MONASH University

Information Technology

FIT5190 Introduction to IT Research Methods

Lecture 8
Quantitative Data Analysis
– Probability and Hypothesis Testing

Slides prepared by

David Green, Frada Burstein, Jacques Steyn, Geoff Webb, Chung-Hsing Yeh

Learning objectives

- Understand
  - The basics of probability theory
  - The basics of hypothesis testing using statistics
- Be able to
  - Define the sample spaces of simple experiments
  - Calculate simple probability
  - Define simple statistical hypotheses
  - Carry out a simple chi-square test of contingency

Overview

- This lecture introduces the use of elementary probability and statistics in hypothesis testing research.

What are statistics?

- Summary information about data
- Descriptive statistics
  - Describe key features and patterns in data
- Inferential statistics
  - Inference about the whole population based on a sample
  - Most commonly used in testing hypotheses

Populations versus samples

- Population
  - An entire set of objects
  - e.g. all people in Melbourne
- Sample
  - A set of objects selected to represent a population
  - e.g. the people interviewed in a survey
- Role of samples
  - Complete surveys are usually impossible
  - Statistics help us make inferences about a population from a sample

Concepts of frequency

- Frequency: number of occurrences
  - e.g. number of times throwing a dice: 1, 2, 3,..., n
- Relative frequency: proportion of the total number of occurrences
  - e.g. from 100 throws, number 3 was thrown 18 times
- Cumulative frequency: add up all sets of occurrences
  - e.g. throw dice 10 times today, 5 times tomorrow, accumulates to 15 throws

Basics of probability

Probability

- What is it?
  - A measure of the chance that a particular outcome will occur at random
- Assumption
  - Unknown factors, which we do not control, even out
- Scale
  - Values lie in the range [0,1]
  - The probabilities of all possible outcomes sum to 1

Sample spaces

- The sample space of an “experiment” is the set all possible outcomes.
- Example: the coin-toss experiment
  - Imagine that you toss a coin 5 times.
  - Write down the results as a string of heads (H) and tails (T).
  - Sample space is all possible sequences, HTTHT.
  - 32 possible sequences (i.e. outcomes) in all.
  - Each outcome is equally probable.
  - So for each outcome the probability p = 1/32.

Events and probability

- An “Event” *E* is a set of outcomes
  - e.g. “4 in a row”, “at least one H”
- The “Probability” *p*(*E*) of an event *E*:
  - It is the sum of the probabilities of all (mutually exclusive) outcomes involved:

$$p(E) = \sum_{x \in E} p(x)$$

- Example: *p*(“4 in a row”) =

  p(HHHHT) + *p*(THHHH) + *p*(TTTTH) + *p*(HTTTT)

  = 4/32 = 1/8

Sample space - coin toss experiment

HHHHH
HHHHT
HHHTH
HHHTT
HHTHH
HHTHT
HHTTH
HHTTT

HTHHH
HTHHT
HTHTH
HTHTT
HTTHH
HTTHT
HTTTH
HTTTT

THHHH
THHHT
THHTH
THHTT
THTHH
THTHT
THTTH
THTTT

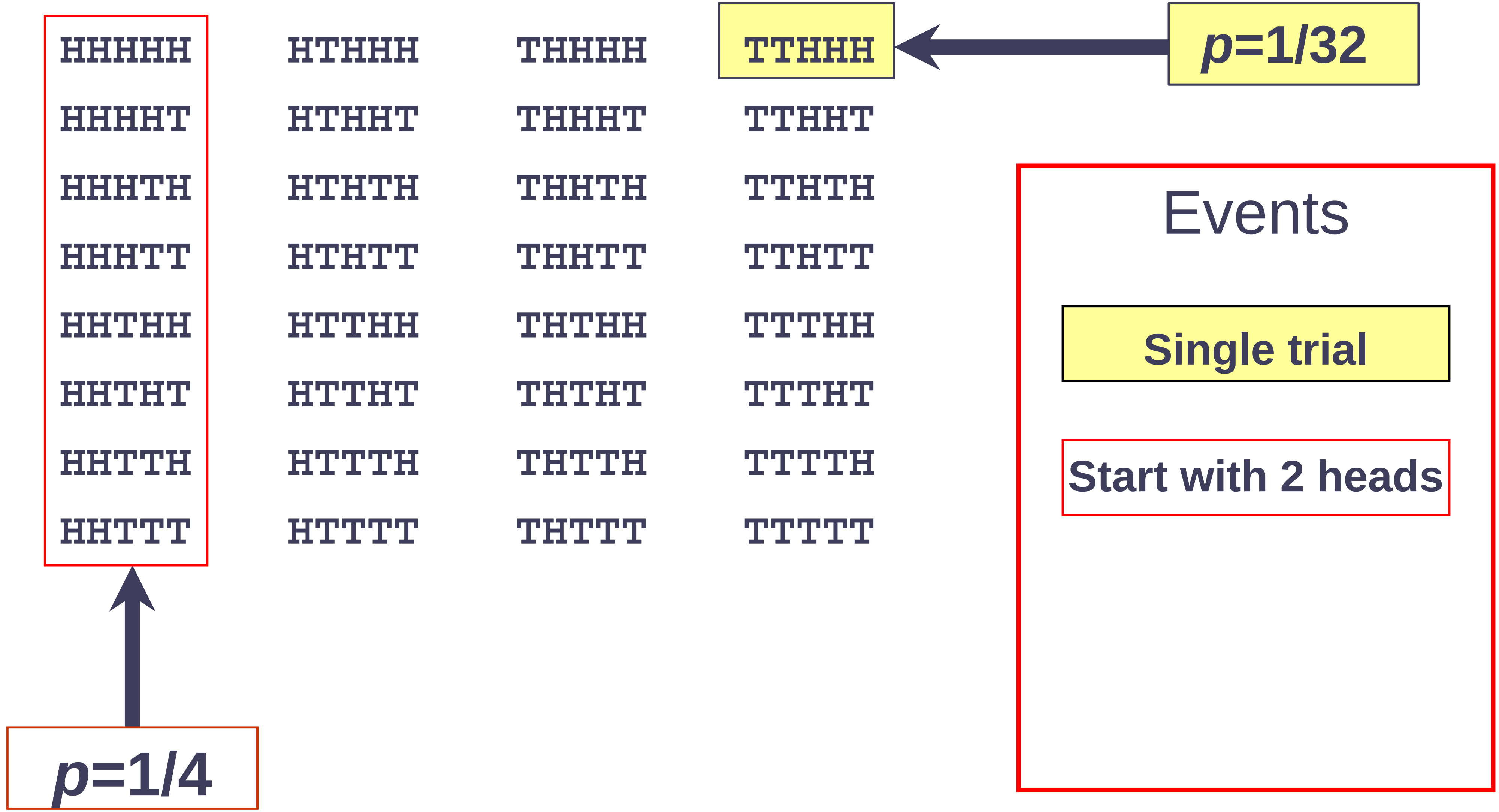
TTHHH
TTHHT
TTHTH
TTHTT
TTTHH
TTTHT
TTTTH
TTTTT

p=1/32

Events

Single trial

Sample space - coin toss experiment



HHHHH
HHHHT
HHHTH
HHHTT
HHTHH
HHTHT
HHTTH
HHTTT

HTHHH
HTHHT
HTHTH
HTHTT
HTTHH
HTTHT
HTTTH
HTTTT

THHHH
THHHT
THHTH
THHTT
THTHH
THTHT
THTTH
THTTT

TTHHH
TTHHT
TTHTH
TTHTT
TTTHH
TTTHT
TTTTH
TTTTT

p=1/32

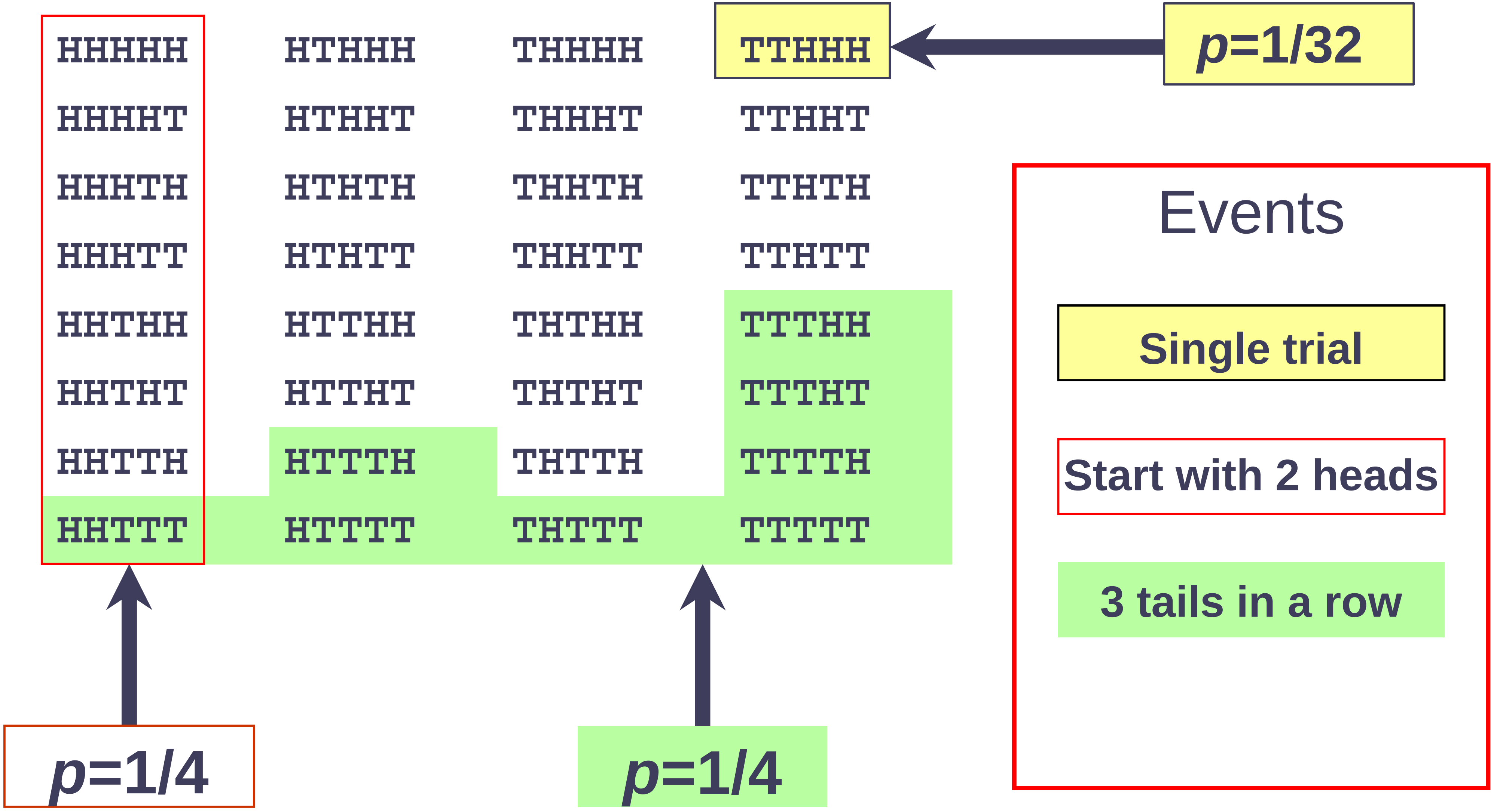
p=1/4

Events

Single trial

Start with 2 heads

Sample space - coin toss experiment



HHHHH	HTHHH	THHHH	TTHHH
HHHHT	HTHHT	THHHT	TTHHT
HHHTH	HTHTH	THHTH	TTHTH
HHHTT	HTHTT	THHTT	TTHTT
HHTHH	HTTHH	THTHH	TTTHH
HHTHT	HTTHT	THTHT	TTTHT
HHTTH	HTTTH	THTTH	TTTTH
HHTTT	HTTTT	THTTT	TTTTT

p=1/32

p=1/4

p=1/4

Events

Single trial

Start with 2 heads

3 tails in a row

Note

- *p*(E) is not the sum of *p*(*x*) for sub-events *x* if *x* events are not mutually exclusive

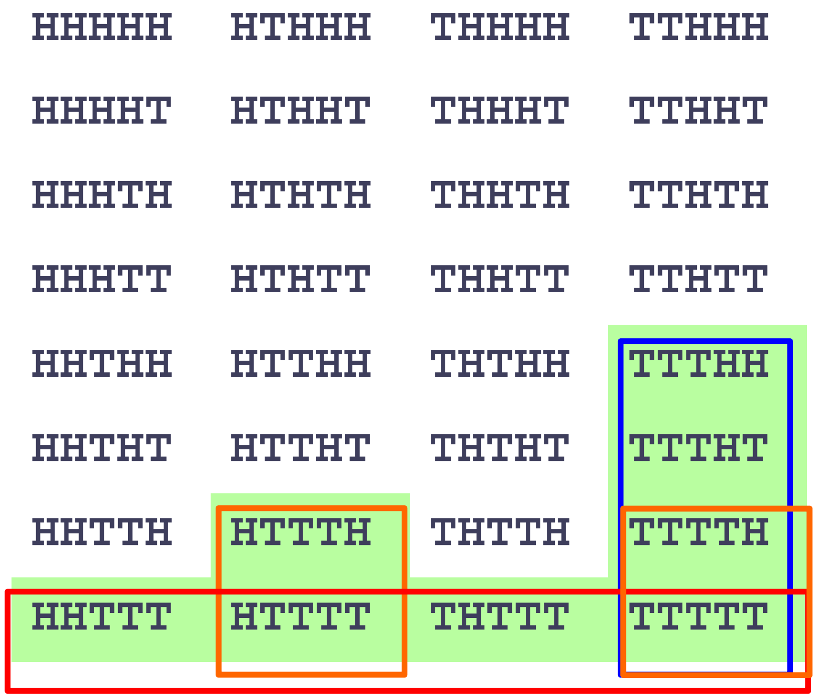
e.g.

–*p*(3 tails in a row) = 8/32

–*p*(3 tails at start) = 4/32

–*p*(3 tails at end) = 4/32

–*p*(3 tails in middle) = 4/32

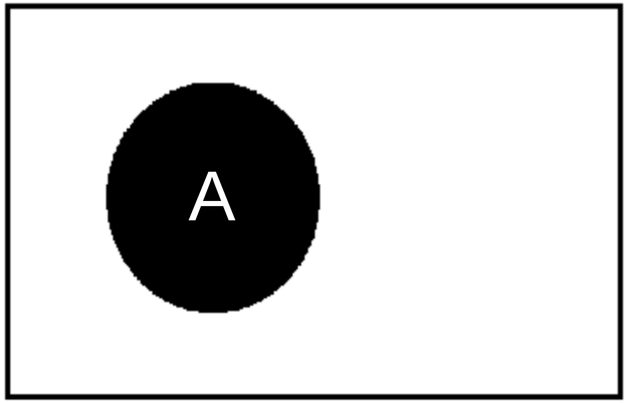


HHHHH	HTHHH	THHHH	TTHHH
HHHHT	HTHHT	THHHT	TTHHT
HHHTH	HTHTH	THHTH	TTHTH
HHHTT	HTHTT	THHTT	TTHTT
HHTHH	HTTHH	THTHH	TTTHH
HHTHT	HTTHT	THTHT	TTTHT
HHTTH	HTTTH	THTTH	TTTTH
HHTTT	HTTTT	THTTT	TTTTT

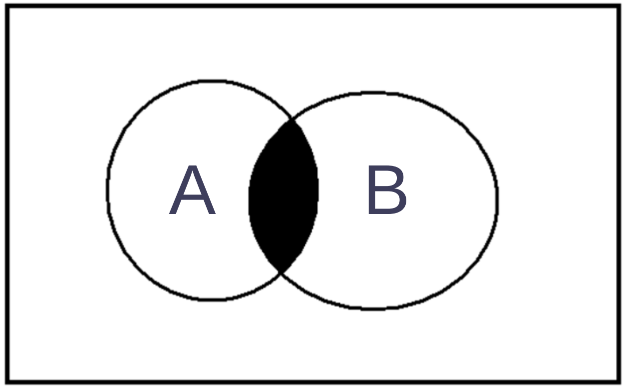
Some basic rules of probability (I)

- Note the link between probability, sample spaces, predicates sets and Venn diagrams, e.g.

$p(\sim A) = 1 - p(A)$



$p(A \text{ or } B) = p(A) + p(B) - p(A\&B)$



Example

- Event(No heads in sequence) = { TTTTT }
  - *p*(No heads) = *p*({ TTTTT }) = 1/32

- Event(At least 1 head) = S – {TTTTT}
  - So *p*(At least 1 head) = 1 - *p*(No heads)

    = 1 – 1/32 = 31/32

Some basic rules of probability (II)

- $p(\Phi) = 0$ $\Phi = \{\} =$ No outcomes
- $p(U) = 1$ U = All possible outcomes
- $p(\sim A) = 1 - p(A)$ ~A = NOT A
- If $B \subset A$, then $p(B) < p(A)$
- $p(A \vee B) = p(A) + p(B) - p(A \& B)$

Event independence

- Events A and B are *independent* if

  p(A & B) = p(A) p(B)

- Are these events independent?
  - H???? and ????H
    - *P(H????)* = 16/32 = ½
    - P(????H) = 16/32 = ½
    - P(H???? & ????H) = 8/32 = ¼ = ½ × ½
  - So these events **are** independent

HHHHH	HTHHH	THHHH	TTHHH
HHHHT	HTHHT	THHHT	TTHHT
HHHTH	HTHTH	THHTH	TTHTH
HHHTT	HTHTT	THHTT	TTHTT
HHTHH	HTTHH	THTHH	TTTHH
HHTHT	HTTHT	THTHT	TTTHT
HHTTH	HTTTH	THTTH	TTTTH
HHTTT	HTTTT	THTTT	TTTTT

Event dependence

- Events A and B are *independent* if

  p(A & B) = p(A) p(B)

- Are the following events independent?
  - H???? and HH???
    - P(H????) = 16/32 = ½
    - P(HH???) = 8/32 = ¼
    - P(H???? & ????H) = 8/32 = ¼

      ≠ ½ × ¼
  - So the events **are not** independent

HHHHH	HTHHH	THHHH	TTHHH
HHHHT	HTHHT	THHHT	TTHHT
HHHTH	HTHTH	THHTH	TTHTH
HHHTT	HTHTT	THHTT	TTHTT
HHTHH	HTTHH	THTHH	TTTHH
HHTHT	HTTHT	THTHT	TTTHT
HHTTH	HTTTH	THTTH	TTTTH
HHTTT	HTTTT	THTTT	TTTTT

Hypothesis testing

Hypothesis

- Research is guided by the specific research problem, question, or hypothesis.
- Once we state the problem and the sub-problems, each sub-problem is then viewed through a construct called a hypothesis.
- A hypothesis is a logical supposition, a reasonable guess, an educated conjecture.

Hypothesis: an example

- You come home after dark, open the door and reach to turn on the lamp. No light.
  - The bulb has burned out
  - The lamp is not plugged into the wall outlet
  - The thunderstorm interrupted the electrical service
  - The wire from the wall to the lamp is defective
  - You forgot to pay your electric bill
- Each of these hypotheses provides a direction for exploration to locate the information that may solve the problem of the malfunction.

Hypothesis testing

Hypothesis testing has two meanings

- The first meaning restricts the word hypothesis to a research problem oriented hypothesis (e.g. the example on the previous slide)
  - Here the hypothesis was a reasonable and logical guess
- The second meaning is limited to a statistically oriented hypothesis
  - Here the word hypothesis refers to a statistically based hypothesis, the null hypothesis
  - Comparison of observed data with the expected results is called testing the hypothesis, or testing the null hypothesis

Studying quantitative data analysis

- Studying quantitative data analysis helps some PhD students, not all however.
  - H_0 : the result is mere chance
  - H_1 : the result is a real difference
  - Set an “acceptance level” at 0.05
  - Assuming the null hypothesis H_0, you find the probability of the result is 0.045.
  - Conclusion: Reject H_0 and accept the alternative H_1.

Hypothesis testing

- Conventional approach:
  - H_0 is a “default” hypothesis, which we retain unless there is sufficient evidence it is wrong.
- The problem
  - Do we retain the null hypothesis H_0?
  - Or reject it in favour of the alternative H_1?

Hypothesis testing

- Hypotheses often amount to statements about numerical differences between samples.
- The problem is that differences are likely to occur by chance.
- We use statistics to determine the probability of a difference being the result of chance.

Probability and hypotheses

- Many experiments have no clear cut outcome.
  - A vaccine may cure some people, but not others.
  - So how do you determine whether the results support your hypothesis?
  - Probability provides a way to decide.
- Statistical Hypotheses
  - Treat hypotheses as statements about a population (e.g. mean, variance).
  - H_0 - Null hypothesis
  - H_1 - Alternative hypothesis

Statistical testing

- The method
  - Use a set of observations
  - Set a "significance level" α
  - Assuming the null hypothesis H_0 is true, calculate the probability p_0 of getting the observed result (*p*-value).
    - If $p_0 > \alpha$ retain the null hypothesis
    - If $p_0 < \alpha$ reject the null hypothesis
- Standard practice
  - Set α low, so we minimise inferential errors.
  - Most studies set $\alpha = 0.01$ or 0.05.
  - Most journals require significance level of a test

* As used in statistics, "significance" does not mean important or meaningful.

Statistical hypothesis tests

- A *statistic* is a number calculated from a sample of observations.
  - e.g. Average crop size
- Assuming the null hypothesis H_0
  - What is the probability of getting that number purely by chance?
  - e.g. What is the probability of getting a bigger crop by chance?
- Statistical significance
  - The probability that the result is not chance.

Example - vaccine trial

- The vaccine cures most people, but not all
  - H_0 : the result is mere chance
  - H_1 : the result is a real difference
- Set a “significance level” α =0.05
  - Assuming the null hypothesis H_0, you find the probability of the result (mere chance) is 0.049, i.e. $< \alpha$
  - Conclusion: Reject H_0 and accept the alternative
- But beware!
  - A probability of 0.05 (=1/20) means that if you repeat an experiment 20 times, you would expect to get a positive result at least once by chance!

Contingency tables

- Hypotheses often imply that results will fall into certain categories.
- These categories form a contingency table of frequencies.

Fertilizer experiment #1

- *Null hypothesis* H_0:
  - “Fertilization makes no difference to plant size”
- *Alternative hypothesis* H_1:
  - “Fertilization makes a difference to plant size”
- Experiment:
  - Grow equal numbers of fertilized and unfertilized plants.
  - Record the results in a **contingency table**.

Contingency table

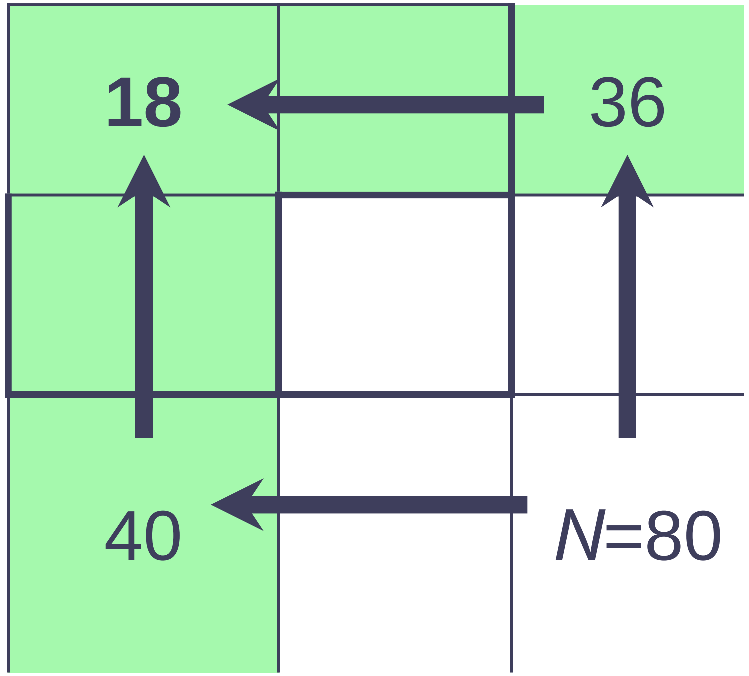
	No Fertilizer	*Fertilizer*	
Small	24	12	**36**
Large	16	28	**44**
	40	**40**	**80**

N = total number of cases

Table entries are **OBSERVED** numbers of plants

Expected frequencies

- According to the Null Hypothesis $H_{0,}$ whether an item falls into a particular row should be *independent* of whether it falls into a particular column.
- This means that we would expect entries to reflect frequencies in rows and columns.



80 X 40/80 X 36/80 = 18

Expected frequencies

The *EXPECTED* number E_{ij} of results in the cell (i,j) is given by the formula

$$E_{ij} = \frac{R_i C_j}{N}$$

where

- R_i = Total number of results in row i
- C_j = Total number of results in column j
- N = Total number of results

Contingency table

	No Fertilizer	*Fertilizer*	**Totals**
Small	**24/18**	**12/18**	**36**
Large	**16/22**	**28/22**	**44**
Totals	**40**	**40**	**80**

N = total number of cases = 80

Table entries here are **OBSERVED / EXPECTED**

Chi-square test

χ^2

To determine whether distributions of categorical variables differ from one another.

Chi-squared distribution

- For a random distribution of results, the sum of the deviations follow a definite distribution.
  - Known as the *chi-squared distribution*

- Chi squared

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O: Observed numbers

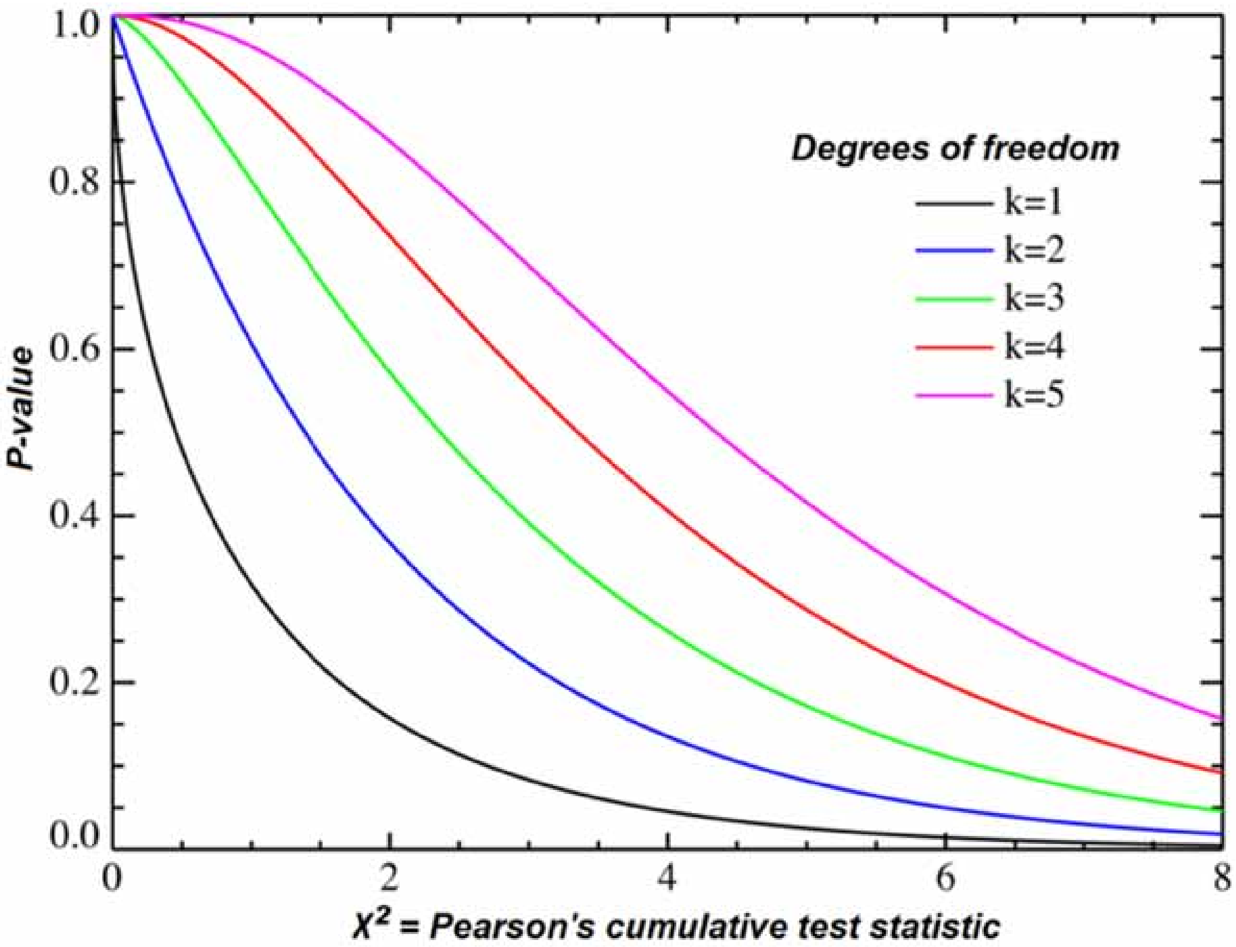
E: Expected numbers

- Example: for O=24, E=18,

$(O\text{-}E)^2/E = (24\text{-}18)^2/18 = 6^2/18 = 36/18 = 2$

Chi-squared distribution

- *p*-values associated with values of χ^2 depend on the degrees of freedom.
- The distribution shows the probability of different levels of variation, assuming that each outcome is equally likely.



Degrees of freedom

- The extent of change depends on the number of *different* ways the entries can be changed.
- This flexibility is called the *degree of freedom* (abbreviated *df*).
- For a contingency table, *df* is the number of entries that you need to know in order to be able to determine the values of all entries.

	No Fertilizer	*Fertilizer*	**Totals**
Small	**x**	**36-x**	**36**
Large	**40-x**	**44-(40-x)**	**44**
Totals	**40**	**40**	**80**

- So *df* = (no of rows – 1)×(no of columns-1)

Chi-square calculation

	No Fertilizer	*Fertilizer*	
Small	24/18 **$(24-18)^2/18=2$**	12/18 **$(12-18)^2/18=2$**	**36**
Large	16/22 **$(16-22)^2/22=1.63$**	28/22 **$(16-22)^2/22=1.63$**	**44**
	40	**40**	**80**

$$\chi^2 = 2 + 2 + 1.63 + 1.63 = 7.26$$

Chi-square calculation

- *Chi-square value*
  - $\chi^2 = 2 + 2 + 1.63 + 1.63 = 7.26$
- *Degrees of freedom*
  - $df = (2-1)*(2-1) = 1$
- *p-value*
  - $\chi^2(7.26,1) = 0.00705$
  - So the chance of this result under H_0 is only 0.007
- Conclusion
  - Reject the Null hypothesis
  - Accept that fertilizer DOES increase plant size

Another example: a medical experiment

	No Treatment	Placebo	Treatment	Totals
Improvement	53	59	61	173
No change	47	41	39	127
Totals	100	100	100	300

Chi-square analysis

	No Treatment	Placebo	Treatment	Totals
Improvement	53 (59.33)	59 (59.33)	61 (59.33)	173
No change	47 (40.67)	41 (40.67)	39 (40.67)	127
Totals	100	100	100	300

χ2 = 3.51

df = (2-1)*(3-1) = 2

p = *0.173*

Example: medical experiment

- The above chi-square analysis implies that the treatment does not work.
- If you repeat the analysis without the placebo, the chi-square analysis gives this result:
  - $\chi^2 = 3.51$
  - $df = (2\text{-}1)*(2\text{-}1) = 1$
  - $p = 0.061$
- In other words, the result implies a marginally significant improvement over no treatment.
- In this case the treatment does not perform significantly better than the placebo.

Readings

- Web Center for Social Research Methods
  - http://www.socialresearchmethods.net/
  - See the section on Selecting Statistics.
- Tutorial on tests of significance
  - http://www.csulb.edu/~msaintg/ppa696/696stsig.htm
- McREL (2004) *Tutorial on Understanding Statistics*. ECS.
  - http://www.ecs.org/html/educationIssues/Research/primer/understandingtutorial.asp
  - See the section on Selecting Statistics.