

**Associations and Correlations**



***The Essential Elements***

A Holistic Strategy for

Discovering the Story of Your Data

LEE BAKER

CEO

Chi-Squared  Innovations

**Contents**

**Introduction**

**Section  1:  Preparing  Your  Data**

**Chapter  1:  Data  Collection  and  Cleaning**

1.1:  Data  Collection

1.2:  Data  Cleaning

**Chapter  2:  Data  Classification**

2.1:  Quantitative  and  Qualitative  Data

**Section  2:  Your  Statistical  Toolbox**

**Chapter  3:  Introduction  to  Associations  and  Correlations**

3.1:  What  are  Associations  and  Correlations?

3.2:  Discovering  the  Story  of  Your  Data

3.3:  Correlation  v  Causation

3.4:  Assessing  Relationships

3.5:  Assessing  the  Strength  of  Relationships

**Chapter  4:  Univariate  Statistics**

4.1:  Correlations

4.2:  Associations

4.3:  Survival  Analysis

**Chapter  5:  Multivariate  Statistics**

5.1  Types  of  Multivariate  Analysis

5.2  Using  Multivariate  Tests  As  Univariate  Tests

5.3  Univariate  v  Multivariate  Analyses

5.4  Limitations  &  Assumptions  of  Multivariate  Analysis

5.5  Step-Wise  Methods  of  Analysis

5.6  Creating  Predictive  Models  with  the  Results  of  Multivariate  Analyses

**Section  3:  The  Story  of  Your  Data**

**Chapter  6:  Visualising  Your  Relationships**

6.1:  A  Holistic  Strategy  to  Discover  Independent  Relationships

6.2:  Visualising  the  Story  of  Your  Data

**Chapter  7:  Bonus:  Automating  Associations  and  Correlations**

7.1:  What  is  the  Problem?

7.2:  CorrelViz

**Summary:  A  Holistic  Strategy  to  Discover  the  Story  of  Your  Data**

**Epilogue**

ii

**1**

**4**

**5**

5

7

**8**

8

**14**

**14**

14

17

19

20

22

**24**

24

27

45

**52**

52

64

64

69

70

74

**79**

**80**

80

90

**95**

95

96

**98**

**99**

**Introduction**

Associations  and  correlations  are  perhaps  the  most  used  of  all  statistical  techniques.  As  a

consequence,  they  are  possibly  also  the  most*mis*-used.

The  problem  is  that  the  majority  of  people  that  work  with  association  and  correlation  tests  are  not

statisticians  and  have  little  or  no  statistical  training.  That’s  not  a  criticism,  simply  an

acknowledgement  that  most  researchers,  scientists,  healthcare  practitioners  and  other  investigators

are  specialists  in  things*other  than*  statistics  and  have  limited  –  if  any  –  access  to  statistical

professionals  for  assistance.

So  they  turn  to  statistical  textbooks  and  perceived  knowledge  amongst  their  peers  for  their  training.

I  won’t  dwell  too  much  on  perceived  knowledge,  other  than  to  say  that  both  the  use  and  misuse  of

statistics  passes  through  the  generations  of  researchers  equally.  There’s  a  lot  of  statistical  mis-

information  out  there…

There  are  many  statistical  textbooks  that  explain  everything  you  need  to  know  about  associations

and  correlations,  but  here’s  the  rub:  most  of  them  are  written  by  statisticians  that  understand  in

great  depth  how  statistics  work  and  they  don’t  understand  why  non-statisticians  have  difficulty  with

stats  –  they  have  little  empathy.  Consequently,  many  of  these  textbooks  are  full  of  highly  complex

equations  explaining  the  mathematical  basis  behind  the  statistical  tests,  are  written  with

complicated  statistical  language  that  is  difficult  for  the  beginner  to  penetrate,  and  they  don’t  take

into  account  that  the  reader  just  might  be  looking  into  statistics  for  the  first  time.

Ultimately,  most  statistics  books  are  written*by*  statisticians*for*  statisticians.

In  writing  this  book,  I  was  determined  that  it  would  be  different.

This  is  a  book  for  beginners.  My  hope  is  that  more  experienced  practitioners  might  also  find  value,

but  my  primary  focus  here  is  on  introducing  the  essential  elements  of  association  and  correlation

analyses.  If  you  want  the  finer  points,  then  you’re  plum  out  of  luck  –  you  won’t  find  them  here.  Just

the  essential  stuff.  For  beginners.

There’s  another  issue  I’ve  noticed  with  most  statistical  textbooks,  and  I’ll  use  a  house  building

analogy  to  illustrate  it.

When  house  builders  write  books  about  how  to  build  houses  they  don’t  write  about  hammers  and

screwdrivers.  They  write  about  how  to  prepare  the  foundations,  raise  the  walls  and  fit  the  roof.

When  statisticians  do  their  analyses,  they  think  like  the  house  builder.  They  think  about  how  to  pre-

process  their  data  (prepare  the  foundations),  do  preliminary  investigations  to  get  a  ‘feel’  for  the  data

(raise  the  walls,  see  what  the  whole  thing  will  look  like)  and  finally  they  deduce  the  story  of  the  data

(making  the  build  watertight  by  adding  a  roof).

1

Unfortunately,  that’s  not  how  they  write  books.  Most  statistical  textbooks  deal  with  statistical  tests

in  isolation,  one-by-one.  They  deal  with  the  statistical  tools,  not  the  bigger  picture.  They  don’t  tend

to  discuss  how  information  flows  through  the  data  and  how  to  create  strategies  to  extract  the

information  that  tells  the  story  of  the  whole  dataset.

Here,  I  discuss  a  holistic  method  of  discovering  the  story  of  all  the  relationships  in  your  data  by

introducing  and  using  a  variety  of  the  most  used  association  and  correlation  tests  (and  helping  you

to  choose  them  correctly).  The  holistic  method  is  about  selecting  the  most  appropriate  univariate

and  multivariate  tests  and  using  them  together  in  a  single  strategic  framework  to  give  you

confidence  that  the  story  you  discover  is  likely  to  be  the  true  story  of  your  data.

The  focus  here  is  on  the  utility  of  the  tests  and  on  how  these  tests  fit  into  the  holistic  strategy.  I

don’t  use  any  complicated  maths  (OK,  well,  just  a  little  bit  towards  the  end,  but  it’s  necessary,

honest…),  I  shed  light  on  complicated  statistical  language  (but  I  don’t  actually*use*  it  –  I  use  simple,

easy-to-understand  terminology  instead),  and  I  don’t  discuss  the  more  complex  techniques  that

you’ll  find  in  more  technical  textbooks.

I  have  divided  the  book  into  three  distinct  sections:







Section  1:  Preparing  Your  Data

Section  2:  Your  Statistical  Toolbox

Section  3:  The  Story  of  Your  Data

The  chapters  in  each  of  these  sections  can  be  read  stand-alone  and  referred  to  at  will,  but  build

towards  the  ultimate  aim  of  the  holistic  strategy  of  discovering  the  story  of  your  data.

Chapter  1  briefly  (*very*  briefly)  introduces  data  collection  and  cleaning,  and  outlines  the  basic

features  of  a  dataset  that  is  fit-for-purpose  and  ready  for  analysis.

Chapter  2  discusses  how  to  classify  your  data  and  introduces  the  four  distinct  types  of  data  that

you’ll  likely  have  in  your  dataset.

Chapter  3  introduces  associations  and  correlations,  explains  what  they  are  and  their  importance  in

understanding  the  world  around  us.

Chapter  4  discusses  the  univariate  statistical  tests  that  are  common  in  association  and  correlation

analysis,  and  details  how  and  when  to  use  them,  with  simple  easy-to-understand  examples.

Chapter  5  introduces  the  different  types  of  multivariate  statistics,  how  and  when  to  use  them.  This

chapter  includes  a  discussion  of  confounding,  suppressor  and  interacting  variables,  and  what  to  do

when  your  univariate  and  multivariate  results  do  not  concur  (spoiler  alert:  the  answer  is  not

‘panic’!).

2

Chapter  6  explains  the  holistic  strategy  of  discovering  all  the  independent  relationships  in  your

dataset  and  describes  why  univariate  and  multivariate  techniques  should  be  used  as  a  tag  team.  This

chapter  also  introduces  you  to  the  techniques  of  visualising  the  story  of  your  data.

Chapter  7  is  a  bonus  chapter  that  explains  how  you  can  discover  all  the  associations  and  correlations

in  your  data  automatically,  and  in  minutes  rather  than  months.

I  hope  you  find  something  of  value  in  this  book.  Whether  you  do  (or  not)  I’d  love  to  hear  your

comments.  Send  me  an  email:



[ebookfeedback@chi2innovations.com](mailto:ebookfeedback@chi2innovations.com)

Just  so  you  know,**you  are  free  to  share  this  book**  with  anyone  –  as  long  as  you  don’t  change  it  or

charge  for  it  (the  boring  details  are  at  the  end).

3

**Section   1:**



**Preparing   Your   Data**

**Chapter  1:  Data  Collection  and  Cleaning**

The  first  step  in  any  data  analysis  project  is  to  collect  and  clean  your  data.  If  you’re  fortunate  enough

to  have  been  given  a  perfectly  clean  dataset,  then  congratulations  –  you’re  well  on  your  way.  For  the

rest  of  us  though,  there’s  quite  a  bit  of  grunt  work  to  be  done  before  you  can  get  to  the  joy  of

analysis  (yeah,  I  know,  I  really  must  get  a  life…).

In  this  chapter,  you’ll  learn  the  features  of  what  a  good  dataset  looks  like  and  how  it  should  be

formatted  to  make  it  amenable  to  analysis  by  association  and  correlation  tests.

Most  importantly,  you’ll  learn  why  it’s  not  necessarily  a  good  idea  to  collect  sales  data  on  ice  cream

and  haemorrhoid  cream  in  the  same  dataset.

If  you’re  happy  with  your  dataset  and  quite  sure  that  it  doesn’t  need  cleaning,  then  you  can  safely

skip  this  chapter.  I  won’t  take  it  personally  –  honest!

**1.1:  Data  Collection**

The  first  question  you  should  be  asking  before  starting  any  project  is  ‘What  is  my  question?’.  If  you

don’t  know  your  question,  then  you  won’t  know  how  to  get  an  answer.  In  science  and  statistics,  this

is  called  having  a  hypothesis.  Typical  hypotheses  might  be:







Is  smoking  related  to  lung  cancer?

Is  there  an  association  between  sales  of  ice  cream  and  haemorrhoid  cream?

Is  there  a  correlation  between  coffee  consumption  and  insomnia?

It’s  important  to  start  with  a  question,  because  this  will  help  you  decide  which  data  you  should

collect  (and  which  you  shouldn’t).

It’s  not  usual  that  you  can  answer  these  types  of  question  by  collecting  data  on  just  those  variables.

It’s  much  more  likely  that  there  will  be  other  factors  that  may  have  an  influence  on  the  answer  and

all  of  these  factors  must  be  taken  into  account.  If  you  want  to  answer  the  question  ‘is  smoking

related  to  lung  cancer?’  then  you’ll  typically  also  collect  data  on  age,  height,  weight,  family  history,

genetic  factors,  environmental  factors,  and  your  dataset  will  start  to  become  quite  large  in

comparison  with  your  hypothesis.

So  what  data  should  you  collect?  Well,  that  depends  on  your  hypothesis,  the  perceived  wisdom  of

current  thinking  and  previous  research  carried  out,  but  ultimately,  if  you  collect  data  sensibly  you

will  likely  get  sensible  results  and*vice  versa*,  so  it’s  a  good  idea  to  take  some  time  to  think  it  through

carefully  before  you  start.

I’m  not  going  to  go  into  the  finer  points  of  data  collection  and  cleaning  here,  but  it’s  important  that

your  dataset  conforms  to  a  few  simple  standards  before  you  can  start  analysing  it.

5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **UniqueID** | **Gender** | **Age** | **Weight** | **Height** | **…** |
| 1 | Male | 31 | 76.3 | 1.80 | … |
| 2 | Female | 57 | 59.1 | 1.91 | … |
| 3 | Male | 43 | 66.2 | 1.86 | … |
| 4 | Male | 24 | 64.1 | 1.62 | … |
| … | … | … | … | … | … |

By  the  way,  if  you  want  a  copy  of  my  book  ‘Practical  Data  Cleaning’,  you  can  get  a  free  copy  of  it  by

following  the  instructions  in  the  tiny  little  advert  for  it  at  the  end  of  this  section…

***1.1.1:  Dataset  checklist***

OK,  so  here  we  go.  Here  are  the  essential  features  of  a  ready-to-go  dataset  for  association  and

correlation  analysis:

Your  dataset  is  a  rectangular  matrix  of  data.  If  your  data  is  spread  across  different

spreadsheets  or  tables,  then  it’s  not  a  dataset,  it’s  a  data*base*,  and  it’s  not  ready  for  analysis.

Each  column  of  data  is  a  single  variable  corresponding  to  a  single  piece  of  information  (such

as  age,  height,  weight,  etc.).

Column  1  is  a  list  of  unique  consecutive  numbers  starting  from  one.  This  allows  you  to

uniquely  identify  any  given  row  and  recover  the  original  order  of  your  dataset  with  a  single

sort  command.

Row  1  contains  the  names  of  the  variables.  If  you  use  rows  2,  3,  4,  etc.  as  the  variable  names

you  won’t  be  able  to  enter  your  dataset  into  a  statistics  programme.

Each  row  contains  the  details  for  a  single  sample  (patient,  case,  test-tube,  etc.).

Each  cell  should  contain  a  single  piece  of  information.  If  you  have  entered  more  than  one

piece  of  information  in  a  cell  (such  as  date  of  birth  and  their  age),  then  you  should  separate

the  column  into  two  or  more  columns  (one  for  date  of  birth,  another  for  age).

Don’t  enter  the  number  zero  into  a  cell  unless  what  has  been  measured,  counted  or

calculated  results  in  the  answer  zero.  Don’t  use  the  number  zero  as  a  code  to  signify  ‘No

Data’.

By  now,  you  should  have  a  well-formed  dataset  that  is  stored  in  a  single  Excel  worksheet,  each

column  is  a  single  variable,  row  1  contains  the  names  of  the  variables,  and  below  this  each  row  is  a

distinct  sample  or  patient.  It  should  look  something  like  Figure  1.1.

**Figure  1.1:  A  Typical  Dataset  Used  in  Association  and  Correlation  Analysis**

For  the  rest  of  this  book,  this  is  how  I  assume  your  dataset  is  laid  out,  so  I  might  use  the  terms

Variable  and  Column  interchangeably,  similarly  for  terms  Row,  Sample  and  Patient.

**1.2:  Data  Cleaning**

Your  next  step  is  cleaning  these  data.  You  may  well  have  made  some  entry  errors  and  some  of  your data  may  not  be  useable.  You  need  to  find  these  and  correct  them.  The  alternative  is  that  your  data

may  not  be  fit  for  purpose  and  may  mislead  you  in  pursuit  of  the  answers  to  your  questions.

Even  after  you’ve  corrected  the  obvious  entry  errors,  there  may  be  other  types  of  errors  in  your  data that  are  harder  to  find.

***1.2.1:  Check  that  your  data  are  sensible***

Just  because  your  dataset  is  clean,  it  doesn’t  mean  that  it  is  correct  –  real  life  follows  rules,  and  your

data  must  do  too.  There  are  limits  on  the  heights  of  participants  in  your  study,  so  check  that  all  data

fit  within  reasonable  limits.  Calculate  the  minimum,  maximum  and  mean  values  of  variables  to  see  if

all  values  are  sensible.

Sometimes,  putting  together  two  or  more  pieces  of  data  can  reveal  errors  that  can  otherwise  be

difficult  to  detect.  Does  the  difference  between  date  of  birth  and  date  of  diagnosis  give  you  a

negative  number?  Is  your  patient  over  300  years  old?

***1.2.2:  Check  that  your  variables  are  sensible***

Once  you  have  a  perfectly  clean  dataset  it  is  relatively  easy  to  compare  variables  with  each  other  to find  out  if  there  is  a  relationship  between  them  (the  subject  of  this  book).  But  just  because  you  can,

it  doesn’t  mean  that  you  should.  If  there  is  no  good  reason  why  there  should  be  a  relationship

between  sales  of  ice  cream  and  haemorrhoid  cream  then  you  should  consider  expelling  one  or  both

from  the  dataset.  If  you’ve  collected  your  own  data  from  original  sources  then  you’ll  have

considered  beforehand  which  data  are  sensible  to  collect  (you  have,  haven’t  you?!??),  but  if  your

dataset  is  a  pastiche  of  two  or  more  datasets  then  you  might  find  strange  combinations  of  variables.

You  should  check  your  variables  before  doing  any  analyses  and  consider  whether  it  is  sensible  to

make  these  comparisons.

So  now  you  have  collected,  cleaned  your  data  and  checked  that  your  data  is  sensible  and  fit  for

purpose.  In  the  next  chapter,  we’ll  go  through  the  basics  of  data  classification  and  introduce  the  four types  of  data.

7

**Chapter  2:  Data  Classification**

In  this  chapter,  you’ll  learn  the  difference  between  quantitative  and  qualitative  data.  You’ll  also  learn

about  ratio,  interval,  ordinal  and  nominal  data  types,  and  what  operations  you  can  perform  on  each

of  them.

**2.1:  Quantitative  and  Qualitative  Data**

All  data  is  either  quantitative;  measured  with  some  kind  of  measuring  implement  –  ruler,  jug,

weighing  scales,  stop-watch,  thermometer  and  so  on,  or  is  qualitative;  an  observed  feature  of

interest  that  is  placed  into  categories  –  gender  (male,  female),  health  (healthy,  sick),  opinion  (agree,

neutral,  disagree).

Quantitative  and  qualitative  data  can  be  sub-divided  into  four  further  classes  of  data:

Quantitative  (measured)

o      Ratio

o      Interval

Qualitative  (categorised)

o Ordinal

o Nominal

The  difference  between  them  can  be  established  by  asking  just  three  questions:







Ordered

o      Can  some  sort  of  progress  be  detected  between  adjacent  data  points  or  categories

or  can  the  data  be  ordered  meaningfully?

Equidistant

o      Is  the  distance  between  adjacent  data  points  or  categories  consistent?

Meaningful  Zero

o Does  the  scale  of  measurement  include  a  unique,  non-arbitrary  zero  value?

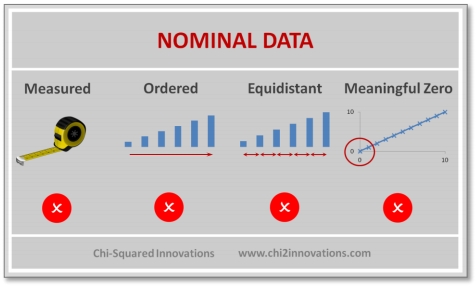
Let’s  go  through  each  of  the  four  classes  to  see  how  they  fit  with  these  questions.

***2.1.1:  Nominal  data***

Nominal  Data  is  observed,  not  measured,  is  unordered,  non-equidistant  and  has  no  meaningful  zero.

We  can  differentiate  between  categories  based  only  on  their  names,  hence  the  title  ‘nominal’  (from

the  Latin*nomen*,  meaning  ‘name’).  Figure  2.1  may  help  you  decide  if  your  data  is  nominal.



**Figure  2.1:  Features  of  Nominal  Data**

Examples  of  nominal  data  include:













Gender  (male,  female)

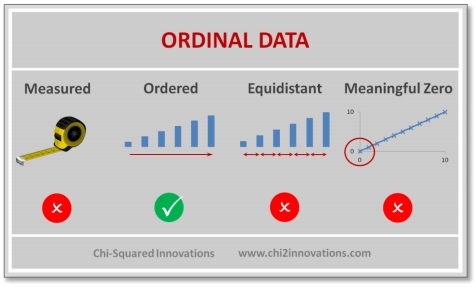
Nationality  (British,  American,  Spanish,...)

Genre/Style  (Rock,  Hip-Hop,  Jazz,  Classical,...)

Favourite  colour  (red,  green,  blue,...)

Favourite  animal  (aardvark,  koala,  sloth,...)

Favourite  spelling  of  ‘favourite’  (favourite,  favorite)



The  only  mathematical  or  logical  operations  we  can  perform  on  nominal  data  is  to  say  that  an

observation  is  (or  is  not)  the  same  as  another  (equality  or  inequality),  and  we  can  determine  the

most  common  item  by  finding  the  mode  (do  you  remember  this  from  High  School  classes?).

Other  ways  of  finding  the  middle  of  the  class,  such  as  median  or  mean  make  no  sense  because

ranking  is  meaningless  for  nominal  data.

If  the  categories  are  descriptive  (Nominal),  like  ‘Pig’,  ‘Sheep’  or  ‘Goat’,  it  can  be  useful  to  separate

each  category  into  its  own  column,  such  as  Pig  [Yes;  No],  Sheep  [Yes;  No],  and  Goat  [Yes;  No].  These

are  called  ‘dummy’  variables  and  they  can  be  very  useful  analytically.  More  on  that  later…

***2.1.2:  Ordinal  data***

Ordinal  Data  is  observed,  not  measured,  is  ordered  but  non-equidistant  and  has  no  meaningful  zero.

Their  categories  can  be  ordered  (1st,  2nd,  3rd,  etc.  –  hence  the  name  ‘ordinal’),  but  there  is  no

consistency  in  the  relative  distances  between  adjacent  categories.  Figure  2.2  shows  the  features  of

ordinal  data.

**Figure  2.2:  Features  of  Ordinal  Data**

Examples  of  ordinal  data  include:











Health  (healthy,  sick)

Opinion  (agree,  mostly  agree,  neutral,  mostly  disagree,  disagree)

Tumour  Grade  (1,  2,  3)

Tumour  Stage  (I,  IIA,  IIB,  IIIA,  IIIB,  etc.)

Time  of  day  (morning,  noon,  night)

Mathematically,  we  can  make  simple  comparisons  between  the  categories,  such  as  more  (or  less)

healthy/severe,  agree  more  or  less,  etc.,  and  since  there  is  an  order  to  the  data,  we  can  rank  them

and  compute  the  median  (or  mode,  but  not  the  mean)  to  find  the  central  value.

It  is  interesting  to  note  that  in  practice  some  ordinal  data  are  treated  as  interval  data  –  Tumour

Grade  is  a  classic  example  in  healthcare  –  because  the  statistical  tests  that  can  be  used  on  interval

data  (they  meet  the  requirement  of  equal  intervals)  are  much  more  powerful  than  those  used  on

ordinal  data.  This  is  OK  as  long  as  your  data  collection  methods  ensure  that  the  equidistant  rule  isn't

bent  too  much.

***2.1.3:  Interval  data***

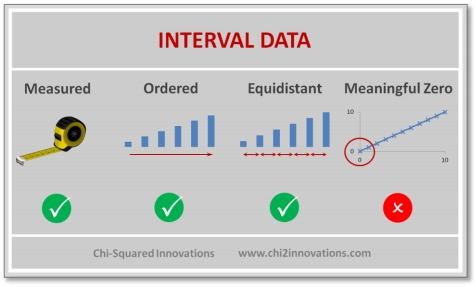
Interval  Data  is  measured  and  ordered  with  equidistant  items,  but  has  no  meaningful  zero.  Interval

data  can  be  continuous  (have  an  infinite  number  of  steps)  or  discrete  (organised  into  categories),

and  the  degree  of  difference  between  items  is  meaningful  (their  intervals  are  equal),  but  not  their

ratio.  Figure  2.3  will  help  in  identifying  interval  data.

10



**Figure  2.3:  Features  of  Interval  Data**

Examples  of  interval  data  include:







Temperature  (°C  or  F,  but  not  Kelvin)

Dates  (1066,  1492,  1776,  etc.)

Time  interval  on  a  12  hour  clock  (6am,  6pm)

Although  interval  data  can  appear  very  similar  to  ratio  data,  the  difference  is  in  their  defined  zero-

points.  If  the  zero-point  of  the  scale  has  been  chosen  arbitrarily  (such  as  the  melting  point  of  water

or  from  an  arbitrary  epoch  such  as  AD)  then  the  data  cannot  be  on  the  ratio  scale  and  must  be

interval.

Mathematically  we  may  compare  the  degrees  of  the  data  (equality/inequality,  more/less)  and  we

may  add/subtract  the  values,  such  as  ‘20°C  is  10  degrees  hotter  than  10°C’  or  ‘6pm  is  3  hours  later

than  3pm’.  However,  we  cannot  multiply  or  divide  the  numbers  because  of  the  arbitrary  zero,  so  we

can’t  say  ‘20°C  is  twice  as  hot  as  10°C’  or  ‘6pm  is  twice  as  late  as  3pm’.

The  central  value  of  interval  data  is  typically  the  mean  (but  could  be  the  median  or  mode),  and  we

can  also  express  the  spread  or  variability  of  the  data  using  measures  such  as  the  range,  standard

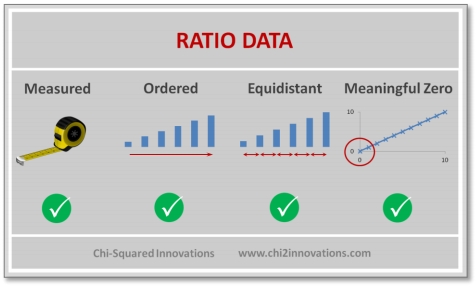
deviation,  variance  and/or  confidence  intervals.

***2.1.4:  Ratio  data***

Ratio  Data  is  measured  and  ordered  with  equidistant  items  and  a  meaningful  zero.  As  with  interval

data,  ratio  data  can  be  continuous  or  discrete,  and  differs  from  interval  data  in  that  there  is  a  non-

arbitrary  zero-point  to  the  data.  The  features  of  ratio  data  are  shown  in  Figure  2.4.



**Figure  2.4:  Features  of  Ratio**

Examples  include:









Age  (from  0  years  to  100+)

Temperature  (in  Kelvin,  but  not  °C  or  F)

Distance  (measured  with  a  ruler  or  other  such  measuring  device)

Time  interval  (measured  with  a  stop-watch  or  similar)

For  each  of  these  examples  there  is  a  real,  meaningful  zero-point  –  the  age  of  a  person  (a  12  year  old

is  twice  the  age  of  a  6  year  old),  absolute  zero  (matter  at  200K  has  twice  the  energy  of  matter  at

100K),  distance  measured  from  a  pre-determined  point  (the  distance  from  Barcelona  to  Berlin  is  half

the  distance  as  Barcelona  to  Moscow)  or  time  (it  takes  me  twice  as  long  to  run  the  100m  as  Usain

Bolt  but  only  half  the  time  of  my  Grandad).

Ratio  data  are  the  best  to  deal  with  mathematically  (note  that  I  didn’t  say*easiest*...)  because  all

possibilities  are  on  the  table.  We  can  find  the  central  point  of  the  data  by  using  any  of  the  mode,

median  or  mean  (arithmetic,  geometric  or  harmonic)  and  use  all  of  the  most  powerful  statistical

methods  to  analyse  the  data.  As  long  as  we  choose  correctly,  we  can  be  really  confident  that  we  are

not  being  misled  by  the  data  and  our  interpretations  are  likely  to  have  merit.

So  this  concludes  the  first  section  of  this  book.

So  far,  you  have  collected,  cleaned,  pre-processed  and  classified  your  data.  It’s  a  good  start,  but

there’s  still  a  lot  more  to  do.  The  next  section  is  all  about  building  a  statistical  toolbox  that  you  can

use  to  extract  real,  meaningful  insights  from  your  shiny  new  dataset.

12

**Section   2:**



**Your   Statistical   Toolbox**

14

**Chapter  3:  Introduction  to  Associations  and  Correlations**

In  this  chapter,  you’ll  start  to  understand  a  little  about  correlations  and  associations,  what  they  are

and  how  you  use  them.  You’ll  also  learn  the  basics  of  how  to  build  a  holistic  strategy  to  discover  the

story  of  your  data  using  univariate  and  multivariate  statistics  and  find  out  why  eating  ice  cream  does

not  make  you  more  likely  to  drown  when  you  go  swimming.

You’ll  learn  about  why  statistics  can*never*  prove  that  there  is  a  relationship  between  a  pair  of

variables,  via  the  null  hypothesis,  p-values  and  a  sensibly  chosen  cut-off  value,  and  learn  about  how

to  measure  the  strength  of  relationships.

**3.1:  What  are  Associations  and  Correlations?**

When  you  can  phrase  your  hypothesis  (question  or  hunch)  in  the  form  of:







Is  smoking*related  to*  lung  cancer?

Is  there  an*association*  between  diabetes  and  heart  disease?

Are  height  and  weight*correlated*?

then  you  are  talking  about  the  relationship  family  of  statistical  analyses.

Typically,  the  terms  correlation,  association  and  relationship  are  used  interchangeably  by

researchers  to  mean  the  same  thing.  That’s  absolutely  fine,  but  when  you  talk  to  a  statistician  you

need  to  listen  carefully  –  when  she  says  correlation,  she  is  most  probably  talking  about  a  statistical

correlation  test,  such  as  a  Pearson  correlation.

There  are  distinct  stats  tests  for  correlations  and  for  associations,  but  ultimately  they  are  all  testing

for  the  likelihood  of  relationships  in  the  data.

When  you  are  looking  for  a  relationship  between  2  continuous  variables,  such  as  height  and  weight,

then  the  test  you  use  is  called  a  correlation.  If  one  or  both  of  the  variables  are  categorical,  such  as

smoking  status  (never,  rarely,  sometimes,  often  or  very  often)  or  lung  cancer  status  (yes  or  no),  then

the  test  is  called  an  association.

In  terms  of  correlations,  if  there  is  a  correlation  between  one  variable  and  another,  what  that  means

is  that  if  one  of  your  variables  changes,  the  other  is  likely  to  change  too.

For  example,  say  you  wanted  to  find  out  if  there  was  a  relationship  between  age  and  percentage  of

body  fat.  You  would  do  a  scatter  plot  of  age  against  body  fat  percentage,  and  see  if  the  line  of  best

fit  is  horizontal.  If  it  is  not,  then  we  can  say  there  is  a  correlation.  In  this  example,  there  is  a  positive

correlation  between  age  and  percentage  of  body  fat;  that  is,  as  you  get  older  you  are  likely  to  gain

increasing  amounts  of  body  fat,  like  that  shown  in  Figure  3.1.

15

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | | |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |  |  |  |  |  |  |  |

**Age (Years)**

**BodyFat(%)**

45

40

35

30

25

20

15

10

5

0

0

10

20

30

40

50

60

70

**Figure  3.1:  Positive  Correlation  with  Line  of  Best  Fit**

For  associations  it  is  all  about  the  measurement  or  counts  of  variables  within  categories.

Let’s  have  a  look  at  an  example  from  Charles  Darwin.  In  1876,  he  studied  the  growth  of  corn

seedlings.  In  one  group,  he  had  15  seedlings  that  were  cross-pollinated,  and  in  the  other  15  that

were  self-pollinated.  After  a  fixed  period  of  time,  he  recorded  the  final  height  of  each  of  the  plants

to  the  nearest  1/8th  of  an  inch.

To  analyse  these  data  you  pool  together  the  heights  of  all  those  plants  that  were  cross-pollinated,

work  out  the  average  height  and  compare  this  with  the  average  height  of  those  that  were  self-

pollinated.  A  statistical  test  can  tell  you  whether  there  is  a  difference  in  the  heights  between  the

cross-pollinated  and  self-pollinated  groups.  This  is  an  example  of  an  association,  and  is  shown  as  a

histogram  in  Figure  3.2.

16

|  |  |  |
| --- | --- | --- |
|  |  | |
|  |
|  |
|  |
|  |
|  |
|  |  |  |

**Method of Pollination**

**Height(inches)**

22

20

18

16

14

12

10

Cross

Self

**Figure  3.2:  Association  of  Heights  Across  Two  Groups**

If  you  don’t  understand  it  yet,  don’t  worry  –  we’ll  talk  more  about  associations  and  correlations  in

the  following  chapters.

**3.2:  Discovering  the  Story  of  Your  Data**

Let’s  say  you  have  your  hypothesis,  you’ve  designed  and  run  your  experiment,  and  you’ve  collected

your  data.

If  you’re  a  medic,  an  engineer,  a  businesswoman  or  a  marketer  then  here’s  where  you  start  to  get

into  difficulty.  Most  people  who  have  to  analyse  data  have  little  or  no  training  in  data  analysis  or

statistics,  and  round  about  now  is  where  the  panic  starts  to  begin.

What  you  need  to  do  is:

1. Take  a  deep  breath

2. Define  a  strategy  of  how  to  get  from  data  to  story

3. Learn  the  basic  tools  of  how  to  implement  that  strategy

17

In  relationship  analyses,  the  strategy  is  a  fairly  straightforward  one  and  goes  like  this:







Clean  &  classify  your  data

o      You  now  have  a  clean  dataset  ready  to  analyse

Answer  your  primary  hypothesis

o *This*  is  related  to*that*

Find  out  what  else  has  an  effect  on  the  primary  hypothesis

o Rarely  is*this*  related  to*that*  without  something  else  having  an  influence

Not  too  difficult  so  far  is  it?

OK,  now  the  slightly  harder  bit  –  you  need  to  learn  the  tools  to  implement  the  strategy.  This  means

understanding  the  statistical  tests  that  you’ll  use  in  answering  your  hypothesis.

Before  we  go  into  depth,  we  need  to  generalise  a  little  bit.

There  are  basically  2  types  of  test  you’ll  use  here:





Univariate  statistics  (Chapter  4)

Multivariate  statistics  (Chapter  5)

They  sound  quite  imposing,  but  in  a  nutshell,  univariate  stats  are  tests  that  you  use  when  you  are

comparing  variables  one  at  a  time  with  your  hypothesis  variable.  In  other  words,  you  compare*this*

with*that*  whilst  ignoring  all  other  potential  influences.

On  the  other  hand,  you  use  multivariate  stats  when  you  want  to  measure  the  relative  influence  of

many  variables  on  your  hypothesis  variable,  such  as  when  you  simultaneously  compare*this*,*that*

and*the  other*  against  your  target.

It  is  important  that  you  use  both  of  these  types  of  test  because  although  univariate  stats  are  easier

to  use  and  they  give  you  a  good*feel*  for  your  data,  they  don’t  take  into  account  the  influence  of

other  variables  so  you  only  get  a  partial  –  and  probably  misleading  –  picture  of  the  story  of  your

data.

On  the  other  hand,  multivariate  stats*do*  take  into  account  the  relative  influence  of  other  variables,

but  these  tests  are  much  harder  to  implement  and  understand  and  you  don’t  get  a  good  feel  for

your  data.

18

My  advice  is  to  do  univariate  analyses  on  your  data  first  to  get  a  good  understanding  of  the

underlying  patterns  of  your  data,  then  confirm  or  deny  these  patterns  with  the  more  powerful

multivariate  analyses.  This  way  you  get  the  best  of  both  worlds  and  when  you  discover  a  new

relationship,  you  can  have  confidence  in  it  because  it  has  been  discovered  and  confirmed  by  two

different  statistical  analyses.

When  pressed  for  time  I’ve  often  just  jumped  straight  into  the  multivariate  analysis.  Whenever  I’ve

done  this  it  always  ends  up  costing  me  more  time  –  I  find  that  some  of  the  results  don’t  make  sense

and  I  have  to  go  back  to  the  beginning  and  do  the  univariate  analyses  before  repeating  the

multivariate  analyses.

I  advise  that  you  think  like  the  tortoise  rather  than  the  hare  –  slow  and  methodical  wins  the  race…

Finally,  when  you  have  the  story  of  your  data  you’ll  need  to  explain  it  to  others,  whether  in  person  or

in  a  publication  of  some  sort  (like  a  report,  thesis  or  white  paper).

If  your  story  is  a  complicated  one  then  it  might  be  useful  to  produce  some  kind  of  visualisation  to

make  it  easier  for  your  audience,  and  some  visualisations  are  more  appropriate  and  easier  to

understand  than  others  are.  We’ll  explore  this  in  Chapter  6.

**3.3:  Correlation  v  Causation**

And  now  a  word  of  warning  that  I’m  sure  you’ve  heard  many  times  before.

Just  because  you’ve  found  a  correlation  (or  association)  between  a  pair  of  variables,  it  does  not

necessarily  mean  that  one  causes  the  other.

For  example,  let’s  say  that  you  discovered  in  your  analysis  that  there  is  a  relationship  between  ice

cream  sales  and  incidences  of  drowning.  What  can  you  infer  from  this?





Ice  cream  sales  cause  people  to  drown

o      Surely  not,  unless  you  have  to  swim  a  mile  out  to  sea  to  reach  the  ice  cream

vendor’s  boat

Drowned  people  eat  ice  cream  in  the  after-life

o      Well,  without  a  thorough  command  of  how  to  travel  to  the  after-life  and  back  again,

you  will  probably  never  know!

19

Most  likely,  there  is  a  variable  that  is  related  to  both  ice  cream  sales  and  the  likelihood  of  drowning.

It  is  easy  to  see  in  this  case  that  the  weather  is  likely  to  play  a  part  –  more  people  will  eat  ice  cream

and  go  for  a  swim  to  cool  off  in  hot  weather  than  in  cold.

Let’s  not  beat  around  the  bush  here  though.  If  you  find  a  relationship  between  a  pair  of  variables

and  you’ve  confirmed  it  with  multivariate  analysis,  then*something*  is  responsible  for  the  relationship

and  there*is*  a  cause.  It’s  your  job  to  find  it.







It  may  be  that  a  variable  in  your  dataset  is  the  cause

o      Then  go  and  find  it  –  the  strategies  detailed  in  this  book  will  help

There  might  be  an  influence  from  something  that  is  not  represented  in  your  dataset

o      Then  you  need  a  new  hypothesis  (and  new  data)  to  find  it

The  result  may  have  occurred  by  chance

o Then  you  need  to  repeat  your  experiment  to  confirm  or  deny  the  result

**3.4:  Assessing  Relationships**

When  you  first  made  your  hypothesis,  you  will  have  asked  yourself  the  following  question:



Is  there  a  relationship  between*this*  and*that*?

o The  answer  is  absolute  –  it  is  either  yes  or  no

This  is  the  question  that  statistical  tests  attempt  to  answer.  Unfortunately,  they  are  not  equipped  to

*definitively*  answer  the  question,  but  rather  they  tell  you  how  confident  you  can  be  about  the  answer

to  your  question.

The  strange  thing  about  statistics  is  that  it  cannot  tell  you  whether  a  relationship  is  true,  it  can  only

tell  you  how  likely  it  is  that  the  relationship  is  false.  Weird,  huh?

What  you  do  is  make  an  assumption  that  there*is  not*  a  relationship  between*this*  and*that*  (we  call

this  the  null  hypothesis).  Then  you  test  this  assumption.

How  do  we  do  this?

All  relationship  tests  give  you  a  number  between  0  and  1  that  you  can  use  to  tell  you  how  confident

you  are  in  the  null  hypothesis.  This  number  is  called  the  p-value,  and  is  the  probability  that  the  null

hypothesis  is  correct.

20

If  the  p-value  is  large,  there  is  strong  evidence  that  the  null  hypothesis  is  correct  and  we  can  be

confident  that  there  is  not  a  relationship  between  the  variables.

On  the  other  hand,  if  the  p-value  is  small,  there  is  weak  evidence  to  support  the  null  hypothesis.  You

can  then  be  confident  in*rejecting  the  null  hypothesis*  and  concluding  that  there  is  evidence  that  a

relationship  may  exist.

OK,  so  if  you  have  a  number  between  0  and  1,  how  does  this  help  you  figure  out  if  there  is  a

relationship  between  your  variables?

Well,  it  depends…

Ah,  you  wanted  a  definitive  answer  didn’t  you?  Well,  you’re  not  going  to  get  one!

Typically,  a  p-value  of  0.05  is  used  as  the  cut-off  value  between  ‘the  null  hypothesis  is  correct’

(anything  larger  than  0.05)  and  ‘not  correct’  (smaller  than  0.05).

If  your  test  p-value  is  0.15,  this  corresponds  to  a  probability  of  15%  that  the  null  hypothesis  is

correct,  we  do  not  reject  the  null  hypothesis  (since  0.15  is  larger  than  0.05)  and  we  say  that  there  is

insufficient  evidence  to  conclude  that  there  is  a  relationship  between*this*  and*that*.

Where  did  the  p-value  of  0.05  come  from?

This  cut-off  value  has  become  synonymous  with**Ronald  A  Fisher**  (the  guy  who  created  the  Fisher’s

Exact  Test,  among  other  things),  who  said  that:

*…it  is  convenient  to  draw  the  line  at  about  the  level  at  which  we  can  say:*

*“Either  there  is  something  in  the  treatment,  or  a  coincidence  has  occurred  such*

*as  does  not  occur  more  than  once  in  twenty  trials”…*

As  1  in  20,  expressed  as  a  decimal,  is  0.05  this  is  the  arbitrarily  chosen  cut-off  point.  Rather  than

calculated  mathematically,  it  was  chosen  simply  to  be  sensible.

21

Since  0.05  is  ‘sensible’,  then  you  should  also  be  sensible  about  test  p-values  that  are  close  to  this

value.

You  should  not  definitively  reject  the  null  hypothesis  when

any  more  than  you  should

definitively  accept  the  null  hypothesis  when .

When  your  test  p-value  is  close  to  0.05,  you  would  be  sensible  to  conclude  that  such  p-values  are

worthy  of  additional  investigation.

For  the  purposes  of  clarity,  from  this  point  on  any  p-value  smaller  than  0.05  we  will  say  things  like

‘there  is  (evidence  for)  a  relationship’,  ‘statistically  significant’  or  something  similar.  Conversely,  p-

values  larger  than  0.05  will  attract  statements  such  as  ‘there  is  not  a  relationship’  or  ‘not  statistically

significant’.

**3.5:  Assessing  the  Strength  of  Relationships**

OK,  so  you  have  tested  whether  the  relationship  is  likely  to  be  true  and  concluded  –  based  on  the  p-

value  –  that  it  is.

Now  you  need  to  test  how  strongly  related  the  variables  are.  We  call  this  the  effect  size,  and  typical

measures  of  the  size  of  the  effect  observed  are:







Odds  Ratio

Risk  Ratio

Hazard  Ratio

These  are  all  positive  values  and  tell  you  the  size  of  the  difference  between  two  groups.

For  example,  say  that  we  have  a  group  of  100  men  who  smoke,  and  that  27  of  them  had  lung  cancer.

The  odds  of  getting  lung  cancer  if  you’re  a  man  are  then

.

Now  let’s  look  at  the  group  of  213  women  who  smoke.  Say  that  57  of  those  had  lung  cancer.  The

odds  of  getting  lung  cancer  in  women  are  then

.

The  ratio  of  these  odds  represents  the  different  risks  of  contracting  lung  cancer  between  the  gender

groups,  which  is

.

22

An  effect  size  close  to  one  indicates  that  there  is  little  or  no  difference  between  the  groups  (and

more  than  likely  the  p-value  would  be  large  and  non-significant),  i.e.  the  odds  are  the  same  in  both

groups.

What  if  the  odds  ratio  is  2?  This  would  tell  you  that  male  smokers  are  twice  as  likely  to  contract  lung

cancer  as  female  smokers  are.

You  can  also  flip  this  odds  ratio  on  its  head  and  say  that  female  smokers  are  half  as  likely  to  contract

lung  cancer  as  male  smokers.

In  the  next  chapter,  we’ll  go  through  the  various  kinds  of  univariate  statistics  that  you’ll  find  in

relationship  analyses.

●

●

●

You   are   free   to   share   this   book

with   anyone,   as   long   as   you   don’t

change   it   or   charge   for   it

●

●

●

23

**Chapter  4:  Univariate  Statistics**

In  this  chapter,  you  will  learn  about  the  different  univariate  statistical  tests  that  you  will  use  in

relationship  analysis.  You  will  learn  about  the  different  correlation,  association  and  survival  analysis

tests  in  univariate  analysis  and  learn  what  to  look  for  in  a  typical  set  of  results  to  decide  whether

there  is  evidence  of  a  relationship  between  the  variables  and  how  strong  is  that  relationship.

**4.1:  Correlations**

When  both  of  the  variables  under  investigation  are  continuous,  you  ask  the  question  ‘is*this*

correlated  with*that*?’.

For  example,  let’s  have  another  look  at  Figure  3.1,  where  we  had  a  look  at  the  relationship  between

age  and  percentage  of  body  fat.  Figure  3.1  is  called  a  scatter  plot,  and  the  best-fit  line  between  them

is  called  a  line  of  regression.

This  is  an  important  graph  because  it  gives  you  a  good  feel  for  your  data  and  what  the  underlying

story  might  be.  When  you  see  a  scatter  plot  such  as  this,  you  should  always  ask  the  following

questions:











Is  there  a  gradient  to  the  regression  line?

o      Is  the  gradient  positive  (slopes  up  from  left-to-right)?

o      Or  negative  (slopes  down  from  left-to-right)?

Do  the  plot  points  fall  close  to  or  far  from  the  regression  line?

Is  the  linear  regression  line  a  good  representation  of  the  data?

o      Would  a  curve  or  other  line  shape  be  a  better  fit?

Are  there  clearly  defined  clusters  of  data?

Are  there  any  outliers?

Now  that  you  have  a  good  sense  of  what’s  going  on  with  the  data  you  can  test  whether  there  is  a

correlation  between  age  and  body  fat.

There  are  2  types  of  correlation  test;  the  Pearson  and  Spearman  correlations.  Figure  4.1  will  help  you

decide  which  is  most  appropriate  for  your  data.

24

|  |  |
| --- | --- |
| **Pearson  Correlation** | **Spearman  Correlation** |
| Assumes  that  both  variables  are  normally  distributed | Makes  no  assumption  about  the  data  distributions |
| Use  this  when  there  is  a  linear  relationship  between  the  variables | Use  this  when  there  is  a  non-linear  relationship  between  the  variables |
| If  the  assumptions  are  met,  Pearson’s  correlation  is  more  powerful  than  Spearman’s | When  the  normality  assumptions  are  not  met,  Spearman’s  correlation  is  more  appropriate |

**Figure  4.1:  Choice  of  Pearson  and  Spearman  Correlation  Tests**

If  a  straight  line  is  not  the  best  fit  for  the  data  then  there  are  methods  that  you  can  try  to  get  a

better  fit,  such  as  data  transformations.  In  this  case,  non-linear  data  may  be  assessed  by  the  Pearson

correlation  rather  than  the  Spearman  but  these  methods  are  beyond  the  scope  of  this  book.

Whether  you  use  the  Pearson  or  Spearman  correlation  test,  the  important  outputs  that  you  will

need  when  you  run  these  tests  on  your  data  are:







**Visualisation**

o **scatter  plot  with  best-fit  regression  line**

**Statistic**

o **p-value**

o **correlation  coefficient*R*  (or*R2*)**

**Effect  Size**

o **equation  of  the  line  (gradient  &  intercept)**

**The  scatter  plot**  is  important  because  it  tells  you  if  there  is  likely  to  be  evidence  of  a  relationship

between  the  pair  of  variables,  and  which  type  of  correlation  test  is  likely  to  be  the  better  choice.

**The  p-value**  is  important  because  it  tells  you  if  there  is  evidence  of  a  relationship  between  the  pair

of  variables.

If  the  p-value  is  not  significant  then  the  correlation  coefficient  and  equation  of  the  line  are

meaningless  –  if  you  have  concluded  that  there  is  not  a  relationship  between  the  variables,  then

measuring  the  strength  of  the  relationship  makes  no  sense.

**The  correlation  coefficient**  gives  you  a  measure  of  how  well  the  plot  points  fit  the  regression  line.

If  the  correlation  coefficient*R*  is  large,  then  the  points  fit  the  line  very  well.

25

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | **Coef** | **SE  Coef** | **T** | **P** |

|  |  |  |
| --- | --- | --- |
| **S** | **R2** | **2**  **R  (adj)** |

Let’s  say  that .  You  can  say  that  93%  (

)  of  the  variation  in  outcome  is

explained  by  the  predictor  variable.  On  the  other  hand,  if ,  then  only  32%  (

)  of  the  variation  is  due  to  the  predictor  and  you  need  to  look  at  other  variables  to  try  to  find

alternative  explanations  for  the  variation  in  outcome.

**The  equation  of  the  line,**

gives  you  a  way  to

predict  one  of  the  variables  if  you  know  the  value  of  the  other.

The  gradient  is  useful  as  it  tells  you  by  how  much  the  outcome  should  increase  for  a  single  unit

change  in  predictor,  and  is  a  measure  of  effect  size.

OK,  let’s  run  a  Pearson  correlation  of  the  age  and  body  fat  percentage  data.  The  results  are  shown  in

Figure  4.2.

Regression equation:

**Body  Fat  =  3.221  +  (0.548  x  Age)**

Summary Table:

**Age**

**Constant**

0.5480

3.2210

0.1056

5.0760

5.19

0.63

<0.001

0.535

Regression Statistics:

**Figure  4.2:  Typical  Results  of  a  Pearson  Correlation**

5.75361

62.7%

60.4%

The  Age  p-value  tell  you  that  there  is  a  significant  correlation  between  age  and  body  fat  (p<0.001)

and  its  coefficient  tells  you  that  the  correlation  is  positive  –  but  of  course  you  knew  that  already

from  the  scatter  plot,  didn’t  you?

The*R2*  value  tells  you  that  62.7%  of  the  variation  in  percentage  of  body  fat  is  explained  by  age  (the

adjusted*R2*  value  is  only  useful  when  assessing  multiple  predictor  variables  and  can  be  ignored  in

univariate  analysis).

26

Finally,  the  equation  of  the  line  tells  you  that  on  average,  body  fat  increases  by  an  additional  0.548%

for  every  single  unit  increase  in  age  (i.e.  for  every  year).

**4.2:  Associations**

While  correlations  concern  only  continuous  data  like  height,  weight,  age,  etc.,  associations  concern

categorical  data  such  as  gender  [male;  female],  status  of  some  sort  [small;  medium;  large]  and  even

continuous  data  that  has  been  arranged  in  categories,  such  as  height  [short;  average;  tall]  or  age

[30s;  40s;  50s;  etc.].

There  are  2  types  of  association  that  you  need  to  be  aware  of:





categorical*versus*  categorical

categorical*versus*  continuous

To  confuse  things  just  a  little  bit  more,  there  are  2  flavours  of  categorical  data:





ordinal

nominal

Recall  from  Chapter  2  that  ordinal  data  is  categorical  data  that  has  a  natural  order  to  the  categories

and  that  the  distance  between  the  categories  is  approximately  equal.  Nominal  data  categories  have

no  natural  order,  so  it  may  be  useful  to  separate  them  into  separate  variables,  each  having  2

categories.

It  is  important  to  remember  this  at  this  stage  because  you  will  deal  here  with  data  that  has  more

than  2  categories.  Such  data*must*  be  ordinal,  otherwise  the  results  you  get  will  be  misleading  at

best,  and  just  plain  wrong  at  worst.

For  data  with  just  2  categories,  it  doesn’t  matter  whether  they  are  nominal  or  ordinal,  their

treatment  and  analysis  is  just  the  same.

***4.2.1:  mxn  categorical***

When  you  are  comparing  2  variables  that  each  have  multiple  categories,  you  can  call  this  an  mxn  (‘m

by  n’)  categorical  analysis.  It’s  also  sometimes  known  as  an  rxc  categorical  analysis.

Let’s  have  a  look  at  a  hypothetical  example.  Say  that  you’re  looking  into  the  possibility  that  the

aggressiveness  of  breast  cancer  is  associated  with  the  number  of  lymph  nodes  to  which  the  cancer

has  spread.  You  have  three  grades  of  cancer,  from  Low  (least  aggressive)  to  High  (most  aggressive)

and  your  lymph  node  data  has  been  categorised  as  0,  1-3  and  3+  lymph  nodes  involved.  You

27

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Number  of  Involved**  **Lymph  Nodes** | | |
| **0**  **Nodes** | **1-3**  **Nodes** | **3+**  **Nodes** |
| **Tumour**  **Aggressiveness** | **Low** | **165** | **33** | **2** |
| **Medium** | **457** | **122** | **51** |
| **High** | **298** | **174** | **103** |

hypothesise  that  the  more  aggressive  the  tumour,  the  more  lymph  nodes  that  the  cancer  will  have

spread  to.

To  visualise  what  these  data  look  like  we  create  a  3x3  grid  and  put  each  of  the  patients  into  the

appropriate  cell.  There  are  many  alternative  names  for  such  a  grid,  including  frequency  table,

contingency  table  and  confusion  matrix,  and  looks  like  Figure  4.3.

**Figure  4.3:  3x3  Contingency  Table**

Although  it  looks  like  an  important  table,  it’s  actually  difficult  to  ‘read’  and  understand  if  there  is

likely  to  be  a  relationship  between  the  2  variables.

What  you  need  is  a  way  to  calculate  the  expected  value  for  each  cell,  then  you  can  figure  out  by  how

much  your  actual  values  deviate  from  what  you  would  expect  by  chance.

You  do  this  by  using  these  equations:

and

When  you  calculate

,  some  of  the  cell  values  will  be  negative.  Make  a  note

of  which  ones,  and  when  you  put  the  results  into  the  table,  add  back  in  the  negative  sign  to  the

appropriate  cells.

28

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Number  of  Involved**  **Lymph  Nodes** | | |
| **0**  **Nodes** | **1-3**  **Nodes** | **3+**  **Nodes** |
| **Tumour**  **Aggressiveness** | **Low** | **8.8** | **-4.1** | **-18.4** |
| **Medium** | **4.8** | **-4.4** | **-5.1** |
| **High** | **-16.4** | **11.5** | **24.0** |

This  will  give  us  what  I  call  the  Chi-Squared  cell  values  ( ),  and  tells  you  whether  and  by  how  much

the  observed  values  in  each  cell  differ  from  what  would  be  expected  by  chance.

A  more  in-depth  explanation  of  Chi-Squared  analysis  can  be  found  in  our[**Discover Stats Blog Series.**](http://chi2innovations.com/blog/discover-stats-blog-series/)

Your  table  of  Chi-Squared  cell  values  will  now  look  like  Figure  4.4.

**Figure  4.4:  3x3  Contingency  Table  of  Chi-Squared  Cell  Values**

This  table  will  now  help  you  to  get  a  good  feel  for  your  data,  and  you  should  always  ask  the  following

questions:









Which  cells  are  positive  (have  more  involved  lymph  nodes  than  expected  by  chance)?

Which  cells  are  negative  (have  fewer  involved  lymph  nodes  than  expected  by  chance)?

Are  there  any  cell  values  that  are  larger  than  the  rest?

o      This  is  the  critical  cell  and  is  the  cell  that  differs  most  from  what  we  would  expect  by

chance

Are  these  results  what  you  expected  to  find?

For  these  data,  you  can  see  that  the  largest  Chi-Squared  cell  value  corresponds  to  the  most

aggressive  tumours  and  highest  number  of  involved  lymph  nodes  and  is  positive.  This  tells  you  that

there  are  more  involved  lymph  nodes  in  the  most  aggressive  tumour  category  than  expected  by

chance.  In  fact,  the  adjacent  cells  suggests  that  for  the  most  aggressive  tumours  it  is  more  likely  than

expected  by  chance  that  the  cancer  will  spread  to  the  lymph  nodes  (cell  values  of  11.5  and  24.0)  and

less  likely  that  the  cancer  will  not  spread  to  the  lymph  nodes  (cell  value  of  -16.4).

29

The  largest  negative  Chi-Squared  cell  value  (-18.4)  suggests  that  the  least  aggressive  tumours  are

less  likely  to  give  rise  to  lots  of  involved  lymph  nodes,  in  fact  any  nodes  at  all  (cell  values  of  -4.1  and  -

18.4),  with  the  least  aggressive  tumours  being  more  likely  to  spread  to  zero  nodes  (cell  value  of  8.8).

Now  that  you  have  a  good  sense  of  what’s  going  on  with  your  data  you  can  test  whether  there  is  an

association  between  tumour  aggressiveness  and  number  of  involved  lymph  nodes,  and  for  this  you

need  the  Pearson  Chi-Squared  Test  –  and  yes,  this  is  the  same  Pearson  that  came  up  with  the

correlation  test.

The  more  you  delve  into  stats,  the  more  often  you  will  bump  into**Karl  Pearson**  and**Ronald  A  Fisher**

(more  on  him  soon!)  –  these  guys  are  Gods  in  the  stats  world!

Well,  as  it  turns  out  you’ve  already  done  most  of  the  work  for  the  Chi-Squared  Test.  All  you  need  to

do  now  is  add  up  all  the  Chi-Squared  cell  values  (ignoring  the  sign)  and  you  have  the  Chi-Squared

value  for  the  whole  table.

This  is  the  number  that  you  get  when  you  run  a  Chi-Squared  Test  on  your  observed  values,  and  it

tells  you  whether  there  is  evidence  of  an  association  between  your  height  and  weight  categories.

The  important  outputs  that  you  will  need  when  you  run  the  Chi-Squared  Test  on  your  data  are:





**Visualisation**

o **mxn  summary  table**

o **mxn  table  of  Chi-Squared  cell  values**

**Statistic**

o **value**

o **number  of  degrees  of  freedom**

o **p-value**

**The  summary  table**  is  important  because  it  gives  you  a  good  feel  for  the  distribution  of  your  data.

However,  it  can  be  difficult  to  tell  if  any  of  the  cells  are  significantly  different  from  what  should  be

expected  by  chance.

**The  Chi-Squared  cell  value  summary  table**  is  important  because  it  tells  you  if  any  of  the  cells  are

different  from  what  should  be  expected  by  chance  and  allows  you  to  easily  identify  the  critical  cell.

30

|  |  |  |
| --- | --- | --- |
|  | **df** | **P** |

**The** **value**  is  not  so  important,  because  on  its  own  it  tells  you  little  about  whether  or  not  there  is

evidence  of  an  association  between  your  variables.

**The  number  of  degrees  of  freedom  (df)**  similarly  is  not  so  important,  but  when  you  put  the

value

and  df  together,  you  can  calculate  the  p-value.

**The  p-value**  is  important  because  it  tells  you  if  there  is  evidence  of  an  association  between  the  pair

of  variables.

Figure  4.5  shows  the  results  of  running  a  Chi-Squared  Test  on  the  lymph  node  data.

Chi-Squared Statistics:

**Figure  4.5:  Typical  Results  of  a  Chi-Squared  Test**

97.55

4

<0.001

The  p-value  tells  you  that  indeed  there  is  a  significant  association  between  tumour  grade  and

number  of  involved  lymph  nodes.  What  the  statistic  doesn’t  tell  you  is  what  is  responsible  for  the

association  or  whether  the  association  is  positive  or  negative.  This  is  why  it’s  important  to  get  your

hands  dirty  with  the  data.

Note  here  that  you  don’t  calculate  an  effect  size.  Although  there  are  effect  size  calculations  that  can

be  done  with  mxn  tables  there  is  a  lot  of  debate  as  to  how  useful  they  are.  What  is  more  useful  is  to

identify  a  portion  of  the  table  from  the  Chi-Squared  cell  values  and  isolate  that  portion  so  that  you

get  a  2x2  table.

From  here,  you  would  then  do  a  standard  analysis  for  2x2  tables,  which  by  a  remarkable  coincidence

is  exactly  what  we’re  looking  at  next!

Read  on…

31

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Age** | |
| **Under  50** | **Over  50** |
| **Metastases** | **No** | **83** | **487** |
| **Yes** | **19** | **61** |

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Age** | |
| **Under  50** | **Over  50** |
| **Metastases** | **No** | **-0.46** | **0.09** |
| **Yes** | **3.31** | **-0.62** |

***4.2.2:  2x2  categorical***

Arranging  the  data  for  2  dichotomous  variables  (that’s  stats-speak  for  ‘having  2  parts’)  is  exactly  the

same  as  for  analysing  mxn  data,  but  this  type  of  analysis  has  a  special  group  of  statistics  that  are

much  more  powerful  and  accurate.

OK,  let’s  look  at  a  different  example.

Let’s  say  that  you’re  investigating  whether  colon  cancer  patients  under  50  years  old  are  more  likely

for  their  primary  cancer  to  metastasise  (spread)  to  other  parts  of  the  body.  Your  data  are  shown  in

Figure  4.6.

**Figure  4.6:  2x2  Contingency  Table**

As  with  the  3x3  table  above,  you  can  work  out  the  Chi-Squared  cell  values  for  this  table  and  this  will

help  you  to  identify  if  any  particular  cell  is  larger  than  the  others,  like  in  Figure  4.7.

**Figure  4.7:  2x2  Contingency  Table  of  Chi-Squared  Cell  Values**

The  critical  Chi-Squared  cell  value  (bottom  left)  tells  you  that  the  metastases  are  more  likely  than

expected  by  chance  in  the  under  50  cohort  (group).  Note  that  the  top-left  to  bottom  right  diagonal

contains  only  negative  Chi-Squared  cell  values,  whilst  the  opposite  diagonal  contains  only  positive

Chi-Squared  cell  values.  This  tells  you  that  there  is  a  negative  association  between  age  and  colon

cancer  metastases  (as  one  variable  increases,  the  other  decreases).

32

With  the  mxn  table,  you  used  Pearson’s  Chi-Squared  Test  to  figure  out  if  there  was  a  relationship

between  the  variables.  While  the  Chi-Squared  Test  is  still  a  fine  test  to  use,  it  is  known  to  be  a  little

inaccurate,  particularly  when  the  counts  in  any  of  the  cells  are  small.  Various  smart  mathematicians

and  statisticians  have  tackled  this  problem  over  the  years  and  come  up  with  a  whole  variety  of  other

ways  to  analyse  the  2x2  table  more  accurately.

Although  all  of  the  newer  methods  are  a  lot  more  complicated,  these  days  it  doesn’t  really  matter

very  much  –  modern  computers  do  all  the  grunt  work  for  you  and  it’s  just  as  quick  to  run  a  Fisher’s

Exact  Test  (see,  I  told  you  we’d  run  into  Fisher  again,  didn’t  I?)  as  a  Chi-Squared  Test.

So,  let’s  have  a  look  at  a  few  stats  that  you  might  use  for  the  2x2  table.

The  Chi-Squared  Test  is  easiest  and  quickest  to  compute,  but  is  the  least  accurate  choice.  A  more

accurate  choice  is  the  Chi-Squared  Test  with  Yates’  Correction.

Better  still  is  the  Fisher’s  Exact  Test,  but  this  has  the  same  drawback  as  the  Chi-Squared  Test,  so

instead  we  can  use  the  Fisher’s  Exact  Test  with  Lancaster’s  mid-p  correction.

It  may  be  a  rather  long  name,  but  this  is  currently  thought  to  be  the  most  accurate  test  for  2x2

tables.  Alternatives  do  exist  (such  as  Barnard’s  Exact  Test),  but  they  are  computationally  much  more

complex,  are  not  more  accurate  and  I  don’t  intend  to  go  into  them  any  further  here.

Figure  4.8  shows  a  selection  of  results  from  each  of  these  tests  for  a  single  2x2  table.  Note  how  the

p-values  become  smaller  the  further  down  the  table  you  go.  Well,  mostly  anyway…

So  if  there  are  different  stats  tests  that  you  could  use  and  they  all  give  slightly  different  results,

which  one  is  right  and  which  one  should  you  use?

Well,  that’s  the  thing  about  stats.  It’s  not  about  right  or  wrong,  it’s  about  confidence.  Statistics  is  the

study  of  uncertainty,  so  you’ll  never  get  answers  that  are  black  or  white.

What  I  can  tell  is  that  the  stats  community  generally  agrees  that  the  results  lower  down  in  Figure  4.8

are  accurate  more  often  than  the  results  above  it.  The  further  you  go  down  the  table,  the  more

confidence  you  will  have  in  the  result.

My  particular  preference  is  the  mid-p  correction  and  is  the  one  I  recommend  you  use,  but  you  may

use  any  of  them  as  long  as  you  understand  their  limitations.

33

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **P-value** | | |
| **Left  tail**  **(negative)** | **Right  tail**  **(positive)** | **2-tail**  **(both)** |
|  | **Test** | N/A               N/A            0.034 | | |
|  | **Test  with  Yates’  Correction** | N/A               N/A            0.051 | | |
|  | **Fisher's  Exact  Test  (FET)** | 0.029            0.986           0.047 | | |
|  | **FET  With  Lancaster's  Mid-P  Correction** | 0.022            0.978           0.040 | | |

|  |  |  |
| --- | --- | --- |
| **Odds**  **Ratio** | **Lower**  **95%  CI** | **Upper**  **95%  CI** |

Statistical Results:

Effect Size:

**Figure  4.8:  Some  Typical  Results  of  Analysis  of  a  2x2  Contingency  Table**

0.547

0.301

1.002

The  important  outputs  that  you  will  need  when  you  run  a  2x2  analysis  on  your  data  are:







**Visualisation**

o **2x2  summary  table**

o **2x2  table  of  Chi-Squared  cell  values**

**Statistic**

o **p-value**

**Effect  Size**

o **odds  ratio**

**The  summary  table**  is  important  because  it  gives  you  a  good  feel  for  the  distribution  of  your  data.

However,  it  can  be  difficult  to  tell  if  any  of  the  cells  are  significantly  different  from  what  should  be

expected  by  chance.

**The  Chi-Squared  cell  value  summary  table**  is  important  because  it  tells  you  if  any  of  the  cells  are

different  from  what  should  be  expected  by  chance  and  allows  you  to  easily  identify  the  critical  cell.

**The  p-value**  is  important  because  it  tells  you  if  there  is  evidence  of  an  association  between  the  pair

of  variables.

34

**The  odds  ratio**  is  important  because  it  tells  you  how  strongly  related  the  variables  are,  and

represents  the  different  odds  between  the  groups.

I’m  sure  you  noticed  that  for  each  of  the  Fisher’s  Exact  Tests  there  are  3  different  p-values;  the  Left

tail,  Right  tail  and  2-tail  p-values.  ‘Which  should  I  use?’,  I  hear  you  ask.

Well,  if  you  made  a  hypothesis  before  you  ran  the  test  then  you  should  select  the  one  that

corresponds  with  your  hypothesis.  Having  a  prior  hypothesis  means  that  you  can  legitimately  ignore

the  2-tailed  p-value  and  choose  either  the  left-tail  or  right-tailed  p-value.  If  you  hypothesised  that  as

one  variable  goes  up  the  other  goes  down,  then  you  hypothesised  a  negative  association  and  you

should  choose  the  left-tailed  p-value.

Alternatively,  if  you  hypothesised  a  positive  association  (as  one  variable  goes  up  so  does  the  other)

then  you  should  use  the  right-tailed  p-value.

If  you  had  no  prior  hypothesis  at  all,  then  I’m  afraid  that  you’re  stuck  with  the  2-tailed  p-value.  No

cheating  now…

You  probably  also  noticed  that  the  odds  ratio  is  less  than  one.  This  confirms  that  there  is  a  negative

association  between  age  and  colon  cancer  metastases.  For  negative  odds  ratios  you  can  take  the

reciprocal  to  give  you  a  number  that  is  more  intuitive,  so

gives  you  an  odds  ratio  of  1.828;

that  is,  colon  cancer  patients  in  the  under  50  cohort  were  almost  twice  as  likely  to  suffer  from

metastases  as  those  in  the  over  50  cohort.

Taking  the  reciprocal  of  the  confidence  intervals  tells  you  that  you  can  be  95%  confident  that  the

true  odds  ratio  lies  in  the  range  [0.998;  3.322].  Since  these  confidence  intervals  span  the  number

one,  you  cannot  be  certain  that  your  variable  will  have  any  effect  on  the  outcome.

Since  the  lower  confidence  interval  is  close  to  one  you  cannot  definitively  accept  or  reject  the  null

hypothesis  based  solely  on  the  odds  ratio  (which  you  shouldn’t  anyway),  but  should  take  a  pragmatic

approach  and  take  all  results  into  account;  all  p-values  and  odds  ratios.

***4.2.3:  2xn  categorical***

Sometimes  you  may  find  a  special  case  of  when  one  of  your  variables  has  2  categories  while  the

other  has  more  than  2  –  but  they  must  be  ordinal!  You  could  analyse  this  by  treating  it  as  mxn  and

simply  running  a  Chi-Squared  Test.

However,  there  is  a  better  way!

What  you’re  looking  for  here  is  a  clear  progression  in  one  group  compared  to  the  other,  and  in  this

case,  the  Chi-Squared  for  Trend  is  a  much  more  accurate  statistic  than  the  plain  vanilla  Pearson  Chi-

Squared  Test  because  the  order  of  categories  is  taken  into  account  as  well  as  the  cell  counts.

35

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Treatment** | |  |
| **Standard** | **New** | **Proportion** |
| **Assessment** | **Large  Improvement** | **4** | **28** | **0.875** |
| **Small  Improvement** | **13** | **10** | **0.435** |
| **No  Change** | **3** | **2** | **0.400** |
| **Small  Deterioration** | **12** | **5** | **0.294** |
| **Large  Deterioration** | **6** | **6** | **0.500** |
| **Death** | **14** | **4** | **0.222** |

You  arrange  your  data  in  a  2xn  contingency  table  as  before,  but  rather  than  work  out  the  Chi-

Squared  cell  values,  you  work  out  the  proportions  in  each  row.

Let’s  have  a  look  at  an  example  of  a  new  treatment  compared  with  the  standard.  The  six  levels  of

assessment  of  patients’  condition  are  tabulated  against  the  standard  and  new  treatments.

In  a  standard  Chi-Squared  test,  you  would  get  the  same  result  irrespective  of  the  ordering  of

categories,  but  with  the  Chi-Squared  for  Trend  test  if  there  is  a  relationship  between  the  new

treatment  and  patient  condition  you  would  expect  a  greater  improvement  for  the  new  treatment

over  the  standard  treatment,  and  this  would  be  reflected  in  the  proportions,  Figure  4.9.

**Figure  4.9:  2x6  Contingency  Table  with  Proportions**

If  you  plot  the  proportions  on  a  scatter  plot  with  a  regression  line,  this  will  help  you  see  if  there  is  a

clear  trend  in  proportions  between  the  groups,  Figure  4.10.

At  this  point,  you  might  be  thinking  ‘aha,  a  scatter  plot  with  a  regression  line  –  I  know  how  to

analyse  this,  it’s  the  Pearson  correlation!’.  If  you  were  thinking  this,  nice  try,  but  get  to  the  back  of

the  class!

The  Pearson  correlation  is  for  when  both  of  your  variables  are  continuous.  Although  one  of  the

variables  in  the  plot  is  now  continuous  (the  proportion),  the  other  is  still  clearly  categorical,  so  you

can’t  use  a  correlation  test.

We  plot  the  proportion  to  help  us  to  understand  the  trend  in  the  data,  but  both  variables  are

categorical  and  so  the  Chi-Squared  for  Trend  is  the  most  appropriate  test  here.

36

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |  |  |  |  |  |  |

**Figure  4.10:  Plot  of  Proportions*versus*  Assessment  Categories**

LargeImprovement

SmallImprovement

NoChange

SmallDeterioration

LargeDeterioration

Death

1.000

0.900

0.800

0.700

0.600

0.500

0.400

0.300

0.200

0.100

0.000

Figure  4.11  shows  the  results  of  running  both  a  Chi-Squared  and  a  Chi-Squared  for  Trend  test  on  the

data  in  Figure  4.9.

In  this  case,  it  doesn’t  really  matter  whether  you  ran  a  Chi-Squared  or  a  Chi-Squared  for  Trend  test,

the  result  is  highly  significant  in  either  case.  You  should  have  noticed  though  that  the  proportion  in

the  Large  Improvement  category  is  substantially  larger  than  in  all  other  categories,  suggesting  that

the  new  treatment  is  significantly  more  effective  than  the  standard  treatment.  It  would  not  be

inappropriate  to  have  split  the  table  below  this  category  to  create  a  2x2  contingency  table  and

running  the  corresponding  tests,  as  in  Section  4.2.2.

To  illustrate  how  important  the  ordering  of  categories  is  in  a  Chi-Squared  for  Trend  test,  try  moving

the  Large  Improvement  row  from  the  top  of  Figure  4.9  to  the  bottom,  below  the  Death  category,

and  re-running  the  analyses.

The  Chi-Squared  test  results  will  be  precisely  the  same  with  a  p-value  of  <0.0001,  but  the  p-value  for

the  Chi-Squared  for  Trend  test  will  be  markedly  different  at  p=0.255.  Just  by  moving  one  row,  your

result  goes  from  highly  significant  to  not-even-close-to-significant.

37

|  |  |  |
| --- | --- | --- |
|  | **df** | **P** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **df** | **P** | **Gradient** | **Intercept** |

Chi-Square Test:

Chi-Square for Trend Test:

26.97

5

<0.0001

**Figure  4.11:  Chi-Squared  Results  of  Analysis  of  a  2x6  Contingency  Table**

17.93

1

<0.0001

-0.0907

0.7718

The  important  outputs  that  you  will  need  when  you  run  the  Chi-Squared  for  Trend  Test  on  your  data

are:







**Visualisation**

o **2xn  summary  table  with  proportions**

o **scatter  plot  of  proportions  with  regression  line**

**Statistic**

o **for  trend  value**

o **p-value**

**Effect  Size**

o **equation  of  the  line  (gradient  &  intercept)**

**The  2xn  summary  table**  is  important  because  it  gives  you  a  good  ‘feel’  for  your  data  and  helps  you

decide  if  there  may  be  a  trend.

**The  scatter  plot**  is  important  because  it  gives  you  a  sense  of  the  scale  of  the  trend  and  whether  the

plot  points  are  close  to  or  far  away  from  the  regression  line.

**The** **value**  is  not  so  important,  because  on  its  own  it  is  not  always  possible  to  tell  whether  the

observed  and  expected  cell  counts  are  different  enough  to  be  statistically  significant.

Note  that  the  number  of  degrees  of  freedom  in  a  Chi-Squared  for  Trend  Test  is  always  1.

**The  p-value**  is  important  because  it  tells  you  if  there  is  evidence  of  an  association  (a  trend  in  the

ratio  of  proportions)  between  the  pair  of  variables.

38

|  |
| --- |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

**The  equation  of  the  line,**

**AverageNumberofEggsLaid**

gives  us  a  way  to  predict

one  of  the  variables  if  we  know  the  value  of  the  other.

***4.2.4:  nxCategorical  v  Continuous***

When  you  have  one  variable  that  is  categorical  while  the  other  is  continuous,  the  best  way  to

analyse  this  is  to  aggregate  all  the  continuous  values  by  categories.  When  you  do  this,  you  can

compare  the  differences  in  the  means  across  the  groups.

Let’s  say  that  you  have  collected  data  on  the  average  number  of  eggs  laid  per  female  fruit  fly  per  day

for  the  first  14  days  of  life.  You  had  25  female  fruit  flies  in  each  of  three  genetic  lines;  25  that  were

selectively  bred  for  resistance  to  a  specific  pesticide,  25  for  susceptibility  to  the  same  pesticide,  and

25  that  were  non-selected,  i.e.  they  were  the  control  group.

So  you  would  now  gather  together  the  average  number  of  eggs  laid  per  female  fruit  fly  per  day  for

the  Resistance  category.  Do  this  for  each  of  the  categories  and  you’ll  have  three  sets  of  continuous

data,  representing  the  Resistant,  Susceptible  and  Non-Selected  categories.

To  help  you  understand  how  the  data  is  distributed,  you  can  plot  box-and-whiskers  plots  of  the

grouped  data,  like  Figure  4.12.

**Figure  4.12:  Box-and-Whiskers  Plot  of  Average  Number  of  Eggs  Laid  by  Fruit  flies**

50

40

30

20

10

Resistant

Susceptible

Non-Selected

39

The  box  tells  you  the  first  and  third  quartiles  of  the  data  (Q1  and  Q3),  with  the  median  somewhere

between.  The  limits  of  the  whiskers  tell  you  which  data  lie  within  1.5  inter-quartile  ranges  (IQR)

away  from  Q1  and  Q3:

and

where

Outliers  are  the  asterisks  that  lie  outwith  the  whiskers.

Box-and-whiskers  plots  are  great  for  getting  a  ‘feel’  for  your  data,  locating  the  central  point  and

deciding  if  the  data  are  distributed  normally,  symmetrically  or  skewed.

Looking  at  the  central  points  and  widths  of  the  distributions,  they  also  give  you  a  sense  of  whether

the  means  are  likely  to  be  different  (far  apart)  or  similar  (close  together).

The  kinds  of  questions  you  should  be  asking  about  these  plots  include:









Are  the  individual  plots  normally  distributed?

o      If  not  normally  distributed,  are  they  at  least  symmetrical?

o      Are  any  of  the  plots  skewed?

Are  the  means  close  together  or  far  apart?

Is  there  a  progression  in  the  means  across  the  groups  (if  appropriate)

o      Are  they  increasing  or  decreasing?

Are  the  distributions:

o Tall  and  thin  (small  variation  in  values)?

o Short  and  fat  (high  variation  in  values)?

If  the  distributions  are  tall  and  thin  then  there  is  a  higher  likelihood  that  the  group  means  are

different,  even  if  they  are  quite  close  together.  On  the  other  hand,  if  the  distributions  are  short  and

fat  it’s  less  likely  they  will  be  different,  even  if  the  means  are  far  apart.

Unless  the  means  are  pretty  much  the  same  (i.e.  clearly  not  different)  or  they  are  far  apart  with

narrow  distributions  (i.e.  clearly  very  different),  then  the  only  way  to  find  out  for  sure  is  to  run  a

statistical  analysis.

For  this,  we  need  either  the  ANOVA  (ANalysis  Of  VAriance)  or  the  Kruskall-Wallis  Test.  Figure  4.13

will  help  you  decide  which  one  to  choose.

40

|  |  |
| --- | --- |
| **ANOVA** | **Kruskall-Wallis  Test** |
| Assumes  that  all  distributions  are  normal | Makes  no  assumption  about  the  data  distributions |
| May  still  be  used  if  the  data  are  non-normal,  but  are  symmetrically  distributed | Use  this  when  the  data  are  non-normal  and  not  symmetrically  distributed |
| If  the  assumptions  are  met,  the  ANOVA  is  more  powerful  than  the  Kruskall-Wallis  Test | When  the  assumptions  are  not  met,  the  Kruskall-Wallis  Test  is  more  appropriate |

**Figure  4.13:  Choice  of  ANOVA  and  Kruskall-Wallis  Tests**

The  ANOVA,  developed  by  our  old  friend  Ronald  A  Fisher,  is  more  accurate  than  the  Kruskall-Wallis

Test  (surprisingly,  not  developed  by  Fisher!)  but  it  depends  on  the  data  in  all  groups  being  normally

distributed  (or  at  least  symmetrically  distributed).  If  any  of  the  groups  fail  this  test,  then  confidence

in  the  result  will  likely  be  lower  and  you  should  use  the  Kruskall-Wallis  Test.

If  you’re  unsure  which  one  to  use,  then  the  Kruskall-Wallis  Test  will  always  be  valid  and  you  can  use

this  (although  you  didn’t  hear  it  from  me  –  my  membership  of  GARGS,  the  Guild  of  Argumentative

Statisticians,  may  be  revoked!).

There  is  a  whole  family  of  different  types  of  ANOVA  tests,  depending  on  the  number  of  factors  that

you’re  investigating.  If  you’re  looking  into  just  one  factor  (the  genetic  line  determines  the  average

number  of  eggs  laid),  then  you  should  use  the  one-way  ANOVA.  If  more  than  one  factor  is  involved

(such  as  the  average  number  of  eggs  laid  being  determined  by  genetic  line,  age,  nutrition,  etc.),  then

there  are  more  complicated  versions  of  the  ANOVA  you  can  use,  but  these  are  beyond  the  scope  of

this  book.

The  important  outputs  that  you  will  need  when  you  run  the  ANOVA  or  Kruskall-Wallis  Test  on  your

data  are:





**Visualisation**

o **box-and-whiskers  plot**

**Statistic**

o **central  values  (means  or  medians)  of  each  group**

o **variation  (SD,  variance,  inter-quartile  range  or  95%  confidence  intervals  of  each**

**group**

o **p-value**

Note  that  here  we  don’t  calculate  an  effect  size.  Although  there  are  effect  size  calculations  that  can

be  done  with  this  type  of  analysis  there  is  a  lot  of  debate  as  to  how  useful  they  are  as  the  calculation

basically  allows  you  to  categorise  the  effect  size  as  small,  medium  or  large.  Not  very  useful  really…

41

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **df** | **SS** | **MS** | **F** | **P** |

|  |  |  |
| --- | --- | --- |
| **S** | ***R2*** | ***2***  ***R*  (adj)** |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **N** | **Mean** | **SD** |

**The  box-and-whiskers  plot**  is  important  because  it  tells  you  where  the  central  points  of  the  data  lie,

whether  the  data  are  skewed  and  which  points  are  outliers.

**The  central  values  and  measures  of  variation**  are  important  because  it  is  precisely  these  that  are

being  compared.

**The  p-value**  is  important  because  it  tells  you  if  there  is  evidence  of  an  association  between  the

categories.  The  other  outputs  can  be  useful  to  understand  your  data,  but  only  the  p-value  can  truly

tell  you  if  there  is  evidence  of  an  association.

Figure  4.14  is  the  output  from  a  one-way  ANOVA  of  the  fruit  fly  data.

One-way ANOVA:

**Error**

**Factor**

72

2

5659.0

1362.2

78.6

681.1

**Total**

8.67

<0.001

74

7021.2

Regression Statistics:

Descriptive Statistics:

8.866

19.40%

17.16%

Susceptible

Resistant

Non-Selected

Pooled  SD

25

25

25

23.628

25.256

33.372

9.768

7.772

8.942

8.866

**Figure  4.14:  Typical  Results  of  a  One-Way  ANOVA**

The  p-value  tells  you  that  there  is  a  significant  difference  between  the  groups  and  that  there  is

strong  evidence  that  the  average  number  of  eggs  laid  per  day  is  different  among  the  genetic  lines.

42

|  |  |
| --- | --- |
| **2-sample  t-test** | **Mann-Whitney  U-test** |
| Assumes  that  both  distributions  are  normal | Makes  no  assumption  about  the  data  distributions |
| May  still  be  used  if  the  data  are  non-normal,  but  are  symmetrically  distributed | Use  this  when  the  data  are  both  non-normal  and  not  symmetrically  distributed |
| If  the  assumptions  are  met,  the  2-sample  t-test  is  more  powerful  than  the  Mann-Whitney  U-test | When  the  assumptions  are  not  met,  the  Mann-Whitney  U-test  is  more  appropriate |

On  its  own  though,  the  ANOVA  result  doesn’t  tell  you  which  of  the  groups  are  different  from  the

others.  The  box-and-whiskers  plot  can  tell  you  this.  It  appears  from  these  plots  that  the  resistant  and

susceptible  genetic  lines  are  both  significantly  different  from  the  non-selected  genetic  line,  but  are

not  different  from  each  other.  This  is  can  be  confirmed  by  analysing  these  genetic  lines  pair  wise,

which  is  precisely  what  we  will  look  at  next.

***4.2.5:  2xCategorical  v  Continuous***

When  your  categorical  variable  has  2  levels,  you  deal  with  the  data  in  precisely  the  same  way  as  if  it

has  multiple  levels.

You  aggregate  all  the  continuous  data  into  their  respective  categories,  plot  histograms  and  box-and-

whisker  plots  to  check  for  normality,  symmetricity  and  skewness,  and  to  assess  the  variability  of  the

data  in  each  group  and  whether  the  central  values  are  likely  to  be  different.

Then  you  run  either  the  2-sample  t-test  (if  both  distributions  are  normal)  or  Mann-Whitney  U-test.

Figure  4.15  will  help  you  decide  which  is  most  appropriate  for  your  data.

**Figure  4.15:  Choice  of  2-Sample  t-test  and  Mann-Whitney  U-test**

FYI  –  you  can  run  an  ANOVA  on  these  data  and  get  exactly  the  same  result  as  the  2-sample  t-test.

Similarly,  the  Mann-Whitney  U-test  and  Kruskall-Wallis  Test  will  also  give  you  the  same  result.

The  2-sample  t-test  is  also  known  as  the  Student’s  t-test.  When**William  Sealy  Gosset**  –  employed  by

Guinness  Breweries  to  improve  the  quality  of  stout  –  was  forbidden  by  his  employer  from  publishing

any  findings,  he  instead  published  his  work  under  the  pseudonym  ‘Student’.  And  a  legend  was

born…

Anyway,  let’s  have  a  look  at  an  example.  Suppose  you  wish  to  find  out  if  US  cars  are  more

economical  than  Japanese  cars,  so  you  measure  the  number  of  miles  per  gallon  for  a  whole  bunch  of

43

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **N** | **Mean** | **SD** | **SE  Mean** |

|  |  |  |
| --- | --- | --- |
| **T** | **df** | **P** |

US  and  Japanese  cars  under  identical  conditions.  The  results  of  a  2-sample  t-test  are  shown  in  Figure

4.16.

Descriptive Statistics:

Japanese

US

79

249

30.48

20.14

6.11

6.41

0.69

0.41

Regression Statistics:

**Figure  4.16:  Results  of  a  2-sample  t-test**

-12.95

136

<0.001

The  results  here  tell  you  that  there  is  a  significant  difference  (p<0.001)  between  US  and  Japanese

cars,  and  that  in  this  dataset  Japanese  cars  run  on  average  50%  further  than  US  cars  per  gallon.

The  important  outputs  that  you  will  need  when  you  run  the  2-sample  t-test  or  Mann-Whitney  U-test

on  your  data  are:





**Visualisation**

o **box-and-whiskers  plot**

**Statistic**

o **central  values  (means  or  medians)  of  each  group**

o **variation  (SD,  variance,  inter-quartile  range  or  95%  confidence  intervals  of  each**

**group)**

o **p-value**

Again,  we  don’t  calculate  an  effect  size  with  this  analysis.  Although  there  are  effect  size  calculations

that  can  be  done,  such  as  Cohen’s  d,  they  are  not  really  very  useful  as  the  calculation  only  allows  you

to  categorise  the  effect  size  as  small,  medium  or  large.

**The  box-and-whiskers  plot**  is  important  because  it  tells  you  where  the  central  points  of  your  data

lie,  whether  the  data  are  skewed  and  which  points  are  outliers.

44

**The  central  values  and  measures  of  variation**  are  important  because  it  is  precisely  these  that  are

being  compared.

**The  p-value**  is  important  because  it  tells  you  if  there  is  evidence  of  an  association  between  the  pair

of  variables.  The  other  outputs  can  be  useful  to  understand  your  data,  but  only  the  p-value  can  truly

tell  you  if  there  is  evidence  for  an  association  between  the  variables.

**4.3:  Survival  Analysis**

Survival  analysis  is  a  method  for  analysing  the  time  to  an  event  rather  than  the  event  itself.

For  example,  if  you  have  a  pair  of  categorical  variables  Gender  [male;  female]  and  Death  [yes;  no],

you  can  simply  run  a  Fisher’s  Exact  Test  to  find  out  if  men  are  more  likely  than  women  to  die

(because  of  some  disease  or  other).

Alternatively,  if  you  have  a  third  variable  ‘Time  To  Event’,  then  you  can  see  how  the  event  in

question  unfolds  over  time  rather  than  just  a  final  snapshot  at  the  end  of  the  study.

Survival  analysis  is  used  extensively  in  medical  studies  to  assess  survival  rates  or  disease  recurrence

rates,  in  engineering  to  assess  failure  rates  of  mechanical  systems  or  parts,  and  in  many  other

sectors,  such  as  economics,  sociology  and  sales  forecasts.

Let’s  look  at  the  2  survival  data  variables  in  question.

**The  ‘event’  variable**.  This  is  the  event  under  study,  for  instance  the  survival  (or  otherwise)  of  a

patient.  In  this  data  column  you  note  whether  the  patient  died  during  the  period  of  the  study

(record  ‘dead’)  or  survived  the  study  period  (record  ‘alive’).

This  is  also  known  as  the  ‘censor’  variable  because  it  is  an  observation  that  is  only  partially  known

(i.e.  death  may  occur  outside  the  study  period).  When  you  record  this  variable,  your  statistics

program  will  ask  you  which  of  the  data  levels  corresponds  to  an  ‘event’,  and  which  to  a  ‘censor’.  It  is

important  to  get  this  right,  otherwise  the  analysis  will  not  be  correct.  Death  is  an  event  (you  can

pinpoint  exactly  when  it  occurred),  whereas  alive  is  not  and  is  the  censor.

**The  ‘time  to  event’  variable**.  It  is  important  that  a  zero  time  point  be  defined  for  each  sample  or

patient  in  the  study.

For  example,  for  survival  data,  you  need  to  figure  out  the  time  from  an  event  (say,  time  from

diagnosis)  to  the  event  in  question  (which  in  this  case  would  be  either  death  or  the  end  of  the  study

period).

45

|  |  |
| --- | --- |
|  |  |

When  collecting  such  data,  you  don’t  collect  time,  you  collect  dates.  For  each  patient  you  collect

CumulativeSurvival(%)

their  date  of  diagnosis.  If  they  die  during  the  study  period  you  also  collect  their  date  of  death  and

note  their  status  (dead)  in  the  event  column.  If  they  remain  alive  at  the  end  of  the  study,  then  you

record  this  in  the  event  column  (alive).  You  will  then  have  2  date  columns  that  you  can  subtract  to

get  the  time  from  diagnosis  to  death  or  end  of  the  study.

This  time-to-event  variable  is  known  as  ‘right-censored’,  since  patients  surviving  beyond  (i.e.*to  the*

*right  of*)  the  end  of  the  study  period  must  be  censored  as  their  time  of  death  is  yet  undetermined.  It

is  important  to  choose  right-censored  survival  analysis  to  avoid  bias  in  your  results.

Once  you  have  the  event  and  time-to-event  columns  you  have  everything  you  need  to  perform  a

survival  analysis.  If  you  have  data  about  a  specific  disease  and  wish  to  know  how  the  time  to  death

unfolds  for  everyone  in  the  study  you  can  do  a  survival  analysis  and  get  a  single  Kaplan-Meier  plot

like  Figure  4.17.

Time  (Years)

100

90

80

70

60

50

40

30

20

10

0

0

2

4 6

8

10

**Figure  4.17:  Kaplan-Meier  Plot**

OK,  if  you’ve  never  seen  a  Kaplan-Meier  plot  before,  the  graph  above  might  look  a  little  strange.

Let’s  go  through  it  step-by-step:

46

















The  x-axis  is  the  time  difference  from  the  first  event  (i.e.  time  of  diagnosis)  to  the  failure

event  (i.e.  death)  or  censor,  and  always  begins  at  time  t=0

The  y-axis  is  the  percentage  (or  proportion)  of  survivors  remaining  in  the  study  at  any  given

time  point  and  the  range  is  always  [0,  100],  or  [0,  1]  if  shown  as  a  proportion

At  the  beginning  of  the  study  (time  =  0)  there  are  no  failure  events

The  graph  always  starts  at  the  plot-point  (0,100)

As  time  proceeds,  the  plot  moves  horizontally  parallel  to  the  x-axis  until  the  next  event  (or

censor)

When  an  event  occurs,  the  plot  drops  vertically  in  the  graph  by  the  proportion  of  the

number  of  people  remaining  in  the  study  at  that  point

o      For  example,  if  at  a  given  time-point  1  patient  out  of  10  remaining  in  the  study  dies,

the  graph  will  drop  by  10%  from,  say,  63%  to  53%

When  a  censor  occurs,  a  ‘+’  is  inserted  onto  the  line  to  denote  a  censor,  and  the  line

continues  horizontally

The  final  point  on  the  graph  is  always  a  ‘+’  (the  final  data  point  is  censored)  or  drops  onto

the  x-axis  (the  final  data  point  is  a  death)

The  real  power  of  this  kind  of  analysis  though  is  that  you  can  use  it  to  compare  the  survival  rates  of

different  groups  or  cohorts  in  the  data.  For  this,  you  will  need  a  third  categorical  variable,  which  can

take  2  or  more  levels,  such  as  Gender  [male;  female]  or  Age  [30s;  40s;  50s;  etc.].

***4.3.1:  nxCategorical  v  Survival***

When  you’re  running  a  survival  analysis  of  a  variable  with  n  categories,  what  you  will  get  is  n  survival

curves  all  on  the  same  set  of  axes,  like  Figure  4.18.

In  this  example,  there  are  3  groups  of  data  corresponding  to  Disease  Severity.  Note  how  the  least

severe  is  the  highest  on  the  graph  (highest  probability  of  survival),  whereas  the  most  severe  has  the

lowest  probability  of  survival.

The  kinds  of  questions  you  should  be  asking  about  these  plots  include:













Is  there  a  clear  separation  between  the  different  groupings?

Is  the  ordering  of  the  groupings  what  you  would  expect?

Do  any  of  the  plots  cross  over?

Are  there  any  large  jumps  in  the  plots  towards  the  end  (indicating  a  decrease  in  confidence)?

Do  any  of  the  groupings  drop  to  zero  on  the  y-axis?

o      Is  this  what  you  would  expect?

Are  there  any  groupings  that  clearly  clump  together?

o      Should  they  be  aggregated,  for  example  to  compare  group  1*versus*  groups  2&3

47

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | |  | | | | | | | |  |
|  |  | |  | | |  | | | | |  | | |  |
|  |  |
|  |
|  |  | | | | | |  |
|  |  | | | | |
|  |  | |  | | | | |  |
|  |
|  |
|  |
|  |
|  |
|  | Disease  Severity  =  1  Disease  Severity  =  2  Disease  Severity  =  3 | | | | | | | | | | |  | |
|  |
|  | | | |  | | | | |  | | |

Time  (Years)

CumulativeSurvival(%)

100

90

80

70

60

50

40

30

20

10

0

0

2

4 6

8

10

**Figure  4.18:  Kaplan-Meier  Plot  with  3  Categories**

The  data  in  Figure  4.18  are  well  behaved,  there  is  a  clear  separation  in  the  groupings,  the

separations  are  consistent  and  there  are  no  large  jumps  in  the  plots.  Actually,  these  data  aren’t  well

behaved;  they  are*very  well*  behaved…

Perhaps  the  most  appropriate  statistic  for  this  analysis  is  the  Log-Rank  Test,  which  tells  us  whether

there  is  a  significant  difference  between  the  plots.

The  important  outputs  that  you  will  need  when  you  run  a  Kaplan-Meier  survival  analysis  on  n  data

categories  are:







**Visualisation**

o **Kaplan-Meier  plot**

**Statistic**

o **p-value**

**Effect  Size**

o **Hazard  Ratio**

48

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Total** | **Event** | **Censor** | **%  Censor** |

|  |  |  |
| --- | --- | --- |
|  | **df** | **P** |

|  |  |
| --- | --- |
|  | **Hazard  Ratio** |

**The  Kaplan-Meier  plot**  is  important  because  it  gives  you  a  good  ‘feel’  for  the  data.

**The  p-value**  is  important  because  it  tells  you  if  there  is  evidence  of  a  relationship  between  the

variable  and  the  probability  of  survival  over  a  given  time  period.

**The  Hazard  Ratio**  is  important  because  it  is  the  ratio  of  probabilities  of  death.  For  example  a  hazard

ratio  of  2  means  that  a  group  has  twice  the  chance  of  dying  than  a  comparison  group.  Although

hazard  ratios  can  only  be  calculated  for  2  groups,  it  may  be  useful  to  compute  the  hazard  ratio  for

each  group  compared  to  a  baseline  group.  For  example  if  you  designate  group  1  as  your  baseline

(most  likely  to  survive),  then  you  can  compute  the  hazard  ratio  for  groups  1  v  2  and  for  groups  1  v  3.

The  analysis  that  accompanies  Figure  4.18  is  shown  in  Figure  4.19.

Censor Table:

**Disease  Severity  2**

**Disease  Severity  1**

**Disease  Severity  3**

**Total**

449

580

82

**1111**

107

52

35

**194**

342

528

47

**917**

76.17

91.03

57.32

**82.54**

Log-Rank Statistics:

Hazard Ratios:

92.32

**Disease  Severity  2**

2

<0.001

**Disease  Severity  1**

**Disease  Severity  3**

2.92

1.00

6.20

**Figure  4.19:  Results  of  a  Kaplan-Meier  Survival  Analysis**

Looking  at  the  %  censor  values  of  the  three  categories,  you  can  see  than  the  chances  of  death

increase  substantially  as  disease  progresses  from  less  severe  to  more  severe.  This  suggests  that

49

there  may  be  a  difference  in  survival  between  the  three  groups.  You  could  do  a  Chi-Squared  or

Fisher’s  Exact  test  of  the  censor  table  and  it  would  give  you  a  useful  snapshot  of  the  study  outcome

at  the  final  time-point,  but  it  wouldn’t  give  you  any  information  of  how  the  study  events  unfolded

over  time.

Running  a  Kaplan-Meier  analysis  tells  you  that  for  the  two  least  severe  disease  categories,  the

likelihood  of  death  is  approximately  constant  over  the  study  period  (the  plots  are  roughly  linear  with

no  large  jumps).  In  the  least  severe  category  more  than  90%  of  patients  survive  beyond  10  years,

while  around  75%  of  the  patients  in  the  next  category  survive  to  10  years  (you  should  ignore  the

steep  jumps  at  the  end  of  the  plot  –  this  is  due  to  having  few  patients  left  in  the  study  and  can  bias

the  end-point).

With  the  most  severe  form  of  the  disease  roughly  30%  of  the  patients  die  within  the  first  two  years

and  50%  within  the  first  seven  years.  However,  after  seven  years  all  of  the  patients  survived  to  10

years.

The  Kaplan-Meier  plot  and  censor  table  suggest  that  there  may  well  be  a  significant  difference  in

survival  between  the  groups,  and  the  log-rank  p-value  statistic  confirms  this  (<0.001).

The  hazard  ratios  tell  you  that  you  are  almost  three  times  as  likely  to  die  if  your  disease  is  of  medium

severity  (2.92),  and  more  than  six  times  as  likely  to  die  if  you  have  the  most  severe  form  of  the

disease  (6.20)  compared  with  the  least  severe  form.

***4.3.2:  2xCategorical  v  Survival***

If  you  have  a  variable  with  2  categories  then  what  you’ll  get  is  a  survival  curve  and  associated  stats

for  2  groups  rather  than  n  groups.

However,  a  useful  analysis  is  when  you  have  n  groups  and  you  wish  to  make  a  more  detailed

comparison  of  one  grouping  of  variables*versus*  another  grouping.

For  example,  there  may  be  a  clear  distinction  between  a  group  of  ‘less  severe’  categories  (perhaps

denoted  as  1-4)  and  ‘more  severe’  (perhaps  categories  5-10).  In  this  case  you  may  wish  to  group

together  categories  1-4  into  a  single  data  category,  similarly  with  categories  5-10,  and  run  a  Kaplan-

Meier  analysis  with  2  categories  of  ‘less  severe’  and  ‘more  severe’.

The  advantage  here  is  that  you  will  get  a  clear  result.  The  p-value  and  hazard  ratio  will  clearly

distinguish  between  the  two  groups  and  will  give  you  a  good  understanding  of  the  difference

between  the  groups.  Perhaps  there  are  different  biological  processes  underpinning  the  two  different

groupings.

This  can  set  you  off  on  an  entirely  new  voyage  of  discovery…

The  important  outputs  that  you  will  need  when  you  run  a  Kaplan-Meier  survival  analysis  on  2  data

categories  are:

50







**Visualisation**

o **Kaplan-Meier  plot**

**Statistic**

o **p-value**

**Effect  Size**

o **Hazard  Ratio**

**The  Kaplan-Meier  plot**  is  important  because  it  gives  you  a  good  ‘feel’  for  the  data.

**The  p-value**  is  important  because  it  tells  you  if  there  is  evidence  of  a  relationship  between  the

variable  and  the  probability  of  survival  over  a  given  time  period.

**The  Hazard  Ratio**  is  important  because  it  is  the  ratio  of  probabilities  of  death  (or  survival).  For

example  a  hazard  ratio  of  3  means  that  a  group  has  three  times  the  chance  of  dying  than  a

comparison  group.

This  concludes  the  chapter  on  univariate  statistics.  Next  comes  multivariate  statistics.  It’s  going  to

get  a  little  bit  harder,  but  wait,  stop  running  –  it’s  really  not  that  difficult.  Honest!

In  chapter  5,  we’ll  go  through  the  difference  between  univariate  and  multivariate  stats,  and  you’ll

get  an  understanding  of  why  you  need  the  more  complex  statistics  to  get  a  strong  understanding  of

what’s  going  on  in  your  data.

51

**Chapter  5:  Multivariate  Statistics**

In  this  chapter  you  will  learn  that,  whilst  in  univariate  relationship  analysis  you  are  analysing  data

pair-wise,  in  other  words*this*  variable  against*that*,  in  multivariate  relationship  analysis  you  are

assessing  many  variables  (predictor  variables)  against  a  single  variable  (target,  outcome  or

hypothesis  variable)  simultaneously;  in  other  words,  testing*this*,*that*  and*the  other*  against  a  single

outcome.

Testing  multiple  variables  simultaneously  has  the  distinct  advantage  that  interactions  between  the

variables  can  be  controlled  for.

OK,  enough  of  the  jargon,  what  does  ‘controlled  for’  really  mean?

Well,  the  bottom  line  is  that  univariate  analyses  do  not  take  into  account  any  factors  other  than  the

ones  in  the  test.

For  example,  if  you  find  a  significant  relationship  between*this*  and*that*  in  univariate  analysis,  you

can’t  be  sure  you’re  seeing  the  full  story,  because  the  influence  of*the  other*  has  not  been  tested  for.

A  univariate  analysis  of*this*  against*that*  tells  you  whether  there  is  a  relationship  between  a  pair  of

variables,  but  it  doesn’t  tell  you  whether  that  relationship  is  independent  of  other  factors.

For  this,  you  need  to  run  a  multivariate  analysis,  which  distinguishes  between  those  variables  that

are:







not  related

dependently  related

o      i.e.  the  relationship  depends  on  other  factors

independently  related

o i.e.  the  relationship  does  not  depend  on  other  factors

Let’s  jump  right  in  and  have  a  look  at  the  different  types  of  multivariate  analyses.  After  that,  we’ll

deal  with  the  difference  between  univariate  and  multivariate  analysis  and  why  you  need  a  holistic

strategy  to  discover  all  the  independent  relationships  in  your  data.

**5.1  Types  of  Multivariate  Analysis**

There  are  3  types  of  multivariate  analysis  that  you  will  use  to  find  associations  and  correlations  in

your  data:







Logistic  Regression

Multivariate  Linear  Regression

Cox’s  Proportional  Hazards  Survival  Analyses

52

**Logistic  regressions**  are  used  when  your  outcome  variable  is  categorical.

When  your  outcome  has  two  categories,  such  as  gender  [male;  female]  or  survival  status  [dead;

alive],  then  you  should  use  the**binary  logistic  regression**.

If  your  outcome  has  more  than  two  categories  you  need  to  decide  whether  there  is  an  order  to  the

categories.  If  there  is,  such  as  when  your  outcome  has  categories  of  size  [small;  medium;  large]  or

grade  [A;  B;  C;  D;  E],  then  you  should  use**ordinal  logistic  regression**.

Alternatively  your  categories  will  have  no  order,  such  as  favourite  colour  [red;  green;  blue]  or

religion  [Hindu,  Buddhist,  Jedi],  and  you  should  use  a**nominal  logistic  regression**  (aka  multinomial

logistic  regression).

**Multivariate  Linear  Regressions**  (aka  multiple  linear  regressions)  are  used  when  your  outcome

variable  is  continuous,  such  as  height,  weight  or  other  such  measurement.

When  you  have  time-to-event  data,  such  as  that  in  a  Kaplan-Meier  analysis,  then  you  should  use

**Cox’s  proportional  hazards  survival  analysis**.

In  each  of  these  cases,  you  can  use  binary,  ordinal,  interval  and/or  ratio  data  as  predictor  variables.

Where  you  have  nominal  data,  it  may  be  useful  to  arrange  them  into  separate  variables,  as  in

Chapter  2.

***5.1.1  Binary  Logistic  Regression***

Binary  logistic  regression  is  a  special  case  of  both  ordinal  and  nominal  logistic  regressions.  If  your

outcome  variable  has  two  categories  it  doesn’t  matter  whether  you  choose  a  binary,  ordinal  or

nominal  logistic  regression,  the  result  will  be  just  the  same.  In  a  sense,  then,  you  might  conclude  that

binary  logistic  regression  is  redundant,  and  you  might  be  right,  but  nevertheless  it’s  a  good  starting

point.

Let’s  take  a  simple  example  to  start  us  off  on  our  multivariate  journey.

Say  you  wish  to  find  out  whether  smoking  and/or  weight  are  related  to  resting  pulse.  Your  variables

have  the  following  properties:





Outcome:

o      Resting  Pulse:  binary,  [high;  low]

Predictors:

o      Smokes:  binary,  [yes;  no]

o      Weight:  continuous,  measured  in  lb

53

|  |  |  |
| --- | --- | --- |
| **Variable** | **Value** | **Count** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | **Coef** | **SE  Coef** | **Z** | **P** | **OR** | **L95** | **U95** |

|  |  |  |
| --- | --- | --- |
| **G** | **df** | **P** |

Among  the  myriad  of  results  that  a  typical  commercial  stats  program  will  spit  out  at  you  (yes,  I  know,

it  can  be  very  confusing  and  immensely  frustrating),  you  should  be  able  to  find  something  that  looks

a  bit  like  Figure  5.1.

Response Information:

High

Resting  Pulse

Low  (Reference)

**Total**

22

70

**92**

Binary Logistic Regression Table:

**Smokes**

**Constant**

-1.987

1.679

-1.18

0.237

**Weight**

-1.192

0.025

0.552

0.012

-2.16

2.04

0.031

0.041

0.30

1.03

0.10

1.00

0.90

1.05

Model Statistics:

**Figure  5.1:  Typical  Results  of  a  Binary  Logistic  Regression**

7.574

2

0.023

The  first  thing  to  look  at  is  whether  your  variables  are  significantly  and  independently  related  to  the

outcome.  For  this,  you  look  to  the  p-values  in  the  table.

In  this  case,  both  smoking  and  weight  have  p-values  <  0.05,  so  they  are  both  independently  related

to  the  outcome  Resting  Pulse.

The  next  thing  is  to  notice  the  sign  of  the  coefficient.  The  coefficient  is  a  measure  of  effect  size  and

the  sign  tells  you  whether  the  variable  is  positively  or  negatively  associated  with  the  outcome.

Also  note  that  you  use  the  coefficient  to  calculate  the  Odds  Ratio  (OR)  with  the  formula:

54

Here,  smoking  is  negatively  associated  with  low  resting  pulse.  That’s  a  bit  awkward  to  wrap  the  brain

cells  around,  so  let’s  flip  it  around  –  smoking  is  positively  associated  with  high  resting  pulse.  That

makes  more  sense.

In  terms  of  the  Odds  Ratio,  we  can  invert  it,  so

.  This  means  that  smokers  are  more

than  3  times  as  likely  as  non-smokers  to  have  a  higher  resting  pulse.

Similarly,  weight  is  positively  associated  with  a  low  resting  pulse,  but  the  coefficient  is  small  (close  to

zero)  and  the  Odds  Ratio  is  also  small  (close  to  one),  so  weight  has  only  a  small  influence  on  resting

pulse.  In  other  words,  a  large  change  in  weight  is  needed  for  a  significant  decrease  in  resting  pulse.

The  Constant  is  only  important  when  creating  a  predictive  model  with  your  result  (more  on  that  in

Section  5.5),  and  while  you  will  use  the  coefficient  the  p-value  is  not  important  and  can  be  ignored.

Final  point:  the  Standard  Error  of  the  Coefficient  is  used  to  calculate  L95  and  U95  –  the  Lower  and

Upper  95%  Confidence  Intervals  (I’m  leaving  the  equations  out  here  –  there  are  plenty  of  other  texts

that  include  them  if  you  really  need  them).  The  confidence  intervals  give  you  the  range  of  values

that  you  can  be  95%  certain  contains  the  true  odds  ratio.

A  tight  range  gives  you  high  confidence,  while  a  wide  range  gives  you  low  confidence.  A  wide

confidence  interval  is  usually  an  indication  that  your  sample  size  is  too  small  for  your  analysis.

If  your  confidence  intervals  span  the  number  one,  then  you  cannot  be  certain  that  your  variable  will

have  any  effect  on  the  outcome  and  you  should  reject  this  result,  even  if  it  is  statistically  significant.

Looking  at  the  model  statistics  below  the  table  you  will  see  a  separate  p-value.  This  is  the  p-value  for

the  whole  analysis  and  tells  you  if  the  model,  in  other  words  the  whole  result,  is  statistically

significant.

The  model  p-value  is  useful  in  comparing  alternative  models.

What  happens  to  this  p-value  if  you  add  or  eliminate  a  variable?  If  the  p-value  becomes  smaller  then

you  have  a  more  accurate  model.

The  important  outputs  that  you  will  need  when  you  run  a  binary  logistic  regression  analysis  are:





**Statistic**

o **individual  p-values**

o **model  p-value**

**Effect  Size**

55

o

o

o

**coefficients**

**odds  ratios**

**confidence  intervals**

**The  individual  p-values**  are  important  because  they  tell  you  when  your  variables  are  significantly

and  independently  related  to  the  outcome.

**The  model  p-value**  is  important  because  it  tells  you  whether  the  model  is  statistically  significant  and

allows  you  to  compare  alternative  models  by  the  addition  and  elimination  of  variables

**The  coefficients**  are  important  because  they  tell  you  how  strongly  the  predictor  variable  is  related  to

the  outcome,  and  whether  the  relationship  is  positive  (an  increase  in  the  predictor  variable  results  in

an  increase  in  the  outcome)  or  negative  (an  increase  in  the  predictor  variable  results  in  a  decrease  in

the  outcome).

**The  odds  ratio**  is  important  because  it  tells  you  by  how  much  the  likelihood  of  the  outcome  changes

with  a  one-unit  change  in  the  predictor  variable.

**The  confidence  intervals**  are  important  because  they  give  you  a  sense  of  how  confident  you  should

be  in  the  odds  ratio.  If  your  95%  confidence  intervals  are  very  wide,  you  might  wish  to  report  90%

confidence  intervals  instead,  or  99%  confidence  intervals  when  they  are  very  narrow.

I  mentioned  earlier  that  a  disadvantage  of  multivariate  analysis  is  that  they  don’t  give  you  a  good

‘feel’  for  your  data.  Take  another  look  at  the  results  in  Table  5.1  above.  Do  you  get  a  good  sense  of

the  mean  value  of  weight?  The  maximum  and  minimum  values?  What  about  how  many  people

smoke?

See  what  I  mean?  Multivariate  analyses  do  have  their  advantages,  but  if  you  want  to  get  a  good  feel

for  what’s  going  on  in  your  data,  you’ll  need  to  do  other  stuff,  like  descriptive  statistics,  graphs,

tables  and  univariate  analysis.

NEVER  jump  straight  into  multivariate  analysis  without  getting  your  hands  dirty  first,  otherwise  I

guarantee  you’ll  regret  it  (been  there,  got  a  truckload  full  of  T-shirts!).

56

***5.1.2  Ordinal  Logistic  Regression***

Ordinal  logistic  regression  is  used  when  your  outcome  variable  is  categorical,  has  more  than  two

levels  and  there  is  some  sort  of  natural  order  to  the  levels  (but  not  necessarily  the  same  distance

between  them).

Let’s  look  at  an  example.

Say  you  wish  to  know  if  there  is  an  association  between  various  factors  on  the  survival  of  water

lizards.  You  split  your  survival  period  into  three  levels  1,  2  and  3  corresponding  to  short,  medium  and

long  term,  and  a  univariate  analysis  has  revealed  that  there  are  two  variables  –  toxicity  and  region  –

that  may  have  an  effect  on  survival.  Your  variables  have  the  following  properties:





Outcome:

o      Survival:  ordinal,  [1;  2;  3]

Predictors:

o Region:  binary,  [1;  2]

o Toxicity:  continuous

Your  results  would  look  something  like  Figure  5.2.

As  before,  look  at  the  p-values  of  the  predictor  variables  first.  Toxicity  is  independently  associated

with  survival  but  region  is  not.  Region  contributes  only  noise  to  the  model,  so  you  should  now

exclude  it  from  the  model  and  run  it  again.  This  would  likely  give  you  a  more  accurate  model  (see

Section  5.5  for  more  details  on  step-wise  modelling  techniques).

The  toxicity  coefficient  is  positive,  telling  you  that  there  is  a  positive  association  between  toxicity

and  survival;  that  higher  toxicity  levels  in  the  environment  tend  to  be  associated  with  lower  values  of

survival.  Note  that  survival  level  3  is  designated  as  the  reference,  so  the  outcome  survival  is

interpreted  as  the  opposite  of  what  is  expected.  If  survival  level  1  was  taken  as  the  reference  then

the  size  of  the  coefficient  would  have  been  precisely  the  same  but  would  have  instead  been  negative

(and  you  would  have  the  reciprocal  odds  ratio).

The  odds  ratio  tells  you  that  there  is  a  13%  increase  in  the  water  lizard  survival  odds  for  each  one

unit  increase  in  toxicity,  i.e.  a  13%  increase  in  odds  of  survival  for  survival  level  2  compared  to  3  or

level  1  compared  to  2.  Note  that  this  is  not  the  same  as  a  13%  increase  in  the  odds  of  absolute

survival  –  you  are  not  measuring  absolute  survival  here,  but  rather  categorising  survival  as  short,

medium  or  long.

57

|  |  |  |
| --- | --- | --- |
| **Variable** | **Value** | **Count** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | **Coef** | **SE  Coef** | **Z** | **P** | **OR** | **L95** | **U95** |

|  |  |  |
| --- | --- | --- |
| **G** | **df** | **P** |

Response Information:

2

Survival

1

3  (Reference)

**Total**

46

15

12

**73**

Binary Logistic Regression Table:

**Constant  (2)**

**Constant  (1)**

-3.724

7.244

1.687

1.879

-2.21

-3.86

0.027

<0.001

**Region**

**Toxicity**

0.201

0.121

0.496

0.034

0.41

3.56

0.685

<0.001

1.22

1.13

0.46

1.06

3.23

1.21

Model Statistics:

**Figure  5.2:  Typical  Results  of  an  Ordinal  Logistic  Regression**

14.713

2

0.001

No  doubt  you  will  have  noticed  that  there  are  two  constant  coefficients  here.  These  correspond  to

the  coefficient  for  survival  level  1  and  survival  level  2.  The  model  could  have  included  a  coefficient

for  survival  3,  but  since  this  is  the  reference,  its  value  would  be  1.000,  so  there’s  no  point  including

it.

Notice  how  similar  the  form  of  the  ordinal  logistic  regression  result  is  to  that  of  binary  logistic

regression.  It  doesn’t  take  a  great  leap  of  imagination  to  see  that  if  you  use  ordinal  logistic

regression  when  your  outcome  variable  has  two  categories,  then  the  result  will  be  precisely  that  of  a

binary  logistic  regression.  This  means  that  when  your  outcome  variable  is  ordinal,  then  it  doesn’t

matter  how  many  outcome  levels  there  are,  the  ordinal  logistic  regression  is  always  applicable,  even

if  there  are  just  two  outcome  levels.

58

***5.1.3  Nominal  Logistic  Regression***

Nominal  logistic  regression  is  used  when  your  outcome  variable  is  categorical,  has  more  than  two

levels  and  there  is  no  natural  order  to  the  levels.

Suppose  that  you  want  to  know  the  favourite  subject  of  a  group  of  school  kids,  and  your  preliminary

tests  have  identified  the  age  and  teaching  method  as  possible  contributing  factors  in  their  choice  of

favourite  subject.  You  data  have  the  following  properties:





Outcome:

o      Subject:  nominal,  [science;  maths;  art]

Predictors:

o Teaching  method:  binary,  [discussion;  lecture]

o Age:  continuous

Your  results  would  look  something  like  Figure  5.3.

This  time  there  are  two  groups  of  results;  a  group  comparing  maths  with  science,  and  a  group

comparing  art  with  science.

In  the  former  group,  according  to  the  p-values,  neither  age  nor  teaching  method  contributed  to  the

choice  of  students’  favourite  subject,  and  this  is  reflected  in  the  odds  ratio  95%  confidence  intervals,

which  span  unity.

However,  when  comparing  science  with  art,  the  teaching  method  did  make  a  significant  contribution

to  their  choice.  The  positive  coefficient  tells  you  that  students  given  lectures  rather  than  discussion-

type  teaching  tended  to  prefer  art  to  science  –  in  fact,  the  odds  ratio  suggests  they  are  almost  16

times  as  likely  to  choose  art  over  science  when  given  lectures  as  opposed  to  discussion.  Note  though

the  wide  span  of  the  confidence  intervals,  suggesting  that  perhaps  there  isn’t  sufficient  data  to  make

any  firm  conclusions.

Also,  take  a  look  at  the  p-value  for  age.  Although  non-significant  at  the  95%  level,  given  that  there

isn’t  much  data  available  it  would  be  reasonable  to  judge  these  data  at  the  90%  level.  These  data

suggest  that  a  larger-scale  study  might  provide  additional  insights.

An  alternative  way  to  have  done  these  analyses  would  have  been  to  create  binary  dummy  variables

for  the  nominal  variable  such  that  you  would  have  a  column  for  science  [1;  0],  one  for  art  [1;  0]  and

one  for  maths  [1;  0].  Then  instead  of  a  nominal  logistic  regression,  you  would  run  binary  or  ordinal

logistic  regressions.  This  would  have  the  advantage  that  you  would  be  comparing  each  of  the

nominal  levels  independently  against  all  of  the  rest,  rather  than  pair-wise  as  in  the  example  of  Figure

5.3.  It  all  depends  on  what  you’re  looking  for.  Do  you  want  a  pair-wise  comparison,  or  one  against

the  rest?

59

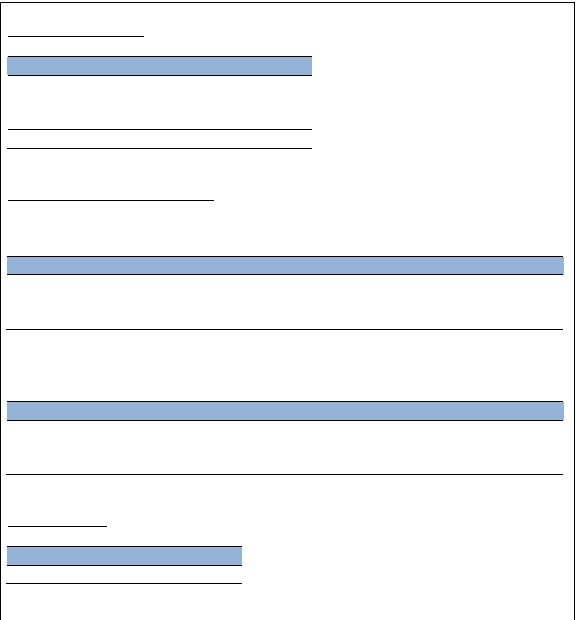
|  |  |  |
| --- | --- | --- |
| **Variable** | **Value** | **Count** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | **Coef** | **SE  Coef** | **Z** | **P** | **OR** | **L95** | **U95** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | **Coef** | **SE  Coef** | **Z** | **P** | **OR** | **L95** | **U95** |

|  |  |  |
| --- | --- | --- |
| **G** | **df** | **P** |

Response Information:



maths

Subject

Science  (Reference)

art

**Total**

11

10

9

**30**

Nominal Logistic Regression Table:

Logit  1:  (maths/science)

**Teaching  (lecture)**

**Constant**

-1.123

4.564

-0.25

0.806

**Age**

-0.563

0.125

0.938

0.401

-0.6

0.31

0.548

0.756

0.57

1.13

0.09

0.52

3.58

2.49

Logit  2:  (art/science)

**Teaching  (lecture)**

**Constant**

-13.849

7.243

-1.91

0.056

**Age**

2.77

1.014

1.372

0.584

2.02

1.73

0.044

0.083

15.96

2.76

1.08

0.88

234.91

8.66

Model Statistics:

**Figure  5.3:  Typical  Results  of  a  Nominal  Logistic  Regression**

12.825

4

0.012

***5.1.4  Multiple  Linear  Regression***

Multiple  linear  regression  is  used  when  your  outcome  variable  is  continuous  and  has  more  than  one

predictor  variable.  Predictor  variables  can  be  continuous  or  categorical.

Let’s  say  that  you  wish  to  assess  the  potential  influence  of  various  factors  on  systolic  blood  pressure.

Your  preliminary  investigation  has  narrowed  the  possibilities  to  age  and  weight.  Your  variables  have

the  following  properties:



Outcome:

o      Systolic  Blood  Pressure:  continuous,  in  mmHg

60

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | **Coef** | **SE  Coef** | **T** | **P** |

|  |  |  |
| --- | --- | --- |
| **S** | ***R2*** | ***2***  ***R*  (adj)** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **df** | **SS** | **MS** | **F** | **P** |



Predictors:

o Age:  continuous,  in  years

o Weight:  continuous,  in  pounds

Your  results  would  look  something  like  Figure  5.4.

Regression equation:

**Systolic  Blood  Pressure  =  30.990  +  (0.861  x  Age)  +  (0.335  x  Weight)**

Summary Table:

**Age**

**Constant**

**Weight**

0.861

30.990

0.335

0.248

11.94

0.131

3.47

2.59

2.56

0.008

0.032

0.034

Regression Statistics:

ANOVA:

2.318

**Error**

97.7%

97.1%

**Factor**

**Total**

8

2

10

42.99

1813.92

1856.91

5.37

906.96

**Figure  5.4:  Typical  Results  of  a  Multiple  Linear  Regression**

168.76 <0.001

The  p-values  tell  you  that  both  age  and  weight  are  significantly  and  independently  associated  with

systolic  blood  pressure,  and  that  systolic  blood  pressure  increases  with  both  age  and  weight

(positive  coefficients).

61

The  adjusted*R2*  value  (which  accounts  for  the  number  of  predictors  in  the  model)  tells  you  that  the

predictors  explain  97.1%  of  the  variance  in  systolic  blood  pressure  and  only  2.9%  remains

unexplained  –  in  other  words,  the  model  fits  the  data  extremely  well.

Note  that  there  aren’t  any  odds  ratios  in  the  table  here.  Odds  ratios  are  only  appropriate  when  your

outcome  variable  is  categorical.  Instead,  when  you  want  a  measure  of  effect  size  you  look  to  the

coefficients.  If  you  remember  back  to  Section  4.1  when  we  discussed  correlations,  you’ll  recall  that

one  of  the  important  measures  in  a  univariate  correlation  test  is  the  equation  of  the  line.  Well,  it’s

no  different  now  we’re  doing  the  multivariate  version.  The  equation  of  the  line  in  a  univariate  linear

regression  takes  the  form:

In  a  multivariate  linear  regression,  there  will  be  multiple  independent  predictors  and  coefficients,

like  this:

The  coefficients  are  a  measure  of  effect  size  and  in  this  multiple  linear  regression  example,  they  tell

you  by  how  much  the  systolic  blood  pressure  increases  for  a  single  unit  change  in  age  or  weight.

You  will  often  find  that  an  ANOVA  has  also  been  run  on  your  data  and  the  p-value  you  find  there  tells

you  how  significant  your  model  is,  allowing  you  to  compare  other  models  by  adding  and  eliminating

variables.

***5.1.5  Cox’s  Proportional  Hazards  Survival  Analysis***

Cox’s  Proportional  Hazards  Survival  Analysis  (aka  Cox’s  regression)  is  used  when  your  outcome  is  a

time-to-event  variable.

For  example,  let’s  say  you  wish  to  find  out  which  variables  contribute  to  survival  of  colorectal  cancer.

Results  of  Kaplan-Meier  survival  analysis  suggest  that  age,  weight  and  gender  may  each  play  a  part,

so  you  want  to  run  a  multivariate  analysis.  Cox’s  regression  is  the  multivariate  equivalent  of  the

Kaplan-Meier  analysis  and  is  the  appropriate  choice.  Your  variables  have  the  following  properties:



Outcome:

o      Survival  status:  binary  [dead;  alive]

o      Time  to  event:  continuous,  measured  in  years

62

|  |  |  |
| --- | --- | --- |
| **Variable** | **Value** | **Count** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | **Coef** | **SE  Coef** | **Z** | **P** | **HR** | **L95** | **U95** |

|  |  |  |
| --- | --- | --- |
| **G** | **df** | **P** |



Predictors:

o Age:  continuous,  measured  in  years

o Weight:  continuous,  measured  in  kg

o Gender:  binary  [male;  female]

Your  results  would  probably  look  a  bit  like  Figure  5.5.

Response Information:

0

Death

1  (Event)

**Total**

349

108

**457**

Cox’s Regression Table:

Weight

Age

Gender  (male)

0.055

0.112

0.074

0.017

0.039

0.188

3.24

2.91

0.39

0.032

0.004

0.694

1.06

1.12

1.08

1.02

1.04

0.75

1.09

1.21

1.56

Model Deviance:

**Figure  5.5:  Typical  Results  of  a  Cox’s  Regression**

12.57

3

0.006

The  p-values  tell  you  that  both  age  and  weight  are  significantly  and  independently  associated  with

death,  and  that  the  risk  of  death  increases  with  both  age  and  weight.  The  hazard  ratios  indicate  that

for  each  year  older  the  risk  of  death  to  the  patient  by  colorectal  cancer  increases  by  12%  and  for

each  additional  kg  the  patient  weighs  the  risk  of  death  also  increases  by  6%.

Here,  gender  is  not  independently  associated  with  death.

If  you  wish  to  accompany  your  Cox’s  regression  results  with  Kaplan-Meier  survival  plots  (and  their

hazard  ratios)  you  may.  It  would  be  appropriate  here  to  publish  Kaplan-Meier  survival  plots  of  age

and  weight  with  your  Cox’s  regression  results,  but  not  gender.  As  gender  is  not  independently

associated  with  survival,  you  can  conclude  that  the  relationship  between  gender  and  survival  is

63

dependent  upon  age  and/or  weight,  so  publishing  a  Kaplan-Meier  plot  of  gender  would  be  a

misrepresentation  of  the  data.

The  strength  of  the  overall  model  is  known  as  the  deviance  and  an  overall  p-value  is  calculated.

**5.2  Using  Multivariate  Tests  as  Univariate  Tests**

You  may  already  have  figured  this  out  by  now,  but  you  can  use  multivariate  analyses  as  univariate

analyses.  All  you  need  to  do  is  put  a  single  predictor  variable  in  your  model,  and  voilà  –  a  univariate

analysis.

The  result  you  get  will  be  the  same  as  (or  at  least,  very  similar  to)  the  result  you  would  get  from  the

equivalent  univariate  analysis  in  terms  of  the  p-value,  coefficients,  odds  ratio,  etc..

Some  researchers  prefer  to  use  multivariate  analysis  tools  for  their  univariate  analysis  before  going

on  to  do  a  full  multivariate  analysis  with  the  same  tools.  The  advantage  to  this  is  that  they  get

consistency  in  the  tools  they  use.

I  prefer  not  to  do  this,  though.  I*like*  using  different  tools  for  univariate  and  multivariate  analysis

mostly  because  the  univariate  tools  give  you  a  better  ‘feel’  for  your  data,  but  also  because  I  feel

more  confident  in  the  results  if  they  concur  and  have  been  computed  by  different  methods.

**5.3  Univariate  v  Multivariate  Analyses**

There  is  often  a  tension  between  the  results  of  univariate  and  multivariate  analysis  that  you  need  to

understand  a  bit  better,  questions  such  as:









Why  do  the  univariate  and  multivariate  results  differ?

What  does  it  mean  when  results  are  significant  in  univariate,  but  not  in  multivariate

analysis?

More  importantly,  what  does  it  mean  when  results  are  significant  in  multivariate,  but  not  in

univariate  analysis?

What  does  it  mean  when  there  are  2  variables  that  are  each  independently  associated  with

the  outcome,  but  are  also  independently  associated  with  each  other?

There  are  basically  3  effects  in  the  data  that  you  need  to  know  about  to  be  able  to  explain  why  your

univariate  and  multivariate  results  differ:







confounding  variables

suppressor  variables

interacting  variables

64

***5.3.1  Univariate  and  Multivariate  Results  that  Concur***

When  the  univariate  and  multivariate  results  concur,  then  we  are  living  in  happy  times.  Everything  is

just  as  it  should  be,  there  is  no  confusion  in  the  results  and  we  can  all  ride  off  into  the  sunset

knowing  that  we’ve  done  a  good  job.

How  often  does  this  happen?  Almost  never.  This  is  why  statisticians  are  often  grumpy.

***5.3.2  Confounding  Variables***

Confounding  variables  typically  account  for  results  that  are  significant  in  univariate  analysis  but  are

non-significant  in  multivariate  analysis.

OK,  let’s  cut  through  the  jargon.  What  does  confounding  mean?

Well  the  easiest  way  to  explain  is  to  look  at  an  example.  Let’s  have  a  look  at  a  classic  epidemiological

example  of  confounding.

Let’s  say  that  in  your  analysis  you  find  that  there  is  a  significant  relationship  between  people  who

have  lung  cancer  and  people  carrying  matches.

The  obvious  inference  is  that  matches  cause  lung  cancer,  but  we  all  know  that  this  explanation  is

probably  not  true.  So  we  need  to  look  deeper  into  the  dataset  to  see  what  is  responsible  for  this

result,  to  find  out  which  particular  variable  is  likely  to*confound*  (cause  surprise  or  confusion  in)  this

result.

Digging  deeper  we  find  that  smoking  is  significantly  associated  with  both  lung  cancer  and  carrying

matches,  Figure  5.6.

The  smoking  variable  now  causes  confusion  in  the  relationship  between  matches  and  lung  cancer,

confounding  our  initial  observations.  The  observation  that  there  is  a  relationship  between  matches

and  lung  cancer  has  become  distorted  because  both  are  associated  with  smoking.  In  this  example

smoking  is  called  the  confounding  variable.

Confounding  can  really  screw  up  your  understanding  of  the  underlying  story  of  the  data.  Without  an

understanding  of  confounding  variables  in  your  analysis  your  results  will  be  at  best  inaccurate  and

biased,  but  at  worst  completely  incorrect.

65

Matches



Lung  Cancer

Smoking

**Figure  5.6:  Illustration  of  Confounding**

Fortunately,  there  are  a  couple  of  analysis  techniques  that  you  can  use  to  control  for  confounding:





stratification

multivariate  analysis

**Stratification**  allows  you  to  test  for  relationships  within  the  strata  (layers)  of  the  confounding

variable.

For  example,  if  you  were  to  take  your  matches  and  lung  cancer  data  and  separate  them  into

separate  layers  (data  subsets),  one  for  ‘smoking’  and  the  other  for  ‘non-smoking’,  the  association

between  matches  and  lung  cancer  can  be  tested  within  the  smoking  population  and  within  the  non-

smoking  population  separately.

In  this  example,  you  will  likely  find  that  there  is  no  relationship  between  matches  and  lung  cancer  in

each  of  the  strata,  telling  you  that  smoking  is  what  is  related  to  lung  cancer  and  not  matches.

Stratification  has  the  drawback  though  of  diminishing  amounts  of  data  as  the  depth  of  your  strata

increase.  Imagine  performing  these  analyses  in  the  strata:











smoking

men  that  smoke

men  over  50  that  smoke

left-handed  men  over  50  that  smoke

left-handed  men  over  50  with  diabetes  that  smoke

To  these  factors,  we  can  also  add  in  fitness  level,  family  history,  body  mass  index,  and  a  whole  host

of  environmental  and  genetic  factors.  By  the  time  you  get  to  the  bottom  of  these  strata  there  will  be

very  few  samples  left  and  in  the  time  it’s  taken  you  to  do  the  analyses  you  will  likely  have  grown  old,

grey  and  very  frustrated!

66

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Non-**  **Smokers** | **Former**  **Smokers** | **Recent**  **Quitters** | **Persistent**  **Smokers** |
| **Relative**  **Risk  of**  **Death** | **Univariate**  **Results** | 1.0 | 1.08 | 0.56 | 0.74 |
| **Multivariate**  **Results** | 1.0 | 1.34 | 1.21 | 1.76 |

**Multivariate  analysis**  allows  you  to  test  for  relationships  while  simultaneously  assessing  the  impact

of  multiple  variables  on  the  outcome  without  having  to  limit  the  pool  of  data.

It  tells  you  of  the  various  risk  factors  and  their  relative  contribution  to  outcome,  and  gets  round  the

issue  of  confounding  by  adjusting  for  it.

Let’s  have  a  look  at  an  example  adapted  from  a  paper  by  Hasdai,  D.,*et  al*  “Effect  of  smoking  status

on  the  long-term  outcome  after  successful  coronary  revascularization”.  N.  Engl.  J.  Med.  336  (1997):

755-761.

They  ran  their  univariate  analysis  comparing  the  risk  of  death  for  four  different  cohorts;  non-

smokers,  former  smokers,  recent  quitters  and  persistent  smokers,  and  found  that  recent  quitters

and  persistent  smokers  had  a*decreased*  risk  of  death  compared  to  non-smokers.

What?  How  does  this  make  sense?  How  can  the  risk  of  death  be*lower*  for  persistent  smokers  than

non-smokers?

To  answer  this  question  they  then  ran  multivariate  analyses  taking  into  account  other  factors  such  as

age,  diabetes,  hypertension  and  other  factors.

The  results  comparing  the  univariate  and  multivariate  analyses  are  shown  in  Figure  5.7.

**Figure  5.7:  Comparison  of  Univariate*versus*  Multivariate  Results**

OK,  the  multivariate  analyses  make  a  lot  more  sense.  All  those  that  had  smoked  at  some  point  in

their  lives  had  a  higher  risk  of  death  than  did  non-smokers  and,  just  as  we  would  expect,  the

persistent  smokers  had  the  highest  risk  of  all.

So  what  went  wrong  with  the  univariate  analyses?  Well,  nothing  really.  It’s  just  that  the  univariate

analyses  were*incomplete*.  As  it  turned  out,  the  recent  quitters  and  persistent  smokers  were  younger

and  had  fewer  underlying  medical  problems  than  the  non-smokers  had,  and  these  factors

confounded  the  univariate  analyses,  biasing  the  univariate  results.

67

Hasdai  and  colleagues  could  have  done  a  stratified  analysis,  but  it  would  have  been  difficult  to

stratify  for  multiple  variables.  They  were  correct  in  choosing  to  run  a  multivariate  analysis.

***5.3.3  Suppressor  Variables***

When  your  results  are  significant  in  multivariate  analysis  but  non-significant  in  univariate  analysis,

then  you  are  most  likely  dealing  with  one  or  more  suppressor  variables  –  variables  that  increase  the

validity  of  another  variable  (or  set  of  variables)  by  its  inclusion  in  the  analysis.

Let’s  have  a  look  at  another  example.  Suppose  you  wish  to  see  if  Height  and  Weight  are  correlated

with  the  times  of  sprinters.  If  you  find  in  a  univariate  analysis  that  Weight  is  correlated  with  Time,

but  Height  is  not,  you  might  be  tempted  to  dismiss  the  influence  of  Height  on  race  times.  The

temptation  would  be  to  exclude  Height  from  the  multivariate  analysis  and  conclude  that  Weight  is

the  only  variable  that  is  independently  correlated  with  Time.

Now  let’s  assume  that  Height  and  Weight  are  correlated  with  each  other.  Some  of  the  variance  that

is  shared  between  Height  and  Weight  may  be  irrelevant  to  the  outcome  Time,  increasing  the  noise  in

the  analysis.  Having  both  Height  and  Weight  in  the  analysis  allows  the  multivariate  technique  to

identify  the  variance  that  is  irrelevant,  suppress  the  noise  and  thereby  lead  to  a  more  accurate

analysis.

This  is  why  it  is  called  a  suppressor  variable,  because  its  presence  suppresses  irrelevant  variance  on

the  outcome,  ridding  the  analysis  of  unwanted  noise.

***5.3.4  Interacting  Variables***

In  the  case  when  you  have  2  variables  that  are  each  independently  associated  with  the  outcome,  but

are  also  independently  associated  with  each  other,  it  is  highly  likely  that  these  variables  are

interacting  with  each  other  in  some  way,  modifying  the  effect  of  the  risk  on  the  outcome.

Let’s  go  back  to  the  confounding  example,  where  the  association  between  carrying  matches  and

lung  cancer  was  disproven  by  the  confounding  variable  ‘smoking’.  The  link  between  matches  and

lung  cancer  was  disproven  because  the  association  did  not  hold  up  in  either  of  the  smoking  and  non-

smoking  cohorts.

Now  suppose  that  the  association  between  matches  and  lung  cancer  holds  up  in  both  the  smoking

and  non-smoking  cohorts.  This  suggests  that  carrying  matches  and  smoking  each  have  a  part  to  play

in  whether  you  get  lung  cancer.

68

If  each  of  these  variables  plays  a  part,  then  there  may  be  an  additive  effect.  In  other  words,  there

would  be  a  higher  risk  of  contracting  lung  cancer  if  you  carried  matches  AND  smoked,  compared

with  carrying  matches  OR  smoked.

If  you  suspect  that  a  pair  of  variables  is  interacting  in  some  way,  then  you  should  test  this  out  by

creating  an  interaction  variable.

The  easiest  way  to  do  this  is  to  create  a  single  variable  that  represents  the  interaction.  Code

‘carrying  matches  AND  smokes’  as  1,  and  code  all  other  possibilities  as  0.

Then  add  this  into  your  multivariate  analysis*in  the  presence  of*  both  variables.

If  there  is  an  interaction  between  the  variables  then  the  interaction  variable  will  be  statistically

significant.  Conversely,  if  the  result  is  non-significant  then  there  is  not  likely  to  be  an  interaction

between  the  variables.

For  the  results  that  are  significant,  if  the  variables  act  synergistically,  then  the  correlation  coefficient

will  be  positive.  Conversely  if  the  relationship  between  the  variables  is  antagonistic  (one  works  to

decrease  the  effect  of  the  other),  then  the  correlation  coefficient  will  be  negative.

If  your  variables  interact,  the  nature  of  their  interaction  can  be  very  complex,  particularly  if  there  are

interactions  between  multiple  variables  and  it  can  be  very  difficult  to  figure  out  what  is  going  on.

Dealing  with  interacting  variables  is  best  left  to  a  specialist,  and  I’m  not  going  to  go  into  it  in  any

greater  detail  here.

If  you  suspect  that  you  have  interacting  variables,  I  would  suggest  you  seek  out  an  experienced

statistician  to  help  you.  Go  armed  with  a  smile  and  a  bottle  of  his  favourite  libation!

**5.4  Limitations  &  Assumptions  of  Multivariate  Analysis**

Although  each  individual  method  of  multivariate  analysis  has  its  own  assumptions  (discussed  at  the

relevant  point  in  the  text),  there  is  one  assumption  that  is  common  to  all,  and  that  is  the  assumption

of  linearity.

The  assumption  is  that  the  outcome  changes  linearly  with  each  predictor  variable.  If  the  predictor

variable  is  linear,  then  the  assumption  is  that  for  a  linear  change  in  the  predictor  variable  there  will

be  a  linear  change  in  the  outcome.  When  the  predictor  is  ordinal,  the  size  of  the  change  in  the

outcome  is  the  same  for  each  unit  change  in  the  predictor.

What  will  likely  change  is  the  scale  of  the  change,  depending  on  the  predictor  variable.  When

variable  A  has  a  greater  effect  on  the  outcome  than  variable  B,  then  a  one-unit  change  in  A  will  lead

to  a  greater  change  in  the  outcome  than  a  one-unit  change  in  B.  Each  predictor  variable  may  be

weighted  differently  (the  coefficients),  but  the  assumption  of  linearity  remains  the  same.

69

So  what  happens  when  one  of  our  predictor  variables  is  not  linearly  related  to  the  outcome?

Well,  if  you  plot  the  predictor  variable  against  the  outcome,  is  the  best-fit  line*mostly*  linear?  If  so,

then  it  may  be  sensible  to  continue  regardless.

On  the  other  hand,  if  the  best-fit  line  were  clearly  non-linear  then  one  way  forward  would  be  to

transform  either  the  predictor  or  outcome  variable.  If  the  outcome  variable  fails  a  standard

normality  test,  then  it  may  be  sensible  to  transform  the  outcome,  alternatively  you  should  consider

transforming  the  predictor  variable.  Either  way,  you  should  repeat  the  univariate  analysis  once  your

variables  have  been  transformed.

Typical  transformations  include:











logarithm  (base  10)

natural  logarithm  (base  e)

square  root

square

reciprocal

There  are  many  other  more  complicated  transformations  and  I’m  not  going  to  delve  deeper  here.

**5.5  Step-Wise  Methods  of  Analysis**

When  you  put  all  predictor  variables  into  a  multivariate  analysis,  what  you’ll  typically  get  is  a  result

where  some  of  the  variables  are  significantly  associated  with  the  outcome  and  some  are  not.  Those

that  are  not  significantly  associated  with  the  outcome  are  adding  noise  into  the  model,  which

obscures  the  true  signal.  If  you  want  to  get  a  clearer  picture,  you  want  more  of  the  signal  and  less  of

the  noise.

If  you  identify  the  most  non-significant  predictor  variable,  exclude  it  and  re-run  the  analysis  you  will

most  likely  remove  some  of  the  noise,  allowing  the  signal  to  come  through  more  strongly.

Conversely,  adding  in  an  additional  predictor  that  is  significantly  associated  with  the  outcome  will

add  to  the  signal  and  improve  the  signal-to-noise  ratio  in  the  model.

Step-wise  methods  of  analysis  are  automatic  procedures  (although  you  can  do  them  manually  using

most  standard  stats  programs)  that  tell  you  which  predictor  variables  should  be  in  your  multivariate

analysis  model.

These  variable  selection  techniques  have  been  much  criticised  as  a  modelling  strategy,  and  rightly  so

–  they  have  been  much  misused  for  decades.  That  doesn’t  mean,  though,  that  you  can’t  or  shouldn’t

70

use  them.  There  are  various  dangers  in  using  step-wise  procedures,  but  if  you  know  them,  you  can

account  for  them  in  your  analysis  to  build  a  strong  model  and  find  all  the  predictor  variables  that  are

independently  related  to  your  hypothesis  variable.

There  are  3  ways  to  use  step-wise  procedures  in  multivariate  analysis:







forward  step-wise

backward  step-wise

hybrid  step-wise

***5.5.1  Forward  Step-Wise  Procedure***

With  the  forward  step-wise  technique,  you  start  with  no  predictor  variables  in  the  model  and  add

them  one-by-one,  like  this:

1. Start  with  the  univariate  results  and  order  the  variables  by  p-value  according  to  their

relationship  with  the  hypothesis  variable

2. Add  in  the  predictor  variable  with  the  lowest  p-value  and  run  the  model

3. If  the  predictor  variable  is  significantly  associated  with  the  outcome,  leave  it  in  the  model  for

the  next  run,  alternatively  kick  it  out

4. Add  in  the  predictor  variable  with  the  next  lowest  p-value  and  run  the  model

5. If  all  predictor  variables  are  significantly  associated  with  the  outcome,  leave  them  in  the

model  for  the  next  run,  alternatively  kick  out  the  non-significant  variables.  You  probably

won’t  be  able  to  predict  which  variables  to  kick  out  in  advance  because  the  multivariate

model  will  take  account  of  confounding  and  how  closely  related  are  the  predictor  variables

with  each  other  –  the  effects  that  univariate  analysis  didn’t  take  account  of!

6. Continue  steps  4-5  until  there  are  no  more  predictor  variables  that  remain  un-analysed

This  procedure  is  good  for  when  you  have  a  small  dataset,  although  in  themselves  small  datasets

may  well  give  you  spurious  results  in  step-wise  analysis.  Tread  carefully  folks!

A  disadvantage  is  that  suppressor  variables  might  be  eliminated  from  the  model  before  they’ve  had

the  chance  to  pair-up  with  their  partners-in-crime.

***5.5.2  Backward  Step-Wise  Procedure***

With  this  technique,  you  start  with  all  your  predictor  variables  in  the  model  and  kick  them  out  one-

by-one,  like  this:

1. Add  in  all  the  predictor  variables.  Typically,  you  would  leave  out  any  variables  that  were  not

significantly  associated  with  your  hypothesis  variable  from  results  of  univariate  analysis.  If

this  makes  you  nervous  (and  it  probably  should),  then  you  could  include  all  those  variables

that  had  a  p-value  <  0.10.  Run  the  model

71

2. Eject  from  the  model  the  predictor  variable  with  the  highest  non-significant  p-value  &  re-run

the  model

3. Continue  step  2  until  all  non-significant  predictor  variables  have  been  kicked  out

This  procedure  gives  suppressor  variables  the  chance  to  do  their  thing,  but  has  the  significant

drawback  in  that  the  early  models  containing  all  predictor  variables  will  have  a  lot  of  noise  in  them

and  there  is  a  real  danger  of  eliminating  an  important  variable  from  the  model.

***5.5.3  Hybrid  Step-Wise  Procedure***

Hybrid  procedures  typically  use  a  combination  of  backward  and  forward  steps  to  try  to  get  round  the

limitations  of  the  backward  and  forward  techniques.

If  there  are  predictor  variables  that  you  know*should*  be  in  the  model,  then  you  can  usually  tell  the

modelling  software  to  start  with  those  variables  in  the  model  and  not  eject  them,  even  if  they  are

non-significant.

Doing  this,  you  can  already  see  that  you  don’t  start  with  an  ‘all  or  none’  model,  but  from  somewhere

in-between,  something  like  this:

1.

2.

3.

4.

5.

Start  with  a  core  of  user-selected  predictor  variables

Add  a  forward  step  by  adding  a  new  predictor  variable  into  the  model

Eliminate  (backward  step)  until  there  are  no  non-significant  variables  in  the  model

Add  in  a  further  new  predictor  variable  and  re-run

Continue  steps  2-4  until  all  predictor  variables  have  been  accounted  for  and  there  are  only

significant  associations  left  in  the  model

There  are  as  many  variations  of  this  as  you  can  dream  of,  and  many  more  additional  checking  steps

you  can  add  in.  For  example,  once  the  model  is  complete  you  might  want  to  have  another  look  at

the  non-significant  predictor  variables  to  see  if  any  of  these  might  act  as  suppressor  variables.

Adding  them  sequentially  to  your  final  model  to  check  their  effect  might  just  give  you  greater

insights  into  their  effect  on  your  model.

***5.5.4  Practical  Tips  for  Step-Wise  Procedures***

If  you  read  around  the  subject  (yeah,  I  know,  I  really  should  get  a  life!)  you’ll  see  all  sorts  of  rules  of

thumb  to  tell  you  how  many  samples  you  should  have  in  a  multivariate  analysis  compared  to  the

number  of  predictor  variables.  The  truth  is  that  none  of  them  are  based  in  fact.

The  bottom  line  is  that  it  depends  on  the  size  of  the  effects  you’re  measuring  and  how  precisely  you

wish  to  measure  them.  Large  effects  can  be  detected  with  relatively  small  amounts  of  data,  while

small  effects  need  large  amounts  of  data.  Some  of  your  predictor  variables  might  have  large  effects

72

on  the  outcome,  whereas  others  may  simultaneously  have  small  effects.  It’s  complex  and  there  is  no

one-size-fits-all  solution  to  determining  sample  size  for  multivariate  analysis.

If  you  suspect  that  your  sample  size  is  small  compared  to  the  number  of  predictor  variables,  one  way

to  effectively  handle  your  hybrid  step-wise  procedure  is  to  separate  your  variables  into  pools.

For  example,  let’s  say  you  have  85  variables:

1. You  might  choose  to  have  no  more  than  10  predictor  variables  in  any  pool,  so  you’ll  have  10

variables  in  each  of  8  pools  and  a  ninth  pool  with  5  variables

2. You  can  then  run  your  favourite  step-wise  procedure  on  each  pool  until  all  non-significant

variables  have  been  eliminated  from  each  pool

3. The  survivors  go  through  to  round  2,  where  new  pools  are  created

4. Continue  steps  1-3  until  all  predictor  variables  have  been  accounted  for

5. Go  back  through  each  eliminated  variable  testing  to  see  if  adding  it  would  improve  the

model

An  issue  that  you’ll  come  across  when  playing  around  with  forwards  and  backwards  elimination

methods  is  that  they  don’t  always  arrive  at  the  same  result.  This  can  cause  serious  headaches  and

may  well  lead  you  to  think  that  the  method  doesn’t  work  and  it’s  more  hassle  than  it’s  worth.  I  think

it’s  fair  to  say  that  we’ve  all  thought  that  at  some  point,  in  fact,  most  of  us  have  thought  that  about

the  whole  of  stats  at  some  point  in  our  learning  curve!

When  both  your  backwards  and  forwards  analysis  leads  you  to  the  same  result,  then  you  have

earned  yourself  the  rest  of  the  day  off.  Relax,  have  a  margarita,  you  deserve  it  –  you  have  a  robust

model  that  you  can  take  to  the  bank.

On  the  other  hand,  when  they  don’t  then  you  still  have  work  to  do.  Most  likely,  there  will  be

suppressor  effects  or  confounders  that  were  present  in  the  backwards  elimination  technique  but

were  kicked  out  of  the  forwards  procedure.  Go  and  find  them.  Account  for  them.

There  are  many  people  who  would  advise  you  not  to  use  step-wise  procedures.  I’m  not  one  of  those

people.

My  advice  would  be  to  learn  how  to  use  multivariate  analysis  techniques  correctly  and  effectively  –

including  the  step-wise  procedures.

There  are  dangers  in  using  all  tools  incorrectly  and  the  best  solution  is  to  learn,  not  walk  away.

Using  your  knowledge  and  experience:



Is  your  hybrid  step-wise  procedure  not  taking  account  of  confounding  variables?

73





o      Then  re-model  your  procedure.

Is  a  suppressor  variable  being  eliminated  from  the  model?

o      Add  it  to  your  core  of  non-eliminable  variables

Are  there  interaction  variables  that  are  not  being  taken  account  of?

o Create  them  and  add  them  into  the  model

There  are  no  rules  to  say  that  the  hybrid  methodology  should  be  completely  automated,  and

nothing  to  stop  you  from  adjusting  the  method  and  going  again.

Experiment,  learn,  adjust,  re-run.  Rinse  and  repeat.

Whenever  there  is  something  itching  at  the  back  of  your  brain  trying  to  tell  you  something  is  just  not

right,  don’t  ignore  it  –  scratch  it!  Find  out  what’s  going  on  and  deal  with  it.

I  guarantee  you’ll  learn  a  lot  and  end  up  with  more  accurate  models  than  blindly  accepting  the  first

result  –  and  far  better  than  walking  away  and  not  doing  it  at  all!

**5.6  Creating  Predictive  Models  with  the  Results  of  Multivariate  Analyses**

So  far,  we’ve  discussed  the  results  of  multivariate  analysis  in  terms  of  correlations  and  associations,

where  the  goal  is  to  understand  and  explain  the  relationships.  The  next  step  is  to  use  these  results  to

create  models  that  can  predict  the  outcome  variable  given  a  new  set  of  predictor  measurements.

It  may  sound  difficult,  but  it’s  really  not.  If  you’ve  run  your  multivariate  analysis  and  got  a  good,

sound  result,  then  you  have  all  you  need  to  be  able  to  make  a  good,  sound  prediction.

So  far  so  good,  but  notice  that  above  I  said  that  the  result  of  finding  the  significant  predictor

variables  in  a  multivariate  analysis  will  be  a*good,  sound  model*.  I  didn’t  say  it  would  be  the  best  or

optimum  model.  Because  the  requirements  of  the  multivariate  technique  are  different  when  finding

relationships  to  making  predictions  using  them,  your  modelling  method  needs  to  be  adjusted  to

account  for  that.

When  seeking  relationships  you  look  at  the  p-values  of  the  individual  predictor  variables  to  try  to

find  only  those  variables  that  are  significantly  and  independently  related  to  the  outcome.  In

determining  the  optimum  model,  you’re  seeking  to  find  the*optimum  model*,  not  the  constituent

parts.  To  find  the  optimum  model  you  need  to  assess  each  forwards  or  backwards  step  in  terms  of

the  model  p-value,  not  the  individual  p-values.

74

Whenever  the  addition  or  elimination  of  a  predictor  variable  decreases  the  model  p-value,  the

model  has  been  improved  and  has  greater  predictive  abilities  than  the  previous  model.

In  most  cases,  eliminating  non-significant  predictor  variables  decreases  the  model  p-value.  However,

there  are  times  when  the  elimination  or  addition  of  variables  has  the  opposite  effect  on  the  model

p-value  than  would  be  expected  by  checking  the  individual  p-value,  and  it  is  here  that  you  need  to

remember  the  purpose  of  the  analysis.

***5.6.1  Predictive  Models  with  Multiple  Linear  Regression***

It  all  starts  with  the  regression  equation,  which  for  a  Multiple  Linear  Regression  looks  like  this:

We  discussed  this  earlier,  you  take  a  new  sample  for  which  you  have  data  on  all  of  the  predictor

variables,  say  height,  weight,  age,  etc.,  multiply  each  of  the  predictor  variables  with  its  coefficient,

add  them  up  and  sum  in  the  constant  coefficient.  This  gives  you  a  prediction  for  your  outcome

variable.  If  you  already  had  a  measurement  for  your  outcome  variable,  but  withheld  it,  you  can

compare  the  predicted  outcome  with  the  actual  outcome  to  give  you  a  measure  of  the  error  of  the

model.

Let’s  have  another  look  at  the  example  of  systolic  blood  pressure  in  Figure  5.4.  The  regression

equation  for  this  was:

The  age  was  measured  in  years  and  weight  in  pounds,  so  a  50  year-old  weighing  160lb  is  predicted

to  have  a  systolic  blood  pressure  of:

Ideally,  systolic  blood  pressure  should  be  below  120mmHg,  so  according  to  the  model,  how  can  the

patient  reduce  their  SBP?  Obviously,  they  cannot  reduce  their  age,  so  they  must  lose  weight.  How

much  weight  should  they  lose  to  lower  their  SBP  to  120mmHg?  Re-arranging  the  equation,  the

answer  is  found  by:

75

The  patient  must  lower  their  weight  to  137lb,  which  is  a  loss  of  23lb.  It  looks  like  a  more  sensible  diet

and  regular  exercise  is  on  the  cards!

If,  rather  than  requesting  a  prediction  of  the  SBP  you  wished  to  find  the*change*  in  SBP  between

circumstances,  you  would  use  the  following:

This  has  the  advantage  that  you  can  get  a  prediction  for  a  change  in  outcome  with  a  change  in  any

predictor  variable  while  keeping  all  other  predictor  variables  constant.  In  this  equation,  this  means

that  any  variable  that  does  not  change  drops  out  of  the  equation  and  simplifies  the  calculation.

***5.6.2  Predictive  Models  with  Logistic  Regression***

OK,  what  about  logistic  regression?  Is  that  the  same  as  Multiple  Linear  Regression?

Well,  nearly.

In  logistic  regression,  the  outcome  variable  is  categorical,  so  instead  we  take  the  logarithm  of  the

odds  of  having  the  outcome  characteristic,  which  is  also  often  called  the  logit  of  p  in  other

textbooks.  For  example,  say  the  outcome  variable  is  lung  cancer  [yes;  no].  The  regression  of  the

predictor  variable  will  give  us  the  natural  logarithm  of  the  odds  of  having  lung  cancer.

Remember  that  the  odds  are  related  to  the  proportion  (or  probability)*p*  of  the  outcome*y*  by:

So,  the  form  of  a  logistic  regression  is  then:

76

Let’s  have  another  look  at  the  Resting  Pulse  example  of  Figure  5.1.  The  regression  equation  of  this

model  is  then:

For  a  non-smoker  weighing  120lb,  the  regression  equation  becomes:

Taking  the  exponent  of  both  sides  of  this  equation  gives:

and

So  someone  who  doesn’t  smoke  and  weighs  120lb  has  a  73%  chance  of  having  a  low  resting  pulse.

Now  try  plugging  Smoking  =  1  into  the  equation  along  with  Weight  =  120  and  see  what  happens.

Since  the  coefficient  for  smoking  is  negative,  this  should  decrease  their  chances  of  having  a  low

resting  pulse,  and  indeed  that  is  exactly  what  happens;  the  probability  drops  to  45.5%.

***5.6.3  Predictive  Models  with  Cox’s  Regression***

As  with  logistic  regression,  the  outcome  of  the  Cox’s  regression  is  a  natural  logarithm,  but  rather

than  being  based  on  odds,  it  is  now  based  on  the  ratio  of  the  hazard  (the  probability  of  dying)

time*t*  to  the  baseline  hazard ,  like  this:

at

77

The  baseline  hazard  is  an  abstract  concept,  and  is  the  hazard  when  all  the  predictor  variables  are  set

to  zero,  such  as  age  =  0,  weight  =  0,  etc..  Obviously,  this  is  meaningless,  so  instead  we  can  choose

our  own  baseline  levels  and  compare  the  hazard  with  these,  like  this:

Wow,  this  is  starting  to  look  awkward.  How  did  we  get  here?  Well,  what  we  did  was  take  a  hazard

with  a  certain  set  of  conditions,  ,  and  set  up  the  ratio  of  the  hazard against  the  baseline  hazard.

We  did  the  same  again  with  a  different  set  of  conditions  (  )  and  took  the  difference  between  them.

This  had  the  effect  of  removing  the  baseline  hazard  and  giving  us  the  hazard  ratio  (HR)  between  a

pair  of  conditions  that  we  specified.

It  may  look  hard  to  work  with,  but  don’t  worry,  it’s  not  as  difficult  as  you  think.

Let’s  go  back  to  Figure  5.5  and  take  another  look  at  the  Cox’s  regression  results.

If  you  took  the  baseline  hazard  to  be  a  patient  aged  50  with  a  weight  of  60kg,  what  would  be  the

likelihood  of  death  for  a  patient  aged  55  weighing  65kg?  Plugging  these  values  into  the  above

equation  would  give  you:

and

This  means  that  a  patient  aged  55  with  a  weight  of  65kg  is  2.3  times  as  likely  to  die  as  a  50-year-old

patient  weighing  60kg  is.  Note  that  you  don’t  say*when*  the  patient  is  likely  to  die.  That’s  not  what

this  means.  The  hazard  ratio  is  the  relative  risk  of  a  death  occurring  at  any  given  time,  and  one  of  the

assumptions  is  that  anything  that  affects  the  hazard  does  so  by  the  same  ratio  at  all  times  (hence

why  it’s  called  the  proportional  hazards  model).

There  is  also  another,  simpler  way  of  calculating  this  hazard  ratio  –  by  using  the  hazard  ratios

reported  in  Figure  5.5.  Since  the  per-year  hazard  of  death  is  1.12,  the  five-year  hazard  is  calculated

as .  Similarly,  the  per-kg  death  hazard  is  1.06,  so  the  five-kg  hazard  is .  Multiply  these

together  and  what  do  you  get? .  Hazards  are  multiplicative!

This  concludes  your  statistical  toolbox.  It’s  now  time  to  put  it  all  together,  and  this  is  exactly  what

we’ll  discover  in  the  next  section.

78

**Section   3:**



**The   Story   of   Your   Data**

79

**Chapter  6:  Visualising  Your  Relationships**

In  Section  1  of  this  book  you  learnt  how  to  collect  and  clean  your  data,  classify  it  and  prepare  it  for

analysis,  and  in  Section  2  you  learnt  about  the  toolbox  of  statistical  tests  that  you  need  to  find  the

relationships  that  characterise  the  story  of  your  dataset.

In  this  chapter,  you  will  learn  how  to  pull  all  of  this  together  to  create  a  holistic  strategy  for

discovering  the  story  of  your  data,  using  univariate  and  multivariate  analysis  techniques  to  discover

which  variables  are  independently  related  and  using  visualisations  to  interrogate  these  relationships

in  an  intelligent,  comprehensive  and  inspiring  way.

**6.1:  A  Holistic  Strategy  to  Discover  Independent  Relationships**

So  far,  the  tools  that  you  have  in  your  statistical  toolbox  have  been  seen  as  stand-alone,  separate

entities,  and  most  statistics  textbooks  will  present  them  to  you  in  that  way.

The  truth  though,  is  that  statisticians  and  experienced  analysts  use  these  tools  in  combination  to

give  them  a  greater  perspective  on  what  the  data  is  trying  to  tell  them.  And  that  is  the  crux  of  the

matter  here  –  data  holds  information  hostage  to  the  extent  that  it  takes  specialist  tools  and  a  joined-

up  strategy  to  extract  the  information  and  do  something  meaningful  with  it  to  change  the  world,

hopefully  for  the  better.

This  is  your  mission,  should  you  choose  to  accept  it.  This  book  will  self-destruct  in  five  seconds…

***6.1.1  Univariate  -  Multivariate  Tag  Team***

The  typical  reaction  to  looking  at  a  whole  dataset  before  starting  any  analysis  is  to  start  to  have  a

panic  attack.  The  whole  thing  is  just  too  big  to  hold  in  your  head  at  once,  so  you  need  to  break  it

down  into  separate  tasks,  then  into  sub-tasks  and  sub-sub-tasks.

Let’s  start  from  the  beginning,  with  the  variable  that  you’re  most  interested  in.  We’ll  call  this

variable  A,  or  varA  for  short.  You  need  to  find  out  which  variables  are  related  to  varA.

Way  back  in  Chapter  3  we  established  that  univariate  statistics  can  tell  you  which  of  the  other

variables  in  your  dataset  are  related  to  varA,  but  they  do  so  pair-wise  without  correcting  for  the

influence  of  other  variables  in  the  dataset.  On  the  other  hand,  multivariate  statistics  can  tell  you

which  variables  are  independently  related  to  varA  –  which  is  exactly  what  you  want  –  but  they  are

very  data  hungry  and  are  sensitive  to  having  too  many  variables  in  the  model  at  once.

So  how  can  we  find  all  the  independent  relationships  with  varA  when  there  is  not  enough

information  going  into  univariate  tests  and  too  much  going  into  multivariate  stats?

80

You  need  a  holistic  strategy,  and  here  it  is:





Run  univariate  analyses  of  all  the  variables  on  varA

o      Eliminate  the  variables  with  non-significant  p-values

Run  multivariate  analyses  of  the  remaining  variables  with  varA

o The  significant  relationships  will  be  independently  related  to  varA

Essentially,  you  use  univariate  analyses  as  a  screening  tool  to  decide  which  of  the  relationships  are

*worthy  of  further  investigation*.  You’re  not  making  definitive  decisions  at  this  stage,  but  rather

narrowing  the  field  for  multivariate  analyses.  You’re  using  univariate  and  multivariate  stats  as  a  tag

team;  the  univariates  ‘soften  up’  the  data,  leaving  them  ready  for  multivariate  stats  to  go  in  for  the

kill.

There  are  four  different  result  possibilities  when  using  univariate  and  multivariate  statistics  in  this

way,  and  here  they  are,  with  explanations  of  why  you  should  accept  or  reject  these  results:

**Non-significance  in  both  univariate  &  multivariate  stats:  Reject  the  relationship**

A  relationship  that  is  non-significant  in  both  univariate  and  multivariate  analysis  is  unlikely  to  be

incorrect*in  this  dataset*,  so  you  can  safely  reject  this  relationship  as  having  no  evidence  to  support

its  existence.  That  doesn’t  mean  that  the  relationship  doesn’t  exist  in  reality  –  your  dataset  may  not

have  sufficient  data  to  capture  the  complexities  of  the  relationship,  in  which  case  you’ll  need  more

data  or  a  bigger  dataset.

**Significance  in  both  univariate  &  multivariate  stats:  Accept  the  relationship**

Similarly,  a  relationship  that  is  significant  in  both  univariate  and  multivariate  analysis  is  also  unlikely

to  be  incorrect*in  this  dataset*,  so  you  can  safely  accept  this  relationship  as  having  some  basis  in

reality.  Again,  that  doesn’t  mean  that  the  relationship  is  correct  in  the  real  world  –  it  may  have

occurred  by  chance,  and  if  you  suspect  that  to  be  the  case  then  you’ll  need  to  run  your  project  again

with  a  new  dataset  to  be  able  to  confirm/deny  this.

**Significance  in  univariate  stats  &  non-significance  in  multivariate  stats:  Reject  the  relationship**

What  about  when  the  relationship  is  significant  in  univariate  analysis  but  non-significant  in

multivariate  analysis?  In  this  case,  we  can  conclude  that  the  relationship  is  not  independent,  but

rather  is  dependent  upon  one  or  more  confounding  variables.  Since  you  are  only  seeking

independent  relationships,  you  can  reject  this  result.

81

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Univariate** | |
| **Non-Significant** | **Significant** |
| **Multivariate** | **Non-Significant** | Reject | Reject |
| **Significant** | Reject | Accept |

**Non-significance  in  univariate  stats  &  significance  in  multivariate  stats:  Reject  the  relationship**

So  what  do  we  do  when  the  relationship  is  non-significant  in  univariate  analysis  but  significant  in

multivariate  analysis?  In  this  case,  most  likely  a  suppressor  variable  is  elevating  the  status  of  a

relationship  from  being  non-significant  to  significant.  Although  suppressor  variables  are  important

when  trying  to  find  the  optimum*model*,  here  you  are  seeking  only  the  independent  relationships

with  varA.  If  varX  is  dependent  upon  a  suppressor  variable  varY  for  its  relationship  with  varA  then  it

is  not  independent  and  should  be  rejected.  There  is  therefore  no  need  to  seek  suppressor  variables

in  the  holistic  relationship  model.

However,  if  what  you  seek  is  the  optimum  model  of  outcome  variable  varA,  then  suppressor

variables  are  important.  They  must  be  identified  and  accounted  for.

My  advice  here  is  to  initially  reject  non-significant  relationships  at  the  univariate  analysis  stage  and

then  do  a  post-hoc  analysis  after  the  multivariate  analyses  if  you  suspect  that  there  are  suppression

effects.  It  can  be  fiendishly  difficult  trying  to  untangle  the  effects  of  suppressor  variables,  and  if

you’re  not  comfortable  (and  I  wouldn’t  be  at  all  surprised)  then  seek  out  an  experienced  statistician

to  help  you.

The  bottom  line  here  is  that  when  seeking  those  variables  that  are  independently  related  to  your

target  variable  varA,  you  first  reject  all  those  variables  that  are  not  significantly  related  to  varA  in  a

univariate  analysis.  Of  those  that  are  significant  in  univariate  analysis,  you  reject  all  those  that  are

not  related  in  a  multivariate  analysis.  The  remaining  variables  are  those  that  are  independently

related  to  varA,  which  is  precisely  what  you’re  looking  for.  Figure  6.1  will  help  you  decide  whether  to

accept  or  reject  relationships  in  the  holistic  relationship  model.

**Figure  6.1:  Hypothesis  Testing  in  the  Holistic  Relationship  Model**

***6.1.2  Tag  Team  Strategy  for  a  Continuous  Variable***

Let’s  say  your  target  variable,  varA  is  continuous.  Right  now,  you  have  no  idea  which  of  the  other

variables  in  your  dataset  is  related  to  varA.  You’ll  probably  have  a  hunch,  you’ll  have  some  prior

knowledge  from  previous  analysis,  textbooks  and  research  publications,  but  these  are  all

relationships  that  were  found  in*other*  datasets.  You  don’t  know  what  you’ll  find  in*your*  dataset.

82

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | **Outcome  Variable** | |
| **Continuous**  **(Normal)** | **Continuous**  **(Non-Normal)** |
|  |  | **Continuous**  **(Normal)** | **Pearson  Correlation**  Scatter  plot  Correlation  Coefficient  Equation  of  the  Line  p-value | **Spearman  Correlation**  Scatter  plot  Correlation  Coefficient  Equation  of  the  Line  p-value |
| **Continuous**  **(Non-Normal)** | **Spearman  Correlation**  Scatter  plot  Correlation  Coefficient  Equation  of  the  Line  p-value | **Spearman  Correlation**  Scatter  plot  Correlation  Coefficient  Equation  of  the  Line  p-value |
| **Ordinal**  **or**  **Nominal**  **(2  categories)** | **2-sample  t-test**  Box-and-Whiskers  plot  Group  means  SD  p-value | **Mann-Whitney  U-test**  Box-and-Whiskers  plot  Group  medians  Inter-quartile  range  p-value |
| **Ordinal**  **or**  **Nominal**  **(>2  categories)** | **ANOVA**  Box-and-Whiskers  plot  Group  means  SD  p-value | **Kruskall-Wallis**  Box-and-Whiskers  plot  Group  medians  Inter-quartile  range  p-value |
|  | **Multiple**  **predictor**  **variables** | **Multiple  Linear  Regression**  Summary  table  Regression  equation  Regression  statistics  ANOVA | |

The  first  step  in  your  holistic  strategy  is  to  use  univariate  analyses  to  find  out  which  variables  are

**PredictorVariable(s)**

**Univariate**

**Multivariate**

related  to  varA.  Remember  that  these  won’t  necessarily  be*independent*  relationships,  and  that

you’re  using  univariate  analysis  as  a  screening  tool  to  narrow  the  field  to  those  variables  that*might*

be  independently  related  to  varA.

Taking  note  of  the  type  of  data  for  varA  (in  this  case,  continuous)  and  for  varX,  select  the  appropriate

univariate  statistical  test.  Figure  6.2  will  help  you  to  decide  correctly.

**Figure  6.2:  Choosing  Appropriate  Analyses  for  a  Continuous  Outcome  Variable**

83

For  example,  if  the  predictor  variable  varX  is  ordinal  and  has  two  categories  you  would  choose  the  2-

sample  t-test  or  Mann-Whitney  U-test,  depending  upon  whether  varA  is  normally  or  non-normally

distributed.

After  choosing  correctly,  run  the  test  and  inspect  your  p-value.  If  you  want  to  be  really  strict  then

you  should  reject  those  variables  for  which  p>0.05.  These  variables  will  not  carry  forward  to  a

multivariate  analysis.  If  you’re  a  bit  nervous  about  using  such  a  strict  p-value  cut-off,  then  you  can

use  one  that  is  a  little  more  lenient  such  as  p>0.10.

Continue  to  do  this  for  varX1,  varX2,  varX3,  etc.,  until  you  have  tested  for  all  possible  relationships

with  varA  in  univariate  analysis.

You’ll  then  have  a  list  of  possible  relationships  with  varA  that  you’ll  need  to  test  with  multivariate

analysis.  The  relationships  with  varA  could  look  something  like  Figure  6.3.

**A**

**B**

**D**

**H**

**J**

**Figure  6.3:  Possible  Relationships  with  varA  After  Univariate  Analysis**

Now  that  you  have  narrowed  down  the  list  of  possible  relationships,  you  can  use  multivariate

statistics  to  confirm  or  deny  these  relationships.  If  your  list  of  possible  relationships  is  small  you  can

jump  straight  into  multivariate  analysis,  otherwise  you  might  need  to  split  the  variables  into  pools,

as  detailed  in  Section  5.5.4.

Using  step-wise  techniques,  eliminate  non-significant  variables  one-by-one  until  you  are  left  with  a

core  of  variables  that  are  all  significantly  and  independently  related  with  the  outcome  variable  varA,

say  varB  and  varH,  Figure  6.4.

This  is  the  story  of  the  data  for  varA  in  this  dataset.  You  have  tested  it  with  univariate  analysis  then

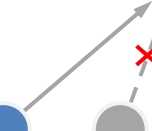
confirmed  it  with  multivariate  analysis.  You  can  be  confident  that  these  are  all  the  independent

relationships  with  varA  in  this  dataset  because  you  have  accounted  and  corrected  for  all

confounding  variables.

84

**A**



**B**

**D**

**H**

**J**

**Figure  6.4:  Relationships  with  varA  After  Multivariate  Analysis**

Variables  B  and  H  are  independently  associated  with  A,  while  D  and  J  are  not  –  they  are  dependent

upon  other  variables  (B  and  H)  for  their  relationship  with  A.  Variables  D  and  J  are  still  related  to  varA

and  may  still  exert  influence  (and  if  you  want  the  optimum  predictive  model,  then  you  may  need  to

have  them  in  your  final  multivariate  analysis),  but  their  effects  are  secondary  to  the  effects  of  B  and

H.

Note  the  directions  of  the  arrows.  One  of  the  powerful  aspects  of  multivariate  analysis  that  you

don’t  get  with  univariate  analysis  is  that  each  of  these  variables  are*predictors*  of  varA,  that  each  of

the  relationships  has  a  direction.  You  now  know  that  each  of  B  and  H  are  independent  predictors  of

A,  but  the  opposite  is  not  necessarily  true  –  you  haven’t  tested  for  the  opposite  yet.

***6.1.3  Tag  Team  Strategy  for  a  Categorical  Variable***

So  now  you  have  the  story  of  your  data  for  variable  A,  let’s  look  at  the  first  of  the  variables  that  are

independently  related  to  it,  varB,  and  let’s  suppose  that  varB  has  categorical  data.

The  strategy  here  is  exactly  the  same  as  before;  use  univariate  analyses  to  filter  out  the  variables

that  are  not  related  to  varB,  and  then  confirm/deny  the  relationships  with  multivariate  analysis.

Figure  6.5  will  help  you  choose  the  appropriate  statistical  tests.

Let’s  say  that  your  predictor  variable  varX  is  nominal  and  has  2  categories  (such  as  Gender  [male;

female]),  you  would  choose  the  Fisher’s  Exact  Test  or  Chi-Squared  Test,  depending  on  whether  your

outcome  variable  has  two  or  more  than  two  categories.

Run  the  appropriate  relationship  test  and  filter  at  your  chosen  p-value  cut-off,  and  do  this  for  varX1,

varX2,  etc.,  until  all  possible  relationships  have  been  screened.  Then  you  can  run  multivariate

analyses  on  the  variables  that  survive  the  cut.

85

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | **Outcome  Variable** | |
| **Categorical**  **(2  categories)** | **Categorical**  **(>2  categories)** |
|  |  | **Continuous**  **(Normal)** | **2-sample  t-test**  Box-and-Whiskers  plot  Group  means  SD  p-value | **ANOVA**  Box-and-Whiskers  plot  Group  means  SD  p-value |
| **Continuous**  **(Non-Normal)** | **Mann-Whitney  U-test**  Box-and-Whiskers  plot  Group  medians  Inter-quartile  range  p-value | **Kruskall-Wallis**  Box-and-Whiskers  plot  Group  medians  Inter-quartile  range  p-value |
| **Ordinal**  **or**  **Nominal**  **(2  categories)** | **Fisher’s  Exact  Test**  2x2  contingency  table  Chi-Squared  cell  values  Odds  Ratio  p-value | **Chi-Squared  Test**  nx2  contingency  table  Chi-Squared  cell  values  Chi-Squared  Value  p-value |
| **Ordinal**  **(>2  categories)** | **Chi-Squared  for  Trend  Test**  2xn  contingency  table  Scatter  plot  of  proportions  Equation  of  the  line  p-value | **Chi-Squared  Test**  nx2  contingency  table  Chi-Squared  cell  values  Chi-Squared  Value  p-value |
| **Nominal**  **(>2  categories)** | **Chi-Squared  Test**  nx2  contingency  table  Chi-Squared  cell  values  Chi-Squared  Value  p-value | **Chi-Squared  Test**  nx2  contingency  table  Chi-Squared  cell  values  Chi-Squared  Value  p-value |
|  | **Multiple**  **predictor**  **variables** | **Ordinal  or  Nominal  Logistic  Regression**  Logistic  regression  table(s)  Coefficients  +  SE  Coefficients  Odds  Ratios  +  95%  confidence  intervals  p-values  Model  statistics | |

**Figure  6.5:  Choosing  Appropriate  Analyses  for  a  Categorical  Outcome  Variable**

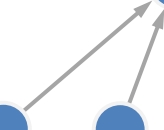
**PredictorVariable(s)**

**Univariate**

**Multivariate**

86

The  relationships  with  varB  would  then  look  like  Figure  6.6.



**B**

**A**

**D**

**F**

**M**

**Figure  6.6:  Relationships  with  varB  After  Multivariate  Analysis**

Figure  6.6  is  the  story  of  the  data  for  varB,  and  all  the  relationships  are  significant  and  independent

of  other  factors.

In  this  case,  note  that  varA  is  an  independent  predictor  of  varB.  In  Figure  6.4  varB  independently

predicted  varA,  so  the  prediction  is  mutual.  If  you  think  about  it  long  enough  you  will  realise  that

that  doesn’t  have  to  be  the  case,  though.  Lung  Cancer  is  a  subset  of  Cancer,  so  Lung  Cancer  data  will

always  predict  Cancer  (Lung  Cancer  ‘Yes’  always  leads  to  a  Cancer  ‘Yes’),  but  it’s  highly  unlikely  that

the  reverse  will  be  true  because  there  are  many  other  subsets  of  Cancer  (Cancer  ‘Yes’  will  lead  to

many  more  Lung  Cancer  ‘No’  than  ‘Yes’).  It’s  a  silly  comparison  to  make  because  it  wouldn’t  make

sense  to  test  for  a  relationship  between  Cancer  and  Lung  Cancer,  but  I  hope  it  illustrates  the  point

that  in  relationship  analysis,  prediction  is  not  necessarily  mutual.

***6.1.4  Tag  Team  Strategy  for  Survival  Variables***

For  survival  variables,  the  strategy  remains  the  same.  In  this  case,  the  univariate  and  multivariate

tests  are  Kaplan  Meier  analysis  (with  the  logrank  test  and  hazard  ratios)  and  Cox’s  regression,

respectively.  For  Kaplan-Meier  analysis,  the  predictor  variable  must  be  categorical.  If  you  have  a

continuous  predictor  variable  then  you  must  use  Cox’s  regression,  as  either  a  univariate  or

multivariate  test,  Figure  6.7

So  let’s  say  that  varC  is  your  survival  variable.  As  before,  use  univariate  analyses  (Kaplan-Meier  for

categorical  predictor  variables  and  Cox’s  regression  for  continuous  predictor  variables)  to  screen  out

the  variables  that  are  not  related  to  varC,  and  then  use  multivariate  analysis  to  confirm/deny  that

the  relationships  are  independent.

87

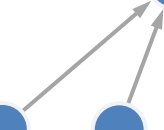
|  |  |  |  |
| --- | --- | --- | --- |
|  | | | **Outcome  Variable** |
| **Time-To-Event** |
|  |  | **Ordinal**  **or**  **Nominal** | **Logrank  Test**  Kaplan-Meier  Plot  Censor  Table  Hazard  Ratio(s)  p-value |
| **Continuous** | **Cox’s  Regression  Analysis**  Response  Information  Cox’s  Regression  Table  Model  Deviance  p-value |
|  | **Multiple**  **predictor**  **variables** | **Cox’s  Regression  Analysis**  Response  Information  Cox’s  Regression  Table  Model  Deviance  p-value |

**Figure  6.7:  Choosing  Appropriate  Analyses  for  a  Time-To-Event  Variable**

**PredictorVariable(s)**

**Univariate**

**Multivariate**



The  relationships  with  varC  might  then  look  like  Figure  6.8.

**C**

**A**

**E**

**J**

**M**

**Figure  6.8:  Relationships  with  varC  After  Multivariate  Analysis**

Here,  varA,  varE  and  varJ  are  all  independent  predictors  of  the  survival  variable  varC,  but  varM  is

dependent  upon  at  least  one  of  A,  E  and  J  for  its  relationship  with  C.

88

***6.1.5  Corrections  for  Multiple  Comparisons***

It’s  round  about  this  point  that  the  more  experienced  analyst  starts  jumping  up  and  down  and

getting  all  hot  under  the  collar  about  multiple  comparisons.

The  crux  of  this  argument  is  that  if  you  run  a  single  hypothesis  test  there  is  a  probability  of  1  in  20

that  a  significant  relationship  will  occur  by  chance.  That  is  to  say,  that  if  you  ran  20  hypothesis  tests,

then  on  average  one  will  be  incorrect.  For  100  tests,  5  will  be  incorrect  and  for  1000  tests,  50  will  be

incorrect.  The  argument  goes  that  the  more  hypothesis  tests  you  run  on  your  data  the  more  false

positives  you  will  get  in  your  result  set.

I’m  not  convinced  by  these  arguments.  While  the  mathematics  may  appear  sound,  they  tend  to

assume  that  chance  is  the  most  common  explanation  for  the  occurrence  of  a  relationship,  whereas

the  universe  is  governed  by  laws  of  nature  and  of  physics.  Most  relationships  are  found  because  they

have  a  basis  in  reality.

That  said,  let’s  look  at  the  issue  in  practical  terms.  You  compare  varA  with  varB.  Now  compare  varX

with  varY.  Does  the  comparison  between  X  and  Y  ‘know’  that  the  previous  comparison  has  been

made?  Does  it  care?  You  cannot  use  quantum  entanglement  to  explain  how  one  hypothesis  test

might  affect  another.

OK,  so  you’re  not  convinced  yet.

Let’s  say  you  have  a  hypothesis  about  a  particular  relationship  in  a  dataset,  you  do  your  analyses,

adjust  for  multiple  comparisons  and  publish  your  paper.  Later,  you  have  another  hypothesis  that  you

can  test  within  the  same  dataset.  In  this  new  set  of  analyses,  should  you  adjust  for  the  previous

multiple  comparisons  you  ran  as  well  as  the  new  ones?  Now  that  you’ve  run  some  new  hypothesis

tests,  should  you  go  back  to  your  previous  analysis  and  re-adjust  according  to  the  new  tests  you’ve

just  run?  Should  you  adjust  for*all*  past  tests  that  have  been  run  on  these  data?  If  you  adjust  for  all

past  tests,  then  surely  you  should  adjust  for  all*future*  tests  too!  What  if  you  have  a  dataset  that  is

available  to  all  researchers  in  your  department  –  should  you  adjust  for  all  past,  present  and  future

tests  made  by  your  fellow  researchers  as  well?

Starting  to  sound  a  little  ridiculous,  isn’t  it?

Well  let’s  go  a  bit  further.  Making  adjustments  for  multiple  comparisons  can  convince  you  to  run  an

insufficient  number  of  hypothesis  tests  on  your  data.  If  the  relationship  between  varA  and  varB  is

borderline  at  p  =  0.046,  then  any  further  tests  will  change  your  p-value  from  significant  to  non-

significant,  so  you  decide  not  to  look  for  confounding  variables  or  other  potentially  important

relationships  so  you  can  preserve  confidence  in  your  result.

This  really  isn’t  a  good  idea.  Why  should  the  fear  of  losing  a  potentially  important  discovery  stop  you

from  being  thorough  in  your  analyses?  It  shouldn’t.  Worse  still,  it  might  lead  to  you  not  making  other

important  discoveries  in  your  data.

89

Perhaps  the  best  way  to  account  for  false  positive  results  is  to  use  your  domain  knowledge  and

experience  to  ask  of  every  relationship  if  there  is  any  reason  to*reject*  the  result.  Typically,  whenever

an  unusual  result  occurs  the  researcher  will  try  to  find  a  reason  why  it  is  plausible  –  and  she  will

usually  succeed.  After  all,  we  all  love  a  winner,  don’t  we?  And  if  she’s  right,  she  can  publish  this  new

finding  and  enhance  her  career,  gain  new  funding  and  climb  onwards  and  upwards.

It  is  better  though  to  do  the  opposite  and  try  to  find  –  for  each  result  –  a  reason  why  it  is

implausible.  This  way  you  are  building  strong  controls  into  your  analyses  and  are  making  reasoned

judgements  that  give  a  stronger  basis  for  your  results  than  arbitrarily  adjusting  p-values.

If  you  still  believe  in  the  concept  of  adjusting  for  multiple  testing  (I  remain  far  from  convinced),  the

question  you  must  ask  yourself  is  this:

Is  it  better  to





Miss  nothing  real  than  to  control  the  number  of  false  positive  results?,  or

Control  the  probability  of  making  a  false  claim?

It  is  worth  noting  that  most  spurious  results  are  accounted  for  by  multivariate  analysis,  and  if  you

insist  on  making  adjustments  for  multiple  comparisons,  then  I  suggest  you  keep  it  simple  and  apply  a

Bonferroni  correction.

Basically,  a  Bonferroni  correction  penalises  you  for  the  number  of  pair-wise  comparisons  by

requiring  that  you  multiply  your  resultant  p-value  by  the  number  of  comparisons  made.  If,  after

making  this  adjustment  your  p-value  remains  <0.05,  then  you  may  reject  the  null  hypothesis  and

accept  that  the  relationship  is  significant.

In  the  holistic  strategy  outlined  here,  I  recommend  that  you  don’t  apply  any  corrective  factors

because  of  univariate  analyses,  since  the  multivariate  analyses  will  naturally  correct  for  these.

You  may  apply  corrective  factors  to  your  multivariate  analyses  if  you  wish,  and  I  recommend  that

you  correct  for  the  number  of  independent  predictors  remaining  in  the  final  model.

For  example,  for  varA,  if  there  are  five  independent  predictors  after  adjusting  for  all  other  factors,

then  multiply  each  of  the  p-values  by  five.  The  variables  that  have  adjusted  p-values  larger  than  0.05

will  then  be  rejected.  Moving  on  to  varB,  if  the  final  model  has  three  predictors,  then  multiply  each

of  the  p-values  by  three,  and  so  on.  Here,  it  doesn’t  make  sense  to  penalise  the  results  of  varB

because  of  the  results  of  varA,  so  I  see  no  reason  to  combine  the  penalties.

**6.2:  Visualising  the  Story  of  Your  Data**

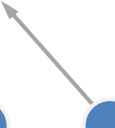
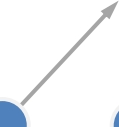
When  you  have  many  relationships  between  the  variables  in  your  dataset  it  can  be  difficult,  if  not

impossible,  to  hold  it  in  your  mind.  It’s  not  usually  difficult  to  visualise  the  story  of  the  data  for  varA,

after  all,  just  how  many  causal  (or*perceived*  causal)  connections  can  there  be  for  a  single  variable?

90

However,  when  you  pull  the  whole  thing  together  it  can  rapidly  get  away  from  you.  Here’s  where



you  need  to  be  able  to  build  a  visual  representation  of  all  the  relationships  in  the  dataset,  and  I’ll

present  a  couple  of  possibilities  to  you  here.

***6.2.1  Node-Link  Diagrams***

So  far,  you’ve  investigated  the  relationships  with  varA  and  you  have  the  story  of  the  data  for  varA

(Figure  6.4).  Similarly,  you’ve  done  the  same  for  varB  (Figure  6.6),  varC  (Figure  6.8)  and  so  on,  and

it’s  now  time  to  pull  the  whole  thing  together,  Figure  6.9.

**C**

**E**

**A**

**J**

**B**

**H**

**D**

**F**

**Figure  6.9:  Node-Link  Diagram  of  the  Relationships  With  varA,  varB  and  varC**

In  Figure  6.9,  I’ve  shown  only  the  results  for  varA,  varB  and  varC,  but  of  course,  the  bigger  picture

will  be  much  larger  than  this  –  we  haven’t  yet  considered  the  results  of  varD,  varE,  varF,  etc.,  but  you

91

get  the  idea.  You  take  the  story  of  A  and  make  the  connections  where  necessary  to  the  other

individual  stories  until  you  have  the  story  of  the  whole  dataset.

There  are  some  amazing  free  visualisation  programs  that  will  do  this  for  you  once  you  input  your

results.  Typically,  all  you  need  to  input  is  a  list  of  the  variable  names  and  the  relationships  and  the

software  does  the  rest.  There  are  also  many  customisation  options  in  node-link  diagram  programs,

such  as:







Adjust  the  thickness  of  links  to  signify  the  strength  of  the  relationship  (effect  size)

Adjust  the  size  of  the  nodes  to  signify  the  importance  in  the  network

Adjust  the  colour  of  nodes  to  signify  certain  data  groupings

Node-link  diagrams  are  a  great  non-linear  structure  of  the  visualisation  of  the  relationships  between

variables,  and  the  available  software  programs  often  allow  force-directed  algorithms  to  spatially

optimise  the  nodes  and  links  in  an  aesthetically  pleasing  conformation.  The  purpose  of  these

algorithms  is  to  position  the  nodes  in  space  so  that  all  the  links  (relationships)  are  of  more  or  less

equal  length  and  there  are  as  few  crossing  lines  as  possible.  By  assigning  forces  among  the  set  of

links  and  the  set  of  nodes,  based  on  their  relative  positions,  the  forces  can  be  used  to  minimise  the

energy  of  the  arrangement.

With  node-link  diagrams,  it  is  straightforward  to  zoom  in  to  inspect  detail,  zoom  out  to  see  the

bigger  picture,  select  individual  nodes  to  visualise  only  their  1st  and  2nd  level  connections  or  select

any  pair  of  nodes  to  inspect  the  possible  pathways  between  them.

However,  as  networks  get  large  and  highly  connected,  node-link  diagrams  often  devolve  into  giant

hairballs  of  line  crossings  and  can  become  practically  impossible  to  work  with.

***6.2.2  Correlation  Matrix***

With  node-link  diagrams,  only  the  significant  independent  relationships  are  displayed.  If  you  want  to

visualise*all*  the  relationships  in  your  dataset,  whether  they  are  significant  or  not,  independent  or

not,  the  correlation  matrix  is  probably  the  best  way  to  go.

The  correlation  matrix  is  a  square  array  with  the  variables  placed  along  horizontal  and  vertical  axes,

like  that  shown  in  Figure  6.10.

92

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Outcome  Variables** | | | | | | | | |
| **varA** | **varB** | **varC** | **varD** | **varE** | **varF** | **varG** | **varH** | **…** |
|  | **varA** |  |  |  |  |  |  |  |  | … |
| **varB** |  |  |  |  |  |  |  |  | … |
| **varC** |  |  |  |  |  |  |  |  | … |
| **varD** |  |  |  |  |  |  |  |  | … |
| **varE** |  |  |  |  |  |  |  |  | … |
| **varF** |  |  |  |  |  |  |  |  | … |
| **varG** |  |  |  |  |  |  |  |  | … |
| **varH** |  |  |  |  |  |  |  |  | … |
| **…** | … | … | … | … | … | … | … | … | … |

**Figure  6.10:  Correlation  Matrix  of  the  Relationships  in  the  Dataset**

**PredictorVariables**

The  outcome  variables  are  on  the  horizontal  axis,  while  the  predictor  variables  are  on  the  vertical

axis.  Each  column  is  a  list  of  all  the  relationships  between  the  predictor  variables  and  the  outcome

variable,  and  each  cell  of  the  matrix  represents  the  relationship  status  between  the  pair  of  variables.

In  this  example  the  status  of  the  relationship  is  represented  by  traffic  light  colours;  green  is

significant  (p<0.05,  and  marked  with  a    to  aid  those  with  red-green  colour  blindness),  red  is  non-

significant  (p≥0.10)  and  amber  tells  you  that  the  result  is  marginal  and  is  perhaps  worthy  of  further

investigation  (0.10<p≤0.05).  These  cut-off  values  can  be  adjusted  to  whatever  suits  your  purpose

and  the  colour  scheme  can  be  varied.

For  example,  we  established  in  Figure  6.8  that  varA,  varE  and  varJ  are  independent  predictors  of

varC.  This  is  reflected  in  the  column  of  varC:  varA  and  varE  are  coloured  in  green  (and  varJ  would

have  been  too  if  I’d  made  Figure  6.10  bigger!).  On  the  other  hand,  since  varC  is  time-to-event  data

we  can  only  analyse  these  data  as  an  outcome,  not  as  a  predictor,  so  the  entire  row  for  varC  is

greyed  out,  signifying  that  no  analyses  have  been  performed  or  were  not  possible.

Correlation  matrices  are  very  informative  because  you  can  instantly  see  which  of  the  variables  in  the

dataset  are  important  and  which  are  not,  for  example,  varG  has  no  significant  relationships  with  any

other  variables  in  the  dataset.

93

Although  for  simplicity  I’ve  shown  the  cells  here  as  being  colours  representing  relationship  status,

you  could  also  add  into  the  cells  the  p-values  giving  rise  to  these  relationships  or  perhaps  Odds

Ratios  to  signify  the  strength  of  the  relationship.

Correlation  matrices  have  an  advantage  over  node-link  diagrams  in  that  line  crossings  are  impossible

with  matrix  views,  although  path  following  is  harder  in  a  correlation  matrix  than  in  a  node-link

diagram.

Diagrams  such  as  these  can  be  very  important  because  they  can  tell  you  which  other  variables  might

be  confounders  (although  the  multivariate  analyses  should  have  accounted  for  these),  or  involved  in

suppression  or  interaction  effects  –  they’re  not  likely  to  be  more  than  one  degree  of  separation

away  from  the  outcome  variable.

They’re  also  important  to  give  you  a  greater  perspective  of  the  story  of  the  data.  So  often,

researchers  worry  when  a  predictor  variable  that  they  expect  to  be  independently  related  to  the

outcome  is  expelled  from  a  multivariate  analysis.  They  think  it’s  a  crisis.  Being  expelled  from  a

multivariate  analysis  is  not  the  end  of  the  world  and  it  doesn’t  mean  that  the  predictor  variable  is

having  no  effect.  Showing  how  all  the  relationships  interact  with  each  other  in  a  node-link  diagram

puts  all  the  variables  in  their  place,  and  should  put  you  more  at  ease  with  the  results.

●

●

●

You   are   free   to   share   this   book

with   anyone,   as   long   as   you   don’t

change   it   or   charge   for   it

●

●

●

94

**Chapter  7:  Bonus:  Automating  Associations  and  Correlations**

In  this  chapter  you’ll  learn  why  it’s  becoming  increasingly  difficult  to  do  relationship  analyses

manually  and  why  automation  of  the  whole  process  is  important,  and  be  introduced  to  just  such  a

product  –  CorrelViz\*.

\*Disclaimer  –  CorrelViz  is  our  own  product  and  is  available  exclusively  via  Chi-Squared  Innovations.

**7.1:  What  is  the  Problem?**

There  is  no  shortage  of  large,  do-it-all  statistics  programs  that  contain  all  the  statistical  tests  you

need  to  do  your  associations  and  correlations  (and  10,000+  other  statistical  tests  all  rolled  up  into

one  huge  package).

The  problem  with  these  is  that  you  have  to  run  the  analyses  manually.  OK,  sure,  it’s  not  terribly

difficult  to  run  a  2-sample  t-test  or  an  ANOVA,  they’re  just  a  few  clicks  away,  but  let’s  have  a  look  at

this  in  the  context  of  the  whole  dataset.

Let’s  take  a  dataset  of  100  variables  (columns).  That’s  pretty  small  by  today’s  standards.  The  number

of  possible  relationships  in  this  dataset  is:

Now  let’s  suppose  that  it  takes  an  average  of  just  five  mouse  clicks  to  locate  and  select  your  chosen

test,  select  your  variables  and  click  GO.  That  means  that  the  story  of  your  data  is  4,950  x  5  =  24,750

mouse  clicks  away.  \*Gulp\*.

But  what  if  you  then  want  to  find  if  the  same  relationships  exist  only  in  the  male  cohort?  That’s

another  4,851  possible  relationships  and  another  24,255  clicks.  Now  in  just  the  female  cohort?

Another  4,851  (24,255  clicks).  What  about  in  the  smoking  cohort,  the  non-smoking  cohort,  the

diabetic  cohort  and  the  non-diabetic  cohort?  That’s  another  4  x  4,851  x  5  =  97,020  mouse  clicks.  You

already  have  34,056  possible  relationships  (with  170,280  mouse  clicks)  and  you’ve  barely  scratched

the  surface.

And  to  think  you  started  out  with  just  100  variables!

How  long  do  you  think  it’s  going  to  take  you  to  make  nearly  a  quarter  of  a  million  clicks?

95

Even  though  most  of  us  will  never  have  to  deal  with  Big  Data,  we  all  have  Small  Data,  it’s  getting

bigger  and  it’s  coming  at  us  faster.  Only  a  decade  ago  analysts  would  often  be  heard  exclaiming

‘you’ve  got  data?!!?’,  now  they’re  all  terribly  over-worked  because  everybody  has  data  and  they  all

want  it  analysing  yesterday.

So  if  you  have  data  and  the  commercial  stats  programs  are  too  slow,  how  do  you  complete  your

analyses  before  dying  of  old  age?  Until  recently,  the  answer  has  been  that  you  compare  a  few

variables  with  your  hypothesis  variable,  check  out  a  few  inconsistencies,  a  couple  of  subset  analyses

and  then  write  your  paper  or  thesis.  Basically,  the  researcher  cherry-picks  the  analyses  that  she

wants  to  do,  because  she  has  neither  the  time  nor  the  will  to  do  a  complete  analysis.  And  it’ll  still

take  months  to  get  all  the  analyses  done!

In  my  experience,  less  than  10%  of  any  dataset  is  adequately  analysed,  and  less  than  1%  is

published!  That’s  an  enormous  amount  of  information  that  we’re  missing.

Just  imagine  the  discoveries  that  we  could  make  with  the  data  that  already  exists.

Just  imagine  the  discoveries  that  you  might  be  missing  in*your*  dataset.

**7.2:  CorrelViz**

In  the  data-rich  world  that  we’re  living  in  the  old  ways  of  doing  analyses  are  increasingly  letting  us

down  and  you  need  a  faster,  more  efficient  way  of  finding  answers  to  your  questions.

Most  of  the  commercial  statistical  programmes  allow  you  to  program  your  own  macros  and  there

are  whole  programming  languages  dedicated  to  allowing  you  to  create  your  own  statistical  routines.

That’s  great,  but  most  people  that  have  to  do  statistics  can’t  program,  have  no  incentive  to  do  so

and  certainly  don’t  have  the  time.

How  many  surgeons  do  you  know  that  can  program  in  R  or  Python?  How  many  microbiologists  can

program  in  SPSS  or  SAS?  I’ve  never  met  any.

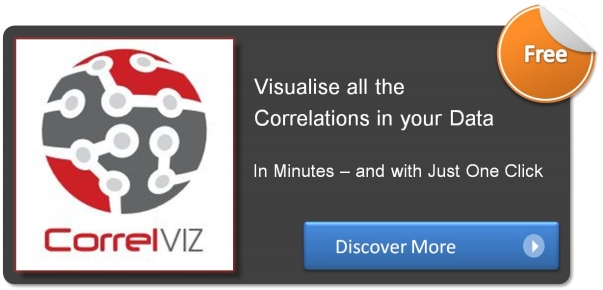
There’s  never  been  a  more  pressing  need  for  a  fully  automated  dedicated  solution  to  the  problem  of

relationship  analysis  than  right  now.

Fortunately,  there  is  a  solution.

96

CorrelViz  is  a  fully  automated  statistical  analysis  and  visualisation  program  that  cleans  and  classifies



your  data,  screens  potential  relationships  by  using  the  appropriate  univariate  statistical  tests,

confirms  or  denies  these  relationships  with  multivariate  stats,  and  presents  you  with  an  intuitive,

interactive  visualisation  of  the  story  of  your  data.

All  this  is  completely  automatic  and  takes  you  from  data  to  story  in  minutes,  not  months.

Best  of  all,  CorrelViz  flips  the  methodology  of  data  analysis  on  its  head.  Instead  of  doing  your  analysis

first  and  then  trying  to  visualise  your  results  last,  in  CorrelViz,  once  your  data  has  been  cleaned  and

classified  all  the  data  analysis  runs  automatically  and  the  first  thing  you  see  is  a  node-link  diagram  of

all  the  independent  relationships  in  the  dataset.  You  get  the  story*first*!

The  visualisation  is  interactive,  so  clicking  on  a  link  will  give  you  the  statistical  details  of  the

independent  relationship  between  the  pair  of  variables.  Clicking  on  a  node  gives  you  all  the  details  of

which  predictor  variables  are  independently  associated  with  your  chosen  outcome  variable.

Through  all  this,  there  is  no  manipulation  of  data,  no  selecting  incorrect  statistical  tests,  no  worries

about  confounding  variables,  and  the  whole  process  takes  just  minutes.  Extract  the  relevant  results

and  your  whole  research  effort  could  be  over  before  lunch!

97

**Summary:  A  Holistic  Strategy  to  Discover  the  Story  of  Your  Data**

Well,  here  we  are  at  the  end.  I  hope  you’ve  found  something  of  value  in  this  book.

Before  we  finish,  let’s  just  recap  on  the  steps  taken  to  discover  the  story  of  your  data:

**Step  1.**  Generate  a  hypothesis  and  start  to*think*  about  the  data  that  you’ll  need  to  collect  in  pursuit

of  the  answer(s)  that  you  seek.  Get  things  wrong  at  this  stage  and  the  rest  of  your  study

could  be  useless.  You  don’t  want  to  go  to  your  statistician  after  3  years  of  study,  only  to  be

told  that  your  dataset  is  not  fit-for-purpose  and  cannot  answer  your  hypothesis.

**Step  2.**  Collect  your  data*sensibly*,  considering  which  relationships  in  your  data  would  be  worthwhile

testing  and  which  would  be  nonsense.  Consider  excluding  the  variables  that  defy  logic.  Make

sure  that  your  dataset  is  in  the  correct  format  and  amenable  to  entry  into  your  chosen

analytics  package.

**Step  3.**  Clean  your  data,  correcting  entry  errors  and  illogical  data  such  as  patients  that  are  -12  or  323

years  old.

**Step  4.**  Classify  your  data,  identifying  those  variables  that  are  ratio,  interval,  ordinal  and  nominal.

The  way  you  identify  your  data  determines  your  choice  of  statistical  tests,  so  it’s  crucial  to

understand  your  data.

**Step  5.**  Understand  the  purpose  of  your  analysis.  If  you  are  seeking  independent  relationships  with

your  hypothesis  variable,  you  will  create  a  different  analytical  strategy  to  that  of  seeking  the

optimum  predictive  model  of  your  hypothesis  variable.

**Step  6.**  Create  your  holistic  analytical  strategy  in  pursuit  of  your  hypothesis.  Your  analysis  will  likely

take  many  months,  so  write  it  down  so  you  understand  what  you’re  doing  and  check  it  with

a  statistician  before  you  start.  Your  strategy  will  likely  involve  using  univariate  statistics  as  a

screening  tool  and  multivariate  statistics  to  confirm  or  deny  the  independent  relationships.

If  you  are  seeking  the  optimum  model,  you  may  need  to  do  a  post-hoc  analysis  to  account

for  suppressor  and/or  interaction  effects.

**Step  7.**  Start  analysing.  First,  use  univariate  statistics  to  weed  out  those  relationships  that  are  not

related  to  the  outcome  variable.  The  survivors  will  pass  to  a  multivariate  test  to  determine

whether  the  relationships  are  independent  (significant)  or  dependent  upon  other  variables

(not  significant).

**Step  8.**  Perform  any  necessary  post-hoc  analyses.

**Step  9.**  Create  a  visualisation  of  the  story  of  your  dataset.

**Step  10.**  Short  cut  steps  3-9  by  using  CorrelViz!

**Step  11.**  Publish  your  paper  in  the  leading  journal  of  your  field  and  be  the  envy  of  your  colleagues!

98

**Epilogue**



Final  point:  try  to  have  fun  with  your  data  and  analyses.  Statistics  is  not  a  closed  shop  and  not  just

for  statisticians.  Ask  ‘what  if…’  of  your  data  and  play  around  with  analyses  in  pursuit  of  knowledge

and  understanding.

If  you  have  categorical  data  that  has,  say,  more  than  a  dozen  or  so  categories,  what  happens  when

you  treat  it  as  continuous  data  and  analyse  it  as  such?  Try  it.  See  what  you  can  learn.  What  happens

when  you  have  an  ordinal  outcome  variable  and  analyse  it  using  a  nominal  logistic  regression?  Give

it  a  go.  You  might  learn  something  new.

Don’t  be  afraid  of  trying  something  out.  If  you’ve  never  done  a  data  transformation  before  but  feel

that  your  analysis  might  benefit  from  it,  then  give  it  a  go.  Play  around  with  a  few  different

transformations  to  get  a*feel*  for  things.

Discuss  your  ideas  and  experiments  with  a  statistician.  Build  your  understanding.  Be  a  better  analyst.

Most  of  all,  don’t  be  afraid  to  make  mistakes  –  you  will  learn  more  from  your  failures  than  from  your

successes.  I  guarantee  you’ll  learn  a  lot  and  you  might  just  enjoy  data  analysis  more  than  just  turning

the  wheel  with  your  nose  to  the  grindstone.

If  you’ve  enjoyed  this  book,  please  feel  free  to  contact  me  and  tell  me  so.  Conversely,  if  you  haven’t

then  I  would*still*  love  to  hear  from  you  (but  be  nice,  please!):



[ebookfeedback@chi2innovations.com](mailto:ebookfeedback@chi2innovations.com)

99

NEXT STEPS



SUBSCRIBE

Why not learn more about data, analysis and statistics by

subscribing to our free newsletter:

**Decimal Points – The CSI Buzz**

You never know, it might not be the worst thing

you do today…

Copyright

The  copyright  in  this  work  belongs  to  the  author,  who  is  solely  responsible  for  the  content.

Please  direct  content  feedback  or  permissions  questions  to  the  [author.](mailto:ebookfeedback@chi2innovations.com)

This  work  is  licensed  under  the  Creative  Commons  Attribution-NonCommercial-NoDerivs  License.

You  are  given  the  unlimited  right  to  print  this  manifesto  and  to  distribute  it  electronically  (via  email,

your  website,  or  any  other  means).

You  can  print  out  pages  and  put  them  in  your  favourite  coffee  shopʼs  windows  or  your  doctorʼs

waiting  room.

You  can  transcribe  the  authorʼs  words  onto  the  sidewalk,  or  you  can  hand  out  copies  to  everyone

you  meet.

You  may  not  alter  this  manifesto  in  any  way,  though,  and  you  may  not  charge  for  it.

