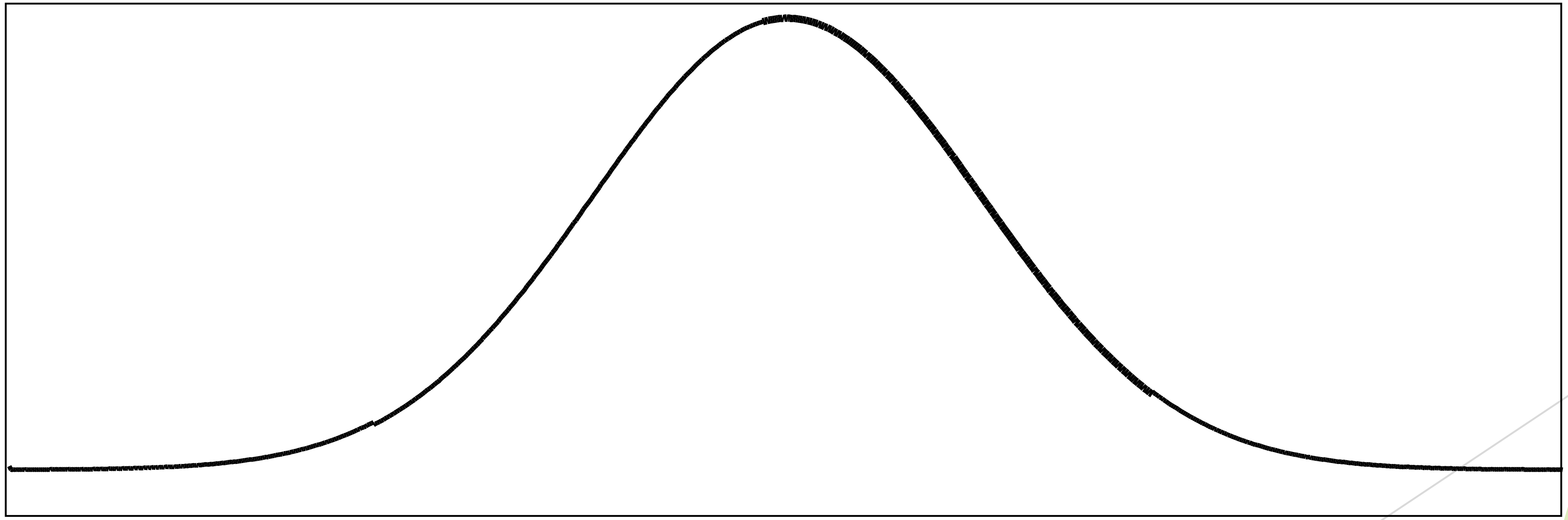
1. The distribution of residuals is normal (at each value of the dependent variable).
2. The variance of the residuals for every set of values for the independent variable is equal.
3. At every value of the dependent variable the expected (mean) value of the residuals is zero
   - No non-linear relationships
4. The expected correlation between residuals, for any two cases, is 0.
   - The independence assumption (lack of autocorrelation)

5. No independent variables are a perfect linear function of other independent variables (no perfect multicollinearity)

Assumption 1: The Distribution of Residuals is Normal at Every Value of the Dependent Variable

Look at Normal Distributions

- A normal distribution
  - symmetrical, bell-shaped (so they say)



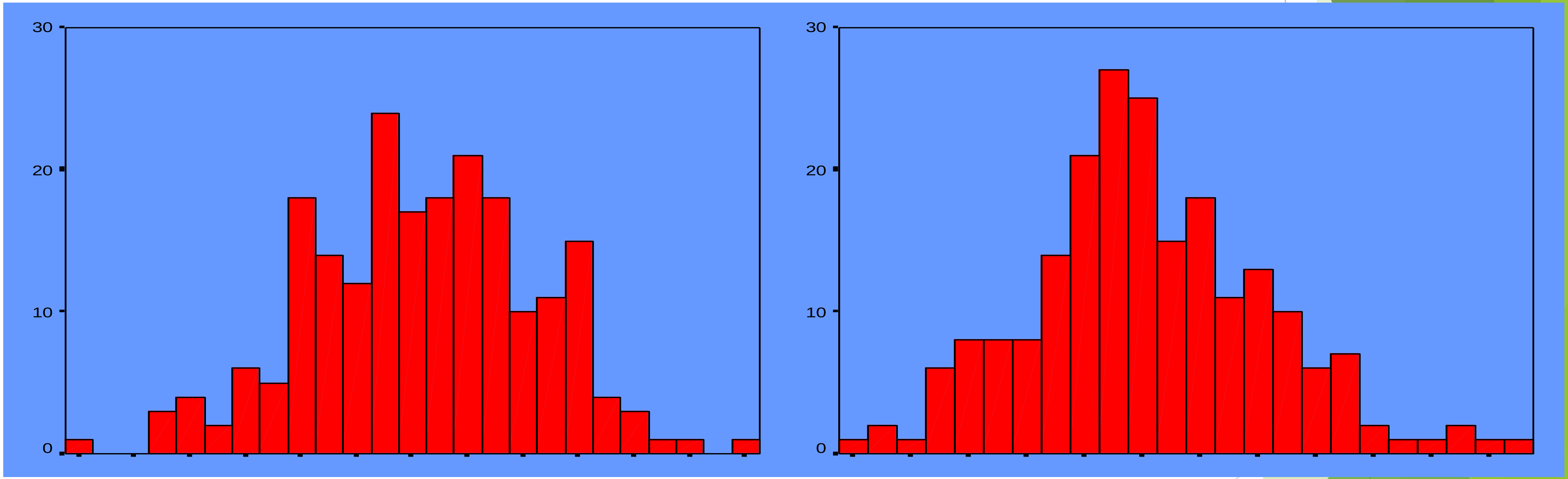
What can go wrong?

- Skew
  - non-symmetricality
  - one tail longer than the other
- Kurtosis
  - too flat or too peaked
  - kurtosed
- Outliers
  - Individual cases which are far from the distribution

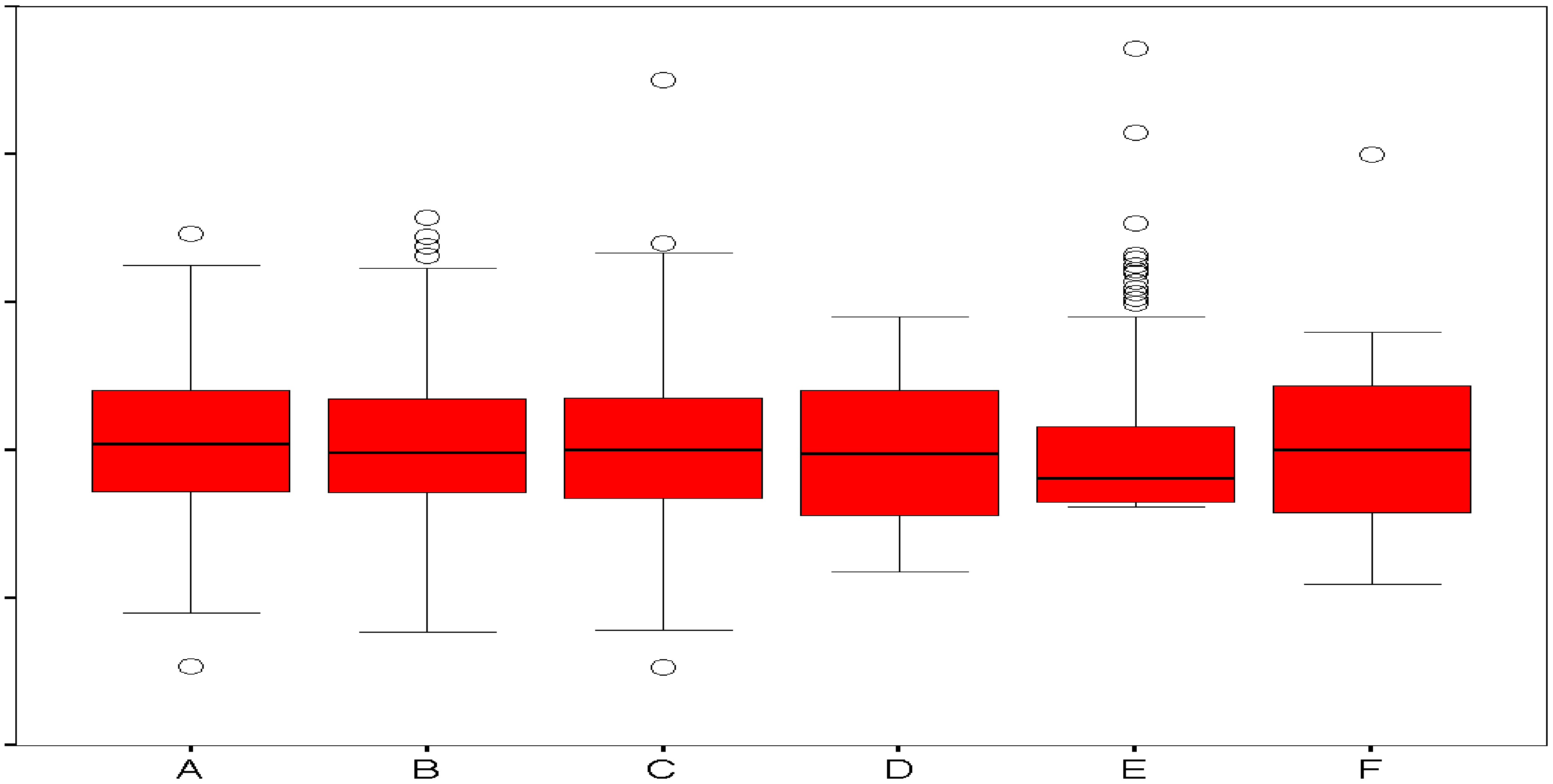
Examining Univariate Distributions

- Histograms
- Boxplots
- P-P Plots
- Kolmogorov-Smirnov Test
- Shapiro-Wilks test

Histograms

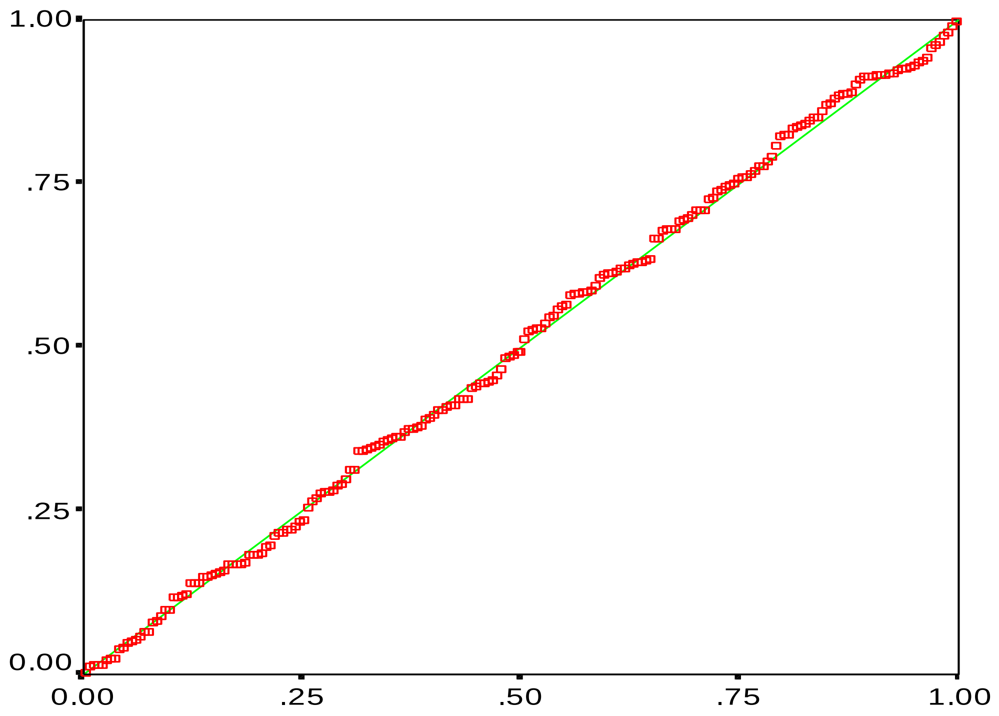


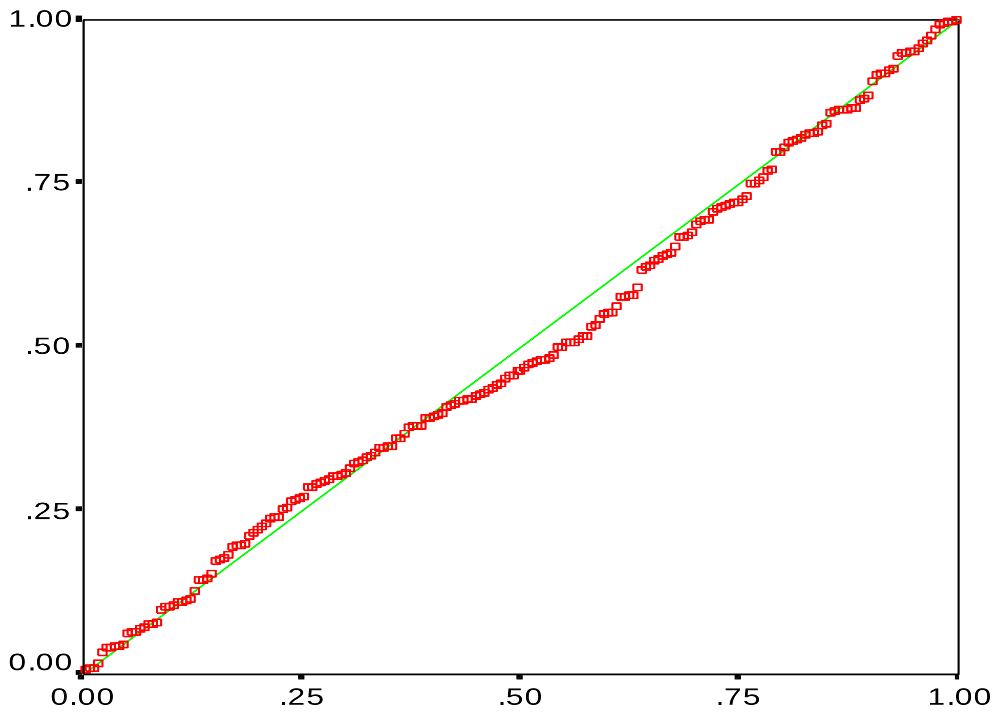
Boxplots



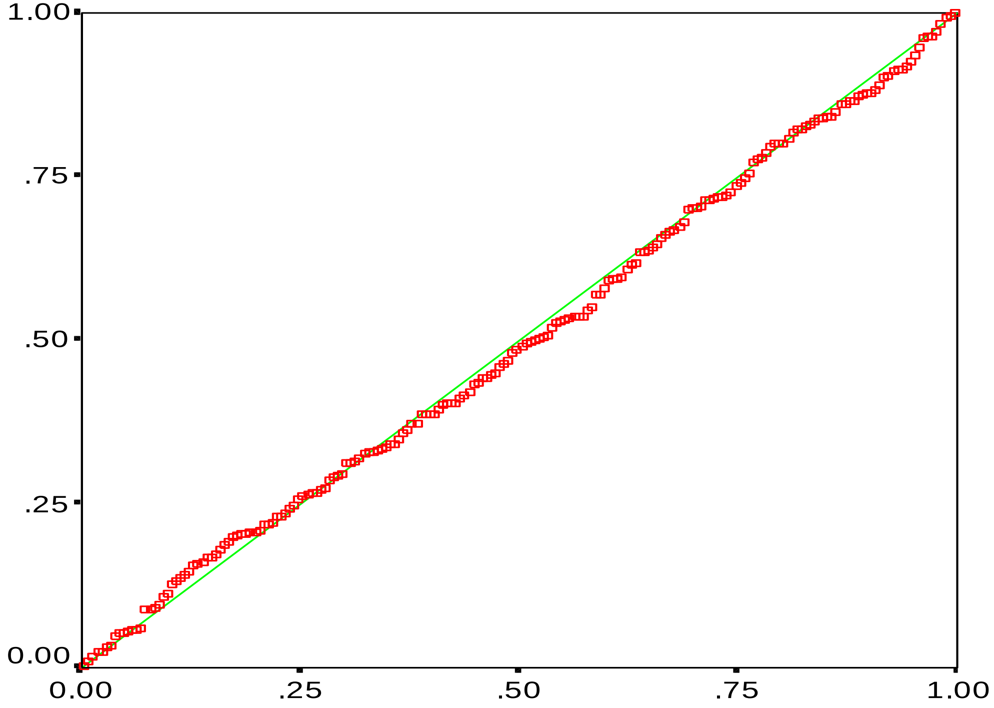
P-P Plots

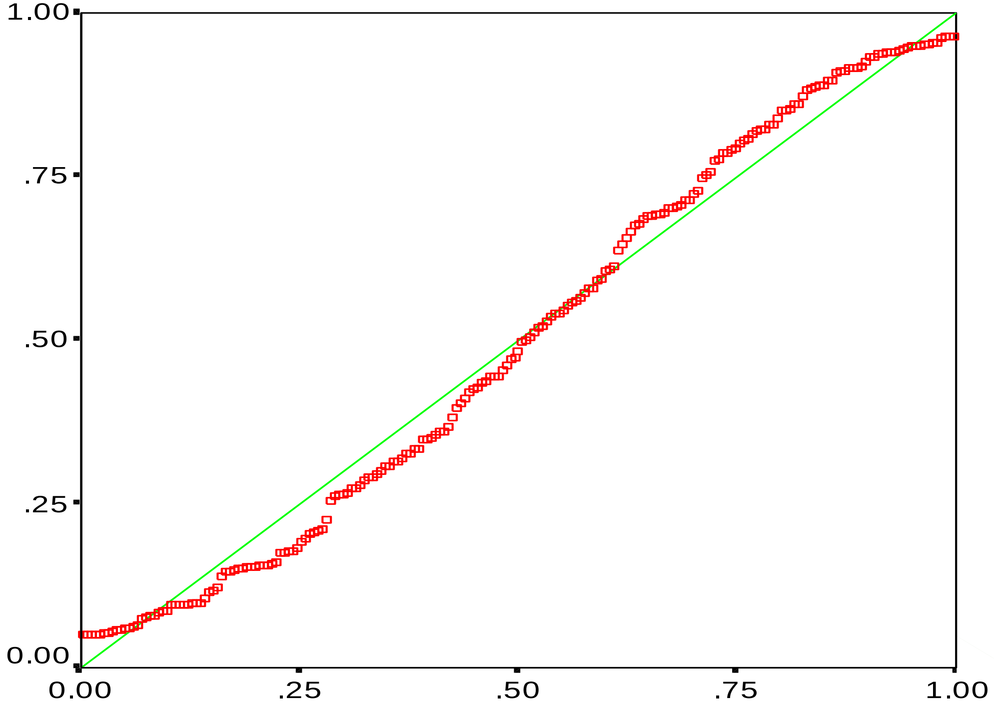
- A & B



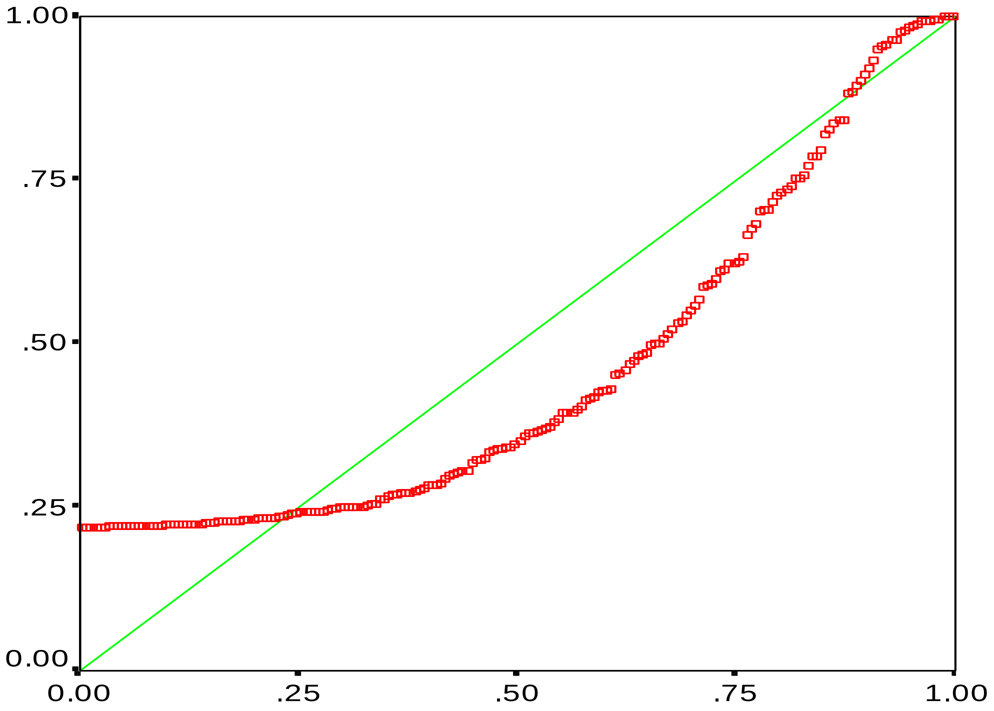


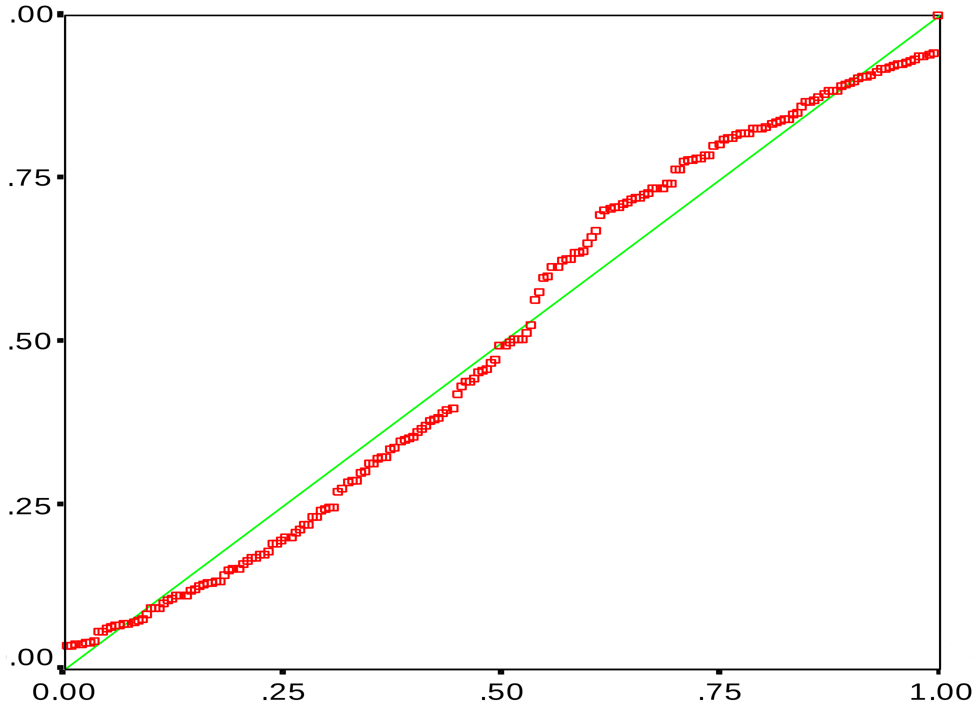
- C & D





- E & F

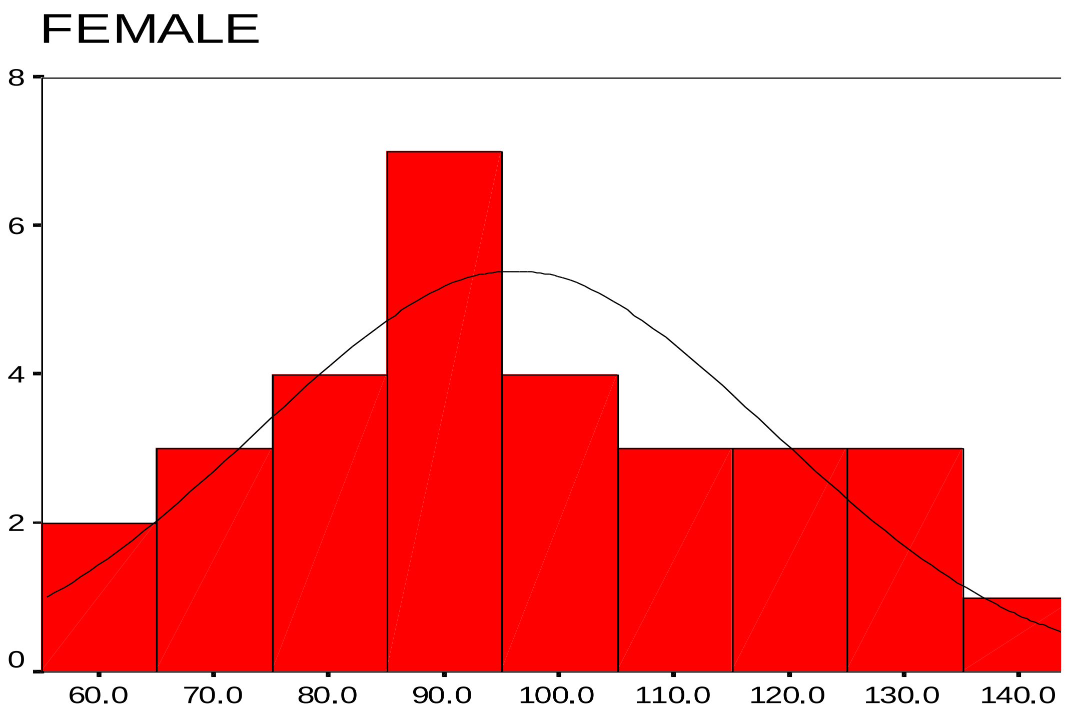


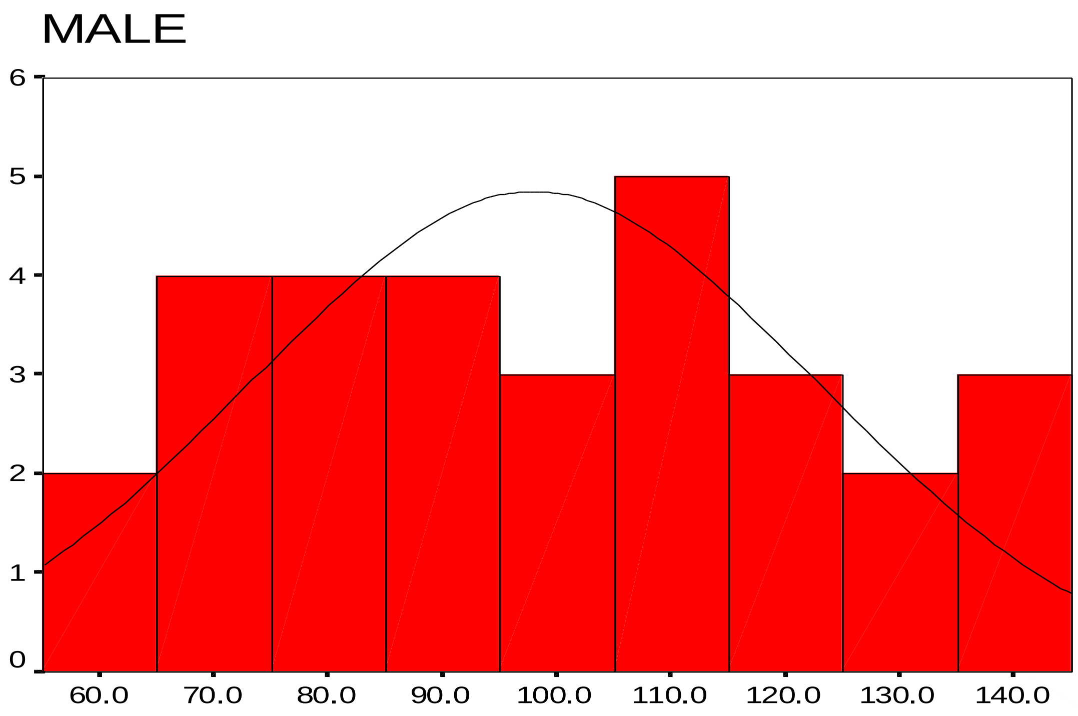


Bivariate Normality

- We didn't just say "residuals normally distributed"
- We said "at every value of the dependent variables"
- Two variables can be normally distributed – univariate,
  - but not bivariate

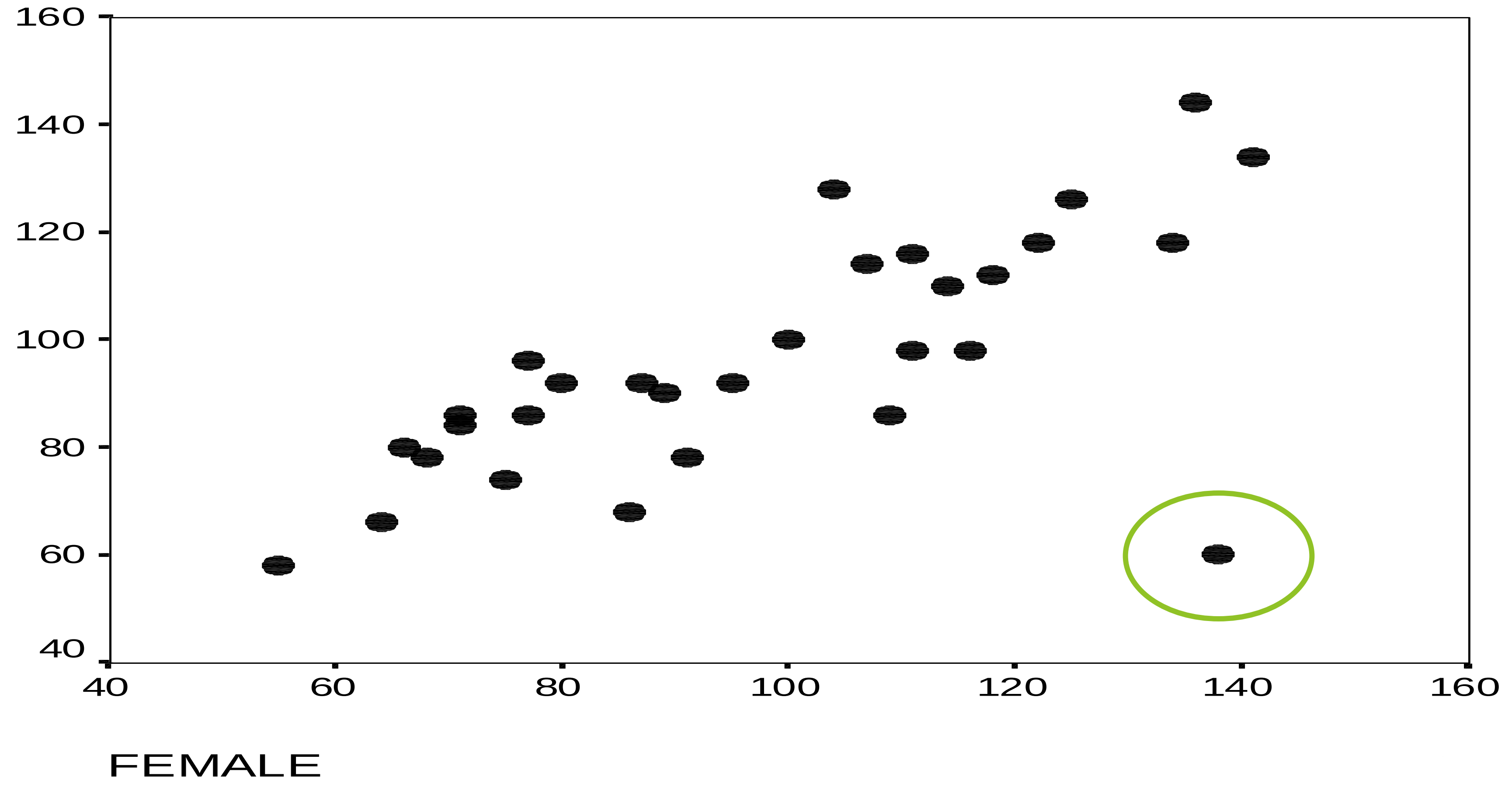
- Couple's IQs
  - male and female





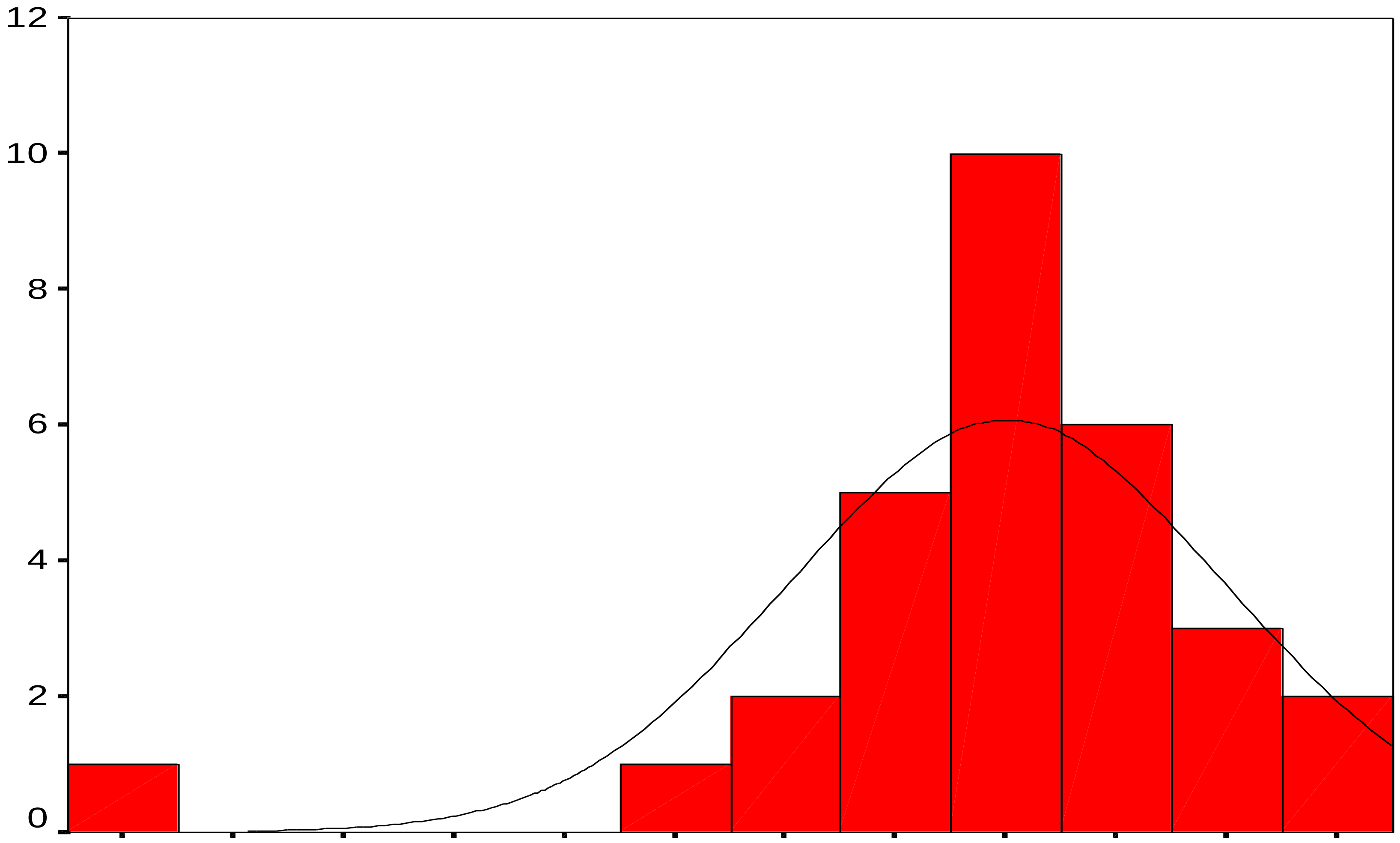
–Seem reasonably normal

- But wait!!



- When we look at bivariate normality
  - not normal – there is an outlier
- So plot X against Y
- OK for bivariate
  - but – may be a multivariate outlier
  - Need to draw graph in 3+ dimensions
  - can’t draw a graph in 3 dimensions
- But we can look at the residuals instead …

- IQ histogram of residuals



What to do about Non-Normality

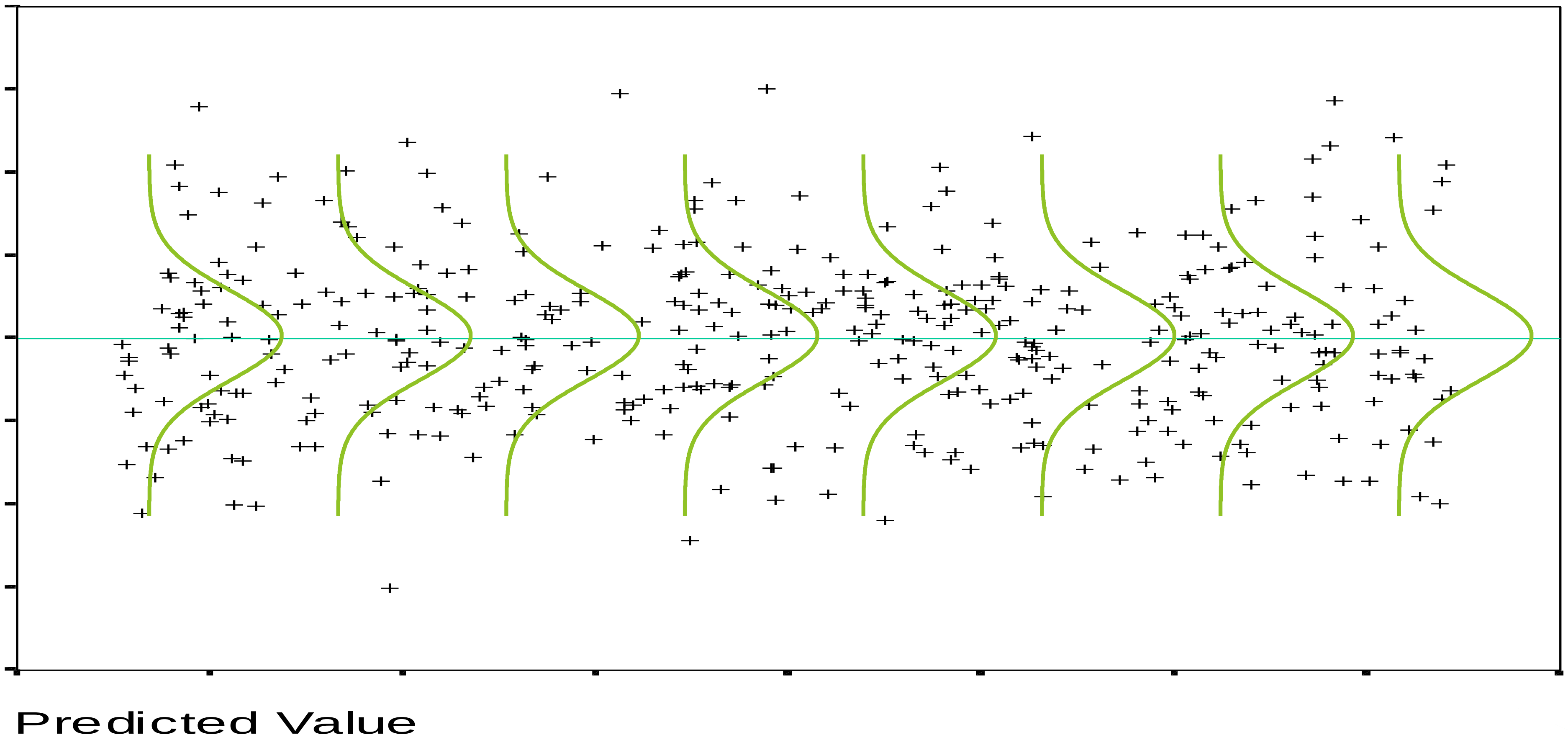
- Skew and Kurtosis
  - Skew – much easier to deal with
  - Kurtosis – less serious anyway
- Transform data
  - removes skew
  - positive skew – log transform
  - negative skew - square

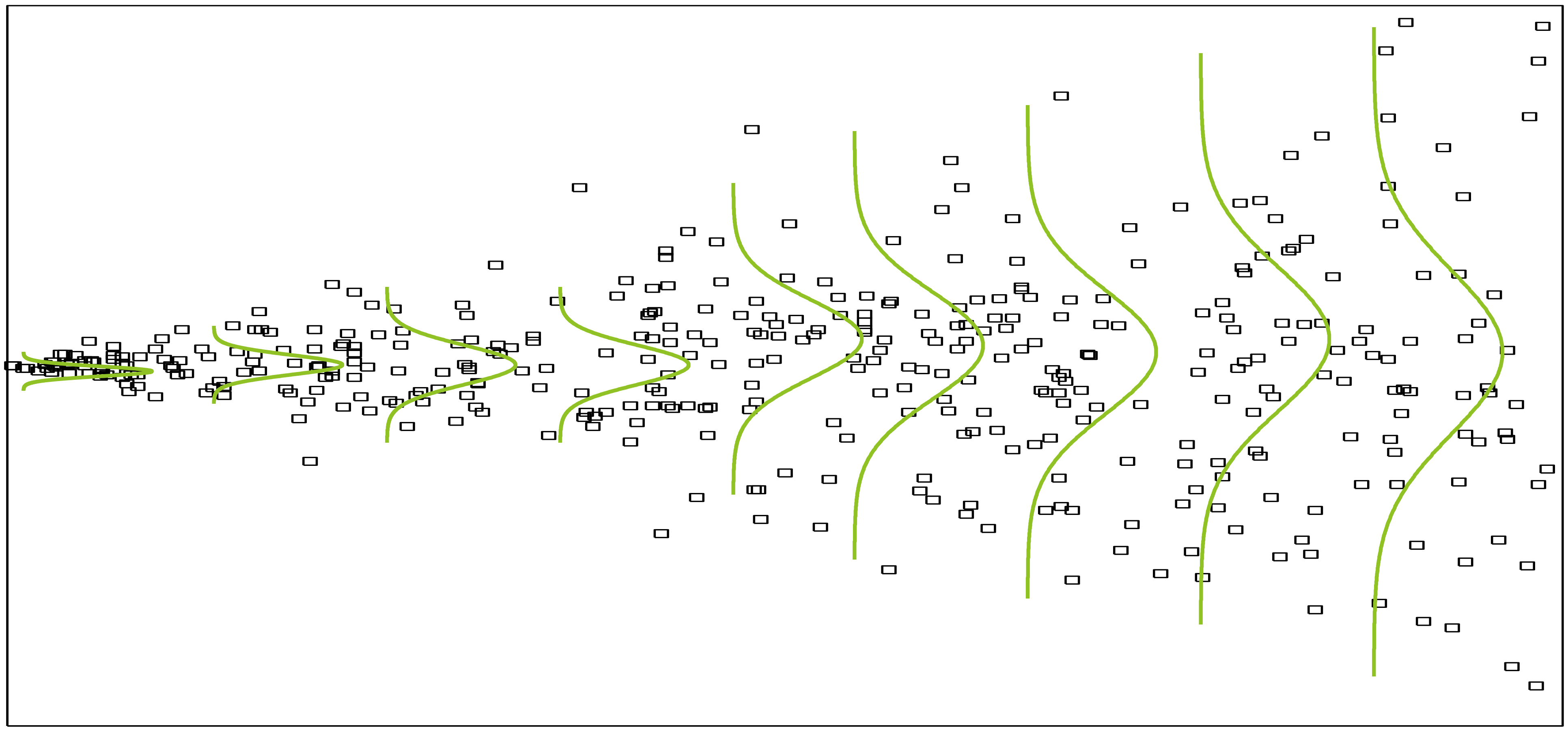
Transformation

- May need to transform IV and/or DV
  - More often DV
    - time, income, symptoms (e.g. depression) all positively skewed
  - can cause non-linear effects (more later) if only one is transformed
  - alters interpretation of unstandardised parameter
  - May alter meaning of variable
  - May add / remove non-linear and moderator effects

- Change measures
  - increase sensitivity at ranges
    - avoiding floor and ceiling effects
- Outliers
  - Can be tricky
  - Why did the outlier occur?
    - Error? Delete them.
    - Weird person? Probably delete them
    - Normal person? Tricky.

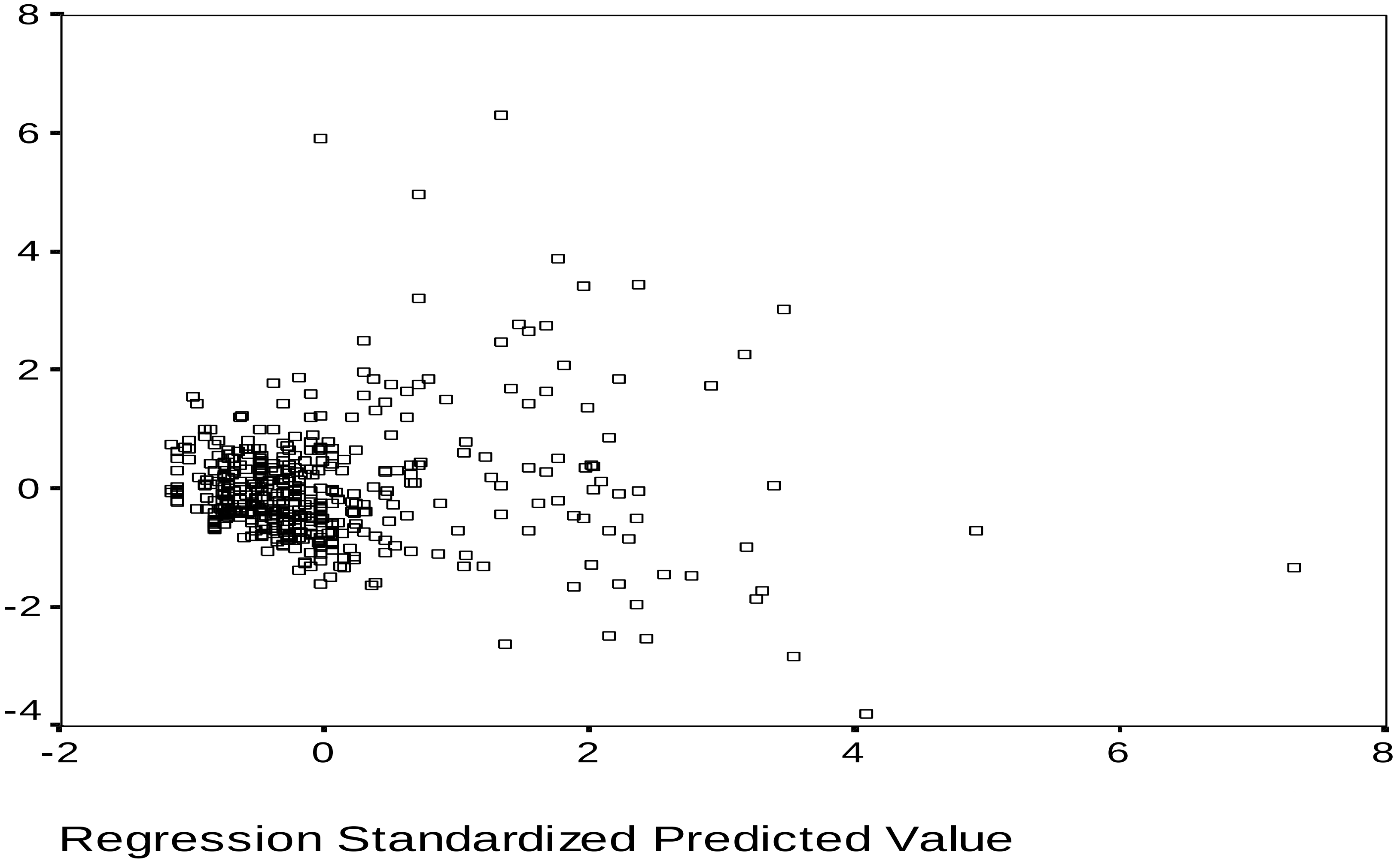
Assumption 2: The variance of the residuals for every set of values for the independent variable is equal.


Predicted Value



Predicted Value

Plot of Pred and Res

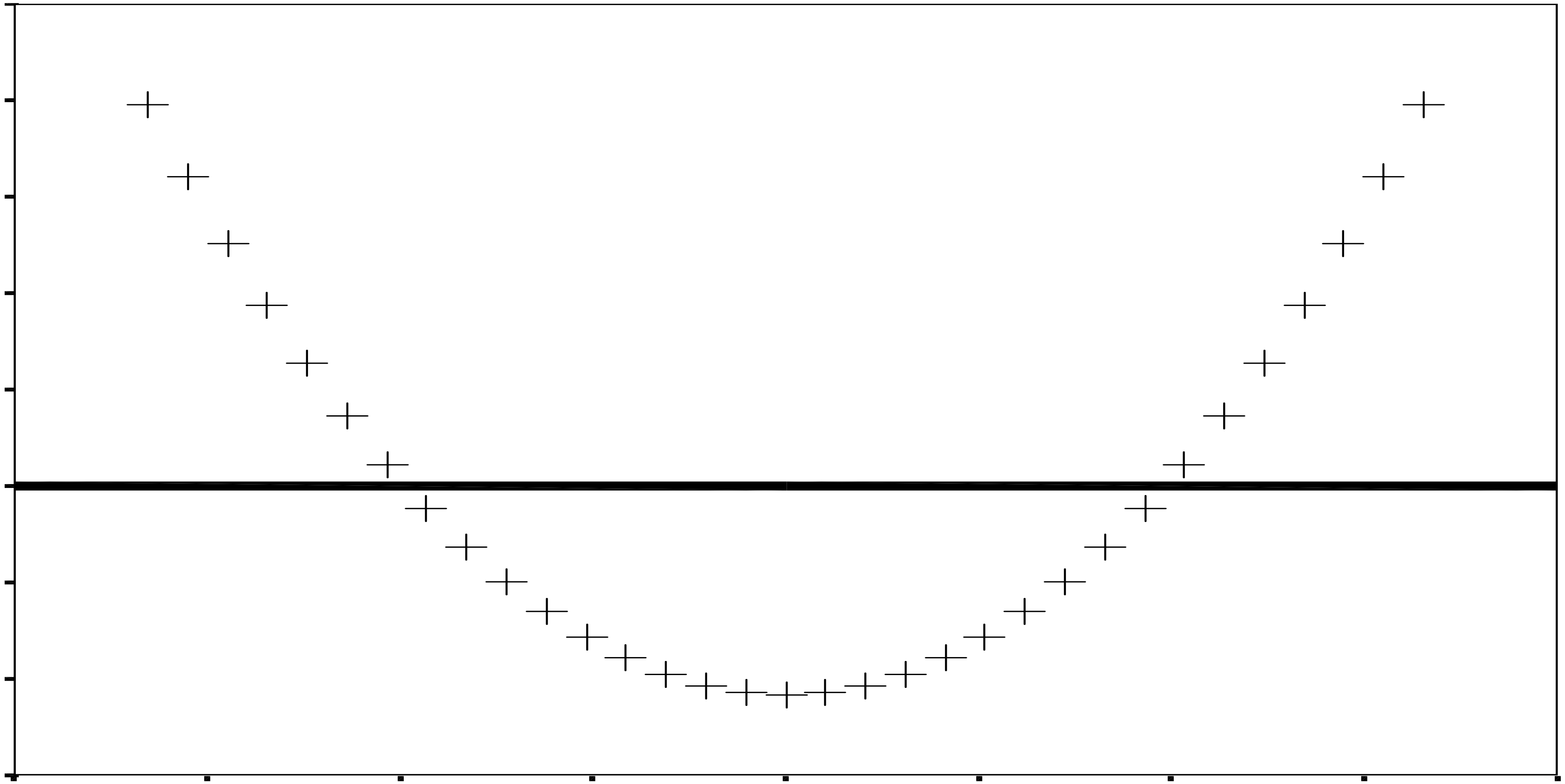


Assumption 3: At every value of the dependent variable the expected (mean) value of the residuals is zero

Linearity

- Relationships between variables should be linear
  - best represented by a straight line

Residual plot



Assumption 4: The expected correlation between residuals, for any two cases, is 0.

Independence Assumption

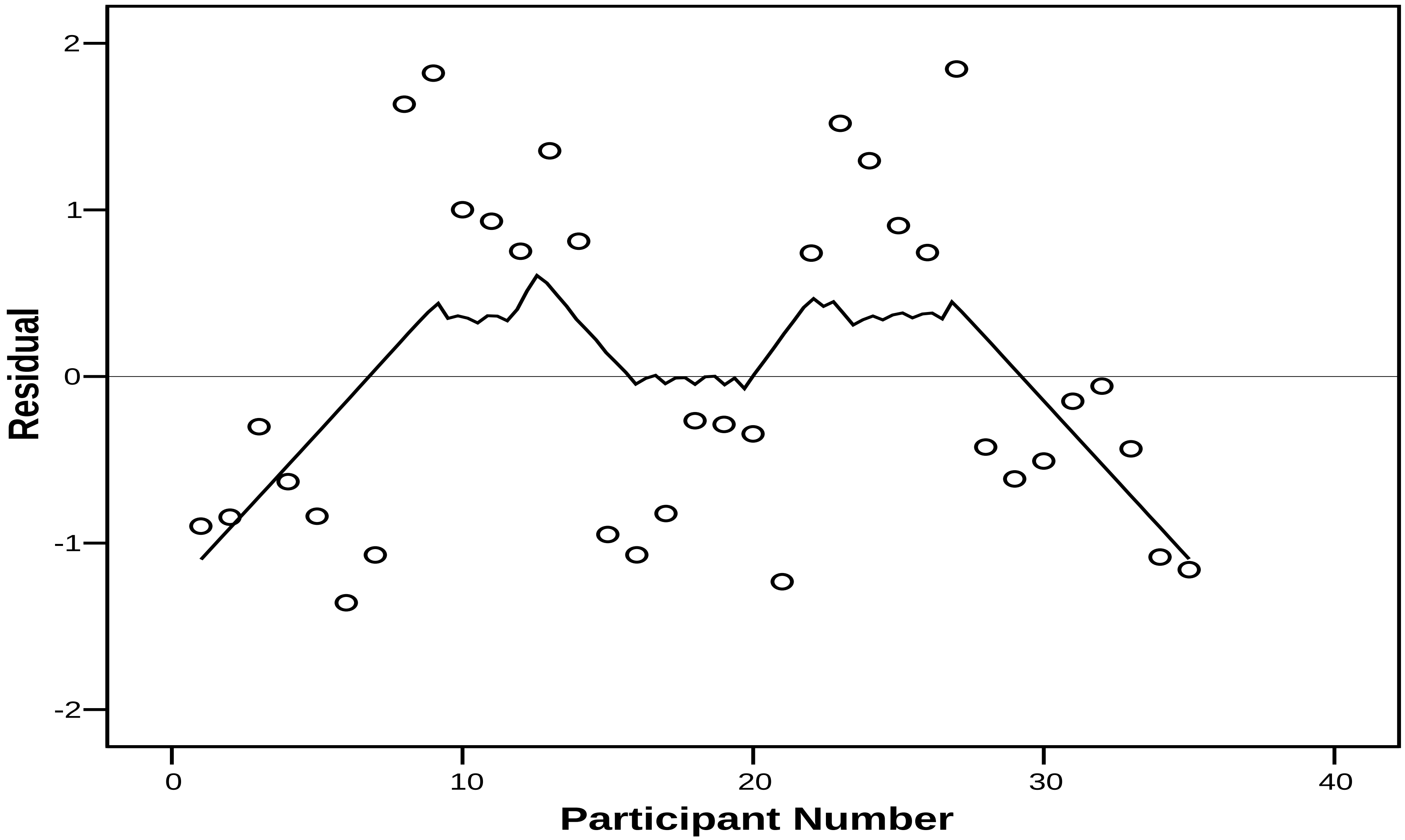
- Also: lack of autocorrelation
- Tricky one
  - often ignored
  - exists for almost all tests
- All cases should be independent of one another
  - knowing the value of one case should not tell you anything about the value of other cases

How is it Detected?

- Can be difficult
  - need some clever statistics (multilevel models)
- Better off avoiding situations where it arises
- Residual Plots
- Durbin-Watson Test
- *d* = 2 indicates no autocorrelation, d<2 there is evidence of positive serial correlation, d>2 successive error terms are negatively correlated.

Residual Plots

- Were data collected in time order?
  - If so plot ID number against the residuals
  - Look for any pattern
    - Test for linear relationship
    - Non-linear relationship
    - Heteroscedasticity



Residual
2
1
0
-1
-2
0
10
20
30
40
Participant Number

How does it arise?

Two main ways

- time-series analyses
  - When cases are time periods
    - weather on Tuesday and weather on Wednesday correlated
    - inflation 1972, inflation 1973 are correlated
- clusters of cases
  - patients treated by three doctors
  - children from different classes
  - people assessed in groups

Why does it matter?

- Standard errors can be wrong
  - therefore significance tests can be wrong
- Parameter estimates can be wrong
  - really, really wrong
  - from positive to negative

Assumption 5: No independent variables are a perfect linear function of other independent variables

No Perfect Multicollinearity

- IVs must not be linear functions of one another
  - matrix of correlations of IVs is not positive definite
  - cannot be inverted
  - analysis cannot proceed
- Have seen this with
  - age, age start, time working
  - also occurs with subscale and total

Test for Goodness of Fit

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$d.f. = \textit{Number of categories} - 1$

$O = \textit{Observed frequency}$

$E = \textit{expected frequency}$

What is Probability?

- In Chapters 2, we used graphs and numerical measures to describe data sets which were usually **samples.**
- We measured "how often" using

Relative frequency = *f*/*n*

- As *n* gets larger,

Sample
And "How often" → **Population**
= Relative frequency → **Probability**

Basic Concepts

- An **experiment** is the process by which an observation (or measurement) is obtained.
- An **event** is an outcome of an experiment, usually denoted by a capital letter.
  - The basic element to which probability is applied
  - When an experiment is performed, a particular event either happens, or it doesn’t!

Basic Concepts

- An event that cannot be decomposed is called a **simple event**.
- Denoted by E with a subscript.
- Each simple event will be assigned a probability, measuring “how often” it occurs.
- The set of all simple events of an experiment is called the **sample space**, **S**.

The Probability of an Event

- The probability of an event A measures "how often" A will occur. We write **P(A).**
- Suppose that an experiment is performed *n* times. The relative frequency for an event A is

$$\frac{\text{Number of times A occurs}}{n} = \frac{f}{n}$$

- If we let *n* get infinitely large,

$$\boxed{P(A) = \lim_{n\to\infty} \frac{f}{n}}$$

The Probability of an Event

- P(A) must be between 0 and 1.
  - If event A can never occur, P(A) = 0. If event A always occurs when the experiment is performed, P(A) =1.
- The sum of the probabilities for all simple events in S equals 1.

- The **probability of an event A** is found by adding the probabilities of all the simple events contained in A.

Counting Rules

- Sample space of throwing 3 dice has 216 entries, sample space of throwing 4 dice has 1296 entries, …
- At some point, we have to stop listing and start thinking …
- We need some counting rules

The *mn* Rule

- If an experiment is performed in two stages, with ***m*** ways to accomplish the first stage and ***n*** ways to accomplish the second stage, then there are ***mn*** ways to accomplish the experiment.
- This rule is easily extended to ***k*** stages, with the number of ways equal to

$$n_1 n_2 n_3 \ldots n_k$$

Example: Toss two coins. The total number of simple events is:

$$2 \times 2 = 4$$

Permutations

- The number of ways you can arrange ***n*** distinct objects, taking them ***r*** at a time is

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

$$\text{where } n! = n(n-1)(n-2)\ldots(2)(1) \text{ and } 0! \equiv 1.$$

Example: How many 3-digit lock combinations can we make from the numbers 1, 2, 3, and 4?

Combinations

- The number of distinct combinations of ***n*** distinct objects that can be formed, taking them ***r*** at a time is

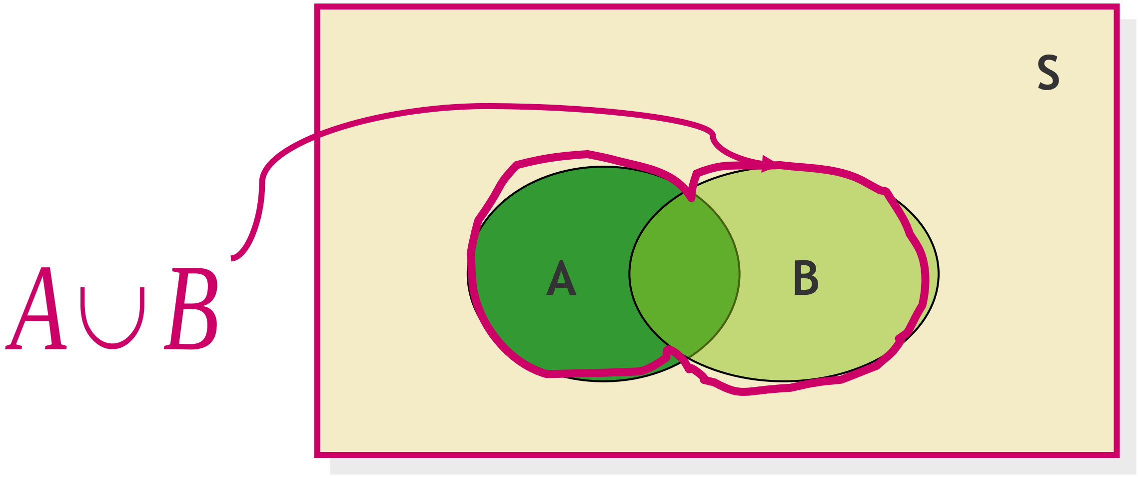
$$C_r^n = \frac{n!}{r!(n-r)!}$$

Example: Three members of a 5-person committee must be chosen to form a subcommittee. How many different subcommittees could be formed?

Event Relations

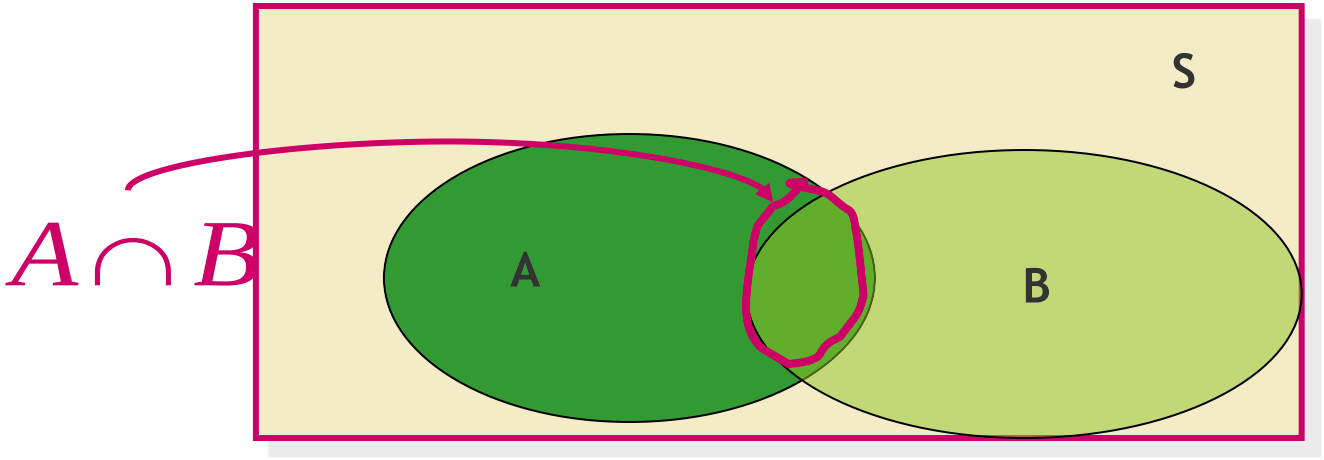
The beauty of using events, rather than simple events, is that we can **combine** events to make other events using logical operations: and, or and not.

The **union** of two events, A and B, is the event that either **A or B or both** occur when the experiment is performed. We write $A \cup B$



Event Relations

The **intersection** of two events, **A** and **B**, is the event that both A **and** B occur when the experiment is performed. We write **A $\cap$ B**.

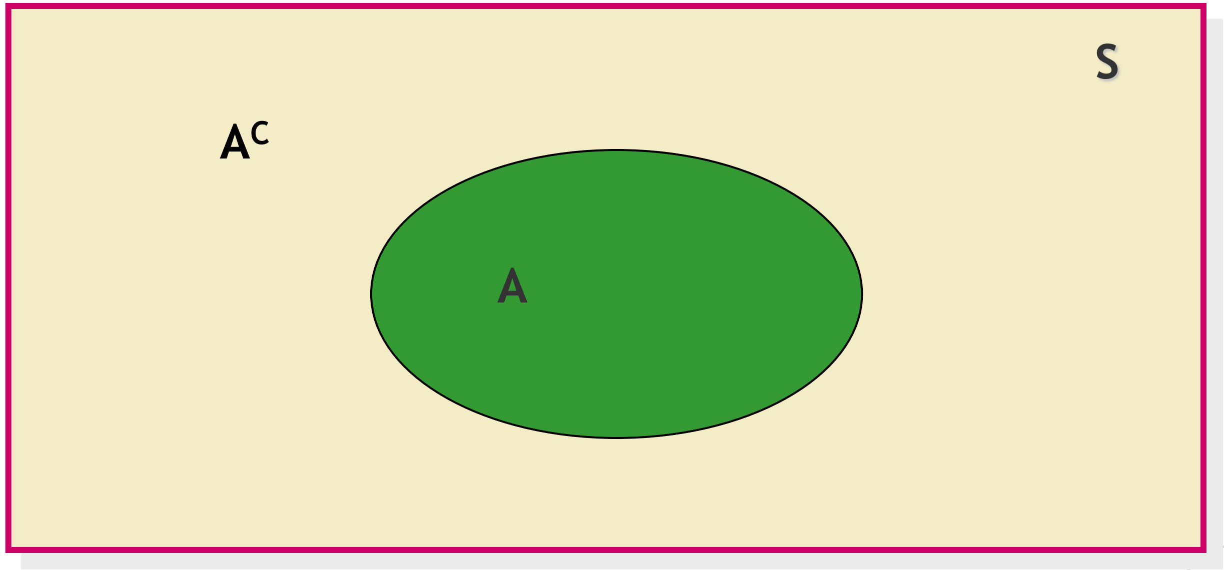


- If two events A and B are **mutually exclusive**, then

P(A $\cap$ B) = 0.

Event Relations

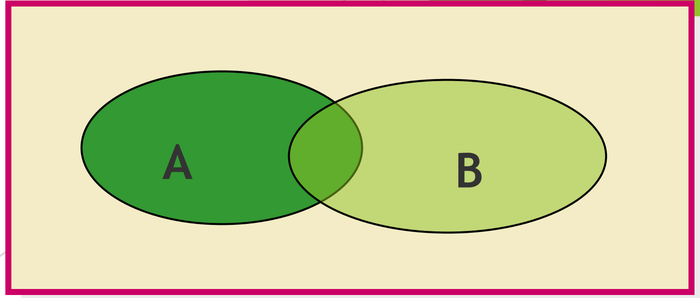
The **complement** of an event A consists of all outcomes of the experiment that do not result in event A. We write A^C.



Calculating Probabilities for Unions and Complements

- There are special rules that will allow you to calculate probabilities for composite events.
- **The Additive Rule for Unions:**
- For any two events, **A** and **B**, the probability of their union, **P(A ∪ B)**, is

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



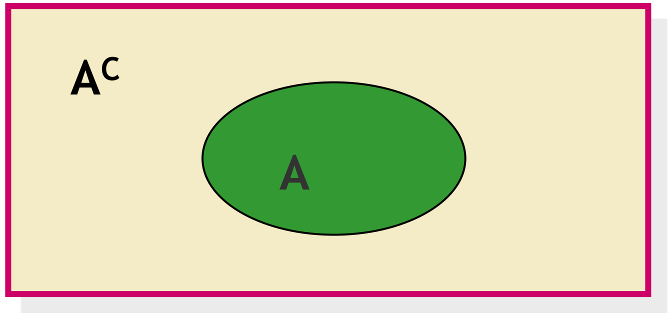
A Special Case

When two events A and B are **mutually exclusive**,
$P(A \cap B) = 0$
and $\mathbf{P(A \cup B) = P(A) + P(B)}$.

Calculating Probabilities for Complements

- We know that for any event **A:**
  - $\mathbf{P(A \cap A^C) = 0}$
- Since either **A** or $\mathbf{A^C}$ must occur,

  $\mathbf{P(A \cup A^C) = 1}$
- so that $\mathbf{P(A \cup A^C) = P(A) + P(A^C) = 1}$



$$\boxed{P(A^C) = 1 - P(A)}$$

Calculating Probabilities for Intersections

In the previous example, we found P(A ∩ B) directly from the table. Sometimes this is impractical or impossible. The rule for calculating P(A ∩ B) depends on the idea of **independent and dependent events.**

Two events, **A** and **B**, are said to be **independent** if the occurrence or nonoccurrence of one of the events does not change the probability of the occurrence of the other event.

Conditional Probabilities

The probability that A occurs, given that event B has occurred is called the **conditional probability** of A given B and is defined as

$$\boxed{P(A\,|\,B) = \frac{P(A \cap B)}{P(B)} \text{ if } P(B) \neq 0}$$

"given"

Defining Independence

- We can redefine independence in terms of conditional probabilities:

Two events A and B are **independent** if and only if

$$P(A|B) = P(A) \qquad \text{or} \qquad P(B|A) = P(B)$$

Otherwise, they are **dependent**.

- Once you've decided whether or not two events are independent, you can use the following rule to calculate their intersection.

The Multiplicative Rule for Intersections

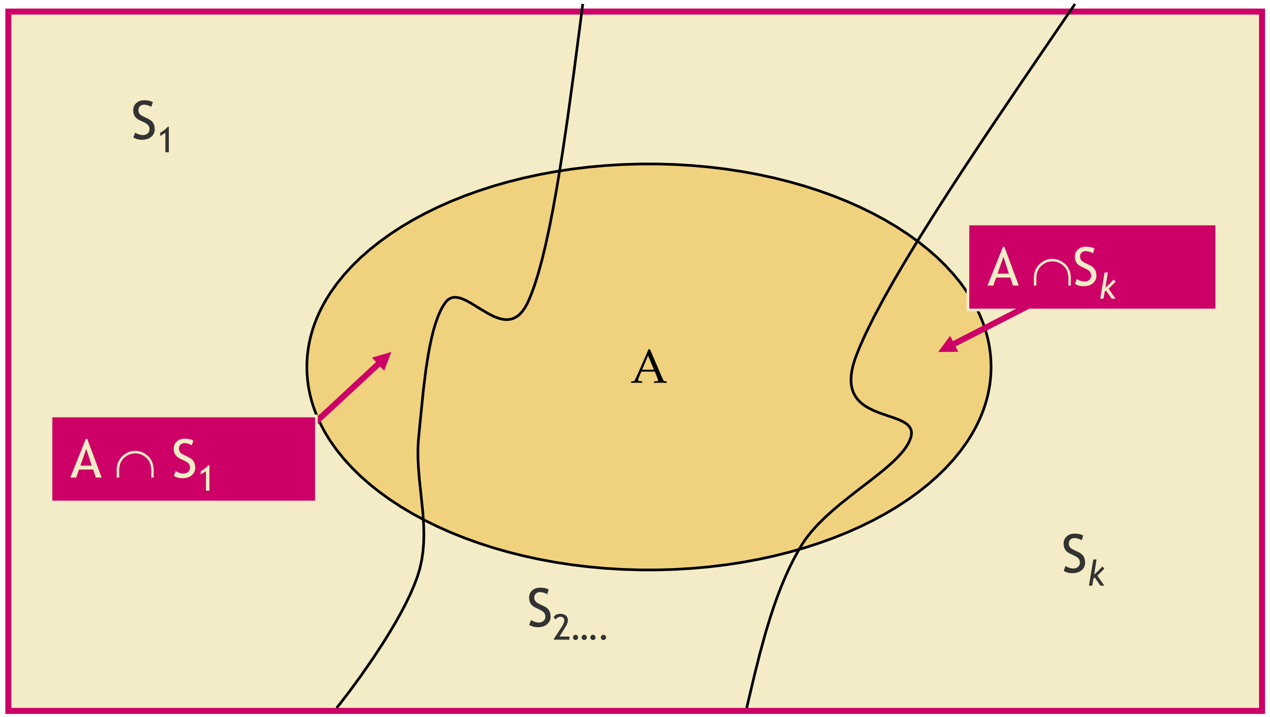
- For any two events, **A** and **B**, the probability that both **A** and **B** occur is

$$\boxed{\begin{aligned} P(A \cap B) &= P(A)\ P(B \text{ given that A occurred}) \\ &= P(A)P(B|A) \end{aligned}}$$

- If the events **A** and **B** are independent, then the probability that both **A** and **B** occur is

$$\boxed{P(A \cap B) = P(A)\ P(B)}$$

The Law of Total Probability



$$P(A) = P(A \cap S_1) + P(A \cap S_2) + \ldots + P(A \cap S_k)$$
$$= P(S_1)P(A|S_1) + P(S_2)P(A|S_2) + \ldots + P(S_k)P(A|S_k)$$

Bayes' Rule

Let S_1 , S_2 , S_3 ,…, S_k be mutually exclusive and exhaustive events with prior probabilities $P(S_1)$, $P(S_2)$,…,$P(S_k)$. If an event A occurs, the posterior probability of S_i, given that A occurred is

$$P(S_i \mid A) = \frac{P(S_i)P(A \mid S_i)}{\sum P(S_i)P(A \mid S_i)} \text{ for } i = 1, 2, \ldots k$$

Example

we know that 49% of the population are female. Of the female patients, 8% are high risk for heart attack, while 12% of the male patients are high risk. A single person is selected at random and found to be high risk. What is the probability that it is a male?

We know:

P(F) =	.49
P(M) =	.51
P(H|F) =	.08
P(H|M) =	.12

$$P(M \mid H) = \frac{P(M)P(H \mid M)}{P(M)P(H \mid M) + P(F)P(H \mid F)}$$

$$= \frac{.51(.12)}{.51(.12) + .49(.08)} = .61$$

Probability Distributions for Discrete Random Variables

The **probability distribution for a discrete random variable *x*** resembles the relative frequency distributions we constructed in Chapter 2. It is a graph, table or formula that gives the possible values of *x* and the probability *p(x)* associated with each value.

We must have

$$0 \le p(x) \le 1 \text{ and } \sum p(x) = 1$$

Key Concepts

I. Experiments and the Sample Space

1. Experiments, events, mutually exclusive events, simple events
2. The sample space

II. Probabilities

1. Relative frequency definition of probability
2. Properties of probabilities
   a. Each probability lies between 0 and 1.
   b. Sum of all simple-event probabilities equals 1.
3. $P(A)$, the sum of the probabilities for all simple events in A

Key Concepts

III. Counting Rules

1. *mn* Rule; extended *mn* Rule
2. Permutations: $$P_r^n = \frac{n!}{(n-r)!}$$
3. Combinations: $$C_r^n = \frac{n!}{r!(n-r)!}$$

IV. Event Relations

1. Unions and intersections
2. Events

   a. Disjoint or mutually exclusive: $P(A \cap B) = 0$

   b. Complementary: $P(A) = 1 - P(A^C)$

Key Concepts

3. Conditional probability: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
4. Independent and dependent events
5. Additive Rule of Probability:

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

6. Multiplicative Rule of Probability:

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

7. Law of Total Probability
8. Bayes' Rule