Ch 3. Basic hypothesis tests

Why this lecture is important

- **Link to dissertations**
  - In 2013 one of my students used related techniques to analyse mergers and acquisitions
  - In 2013 one of my students used a paired t-test to analyse accounting ratios
  - In 2014 techniques used by my students who got distinctions ranged from time series econometrics to unpaired t-tests
- **Link to regression**
  - Regression is the prototype model for all applied statistics and econometrics
  - Underpins much research in finance, economics, social sciences
  - Content of today's lecture will help you understand and apply advanced regression techniques

The golden rule – don't panic!

- **Math is usually simpler than it seems at first**
- **Often the questions you ask are quite simple**
- **Math is usually designed to make sense and to convey useful information**

Simple but useful questions

Generic data

- What is a "typical" observation
  - What is the mean?
- How spread out is the data?
  - What is the variance?

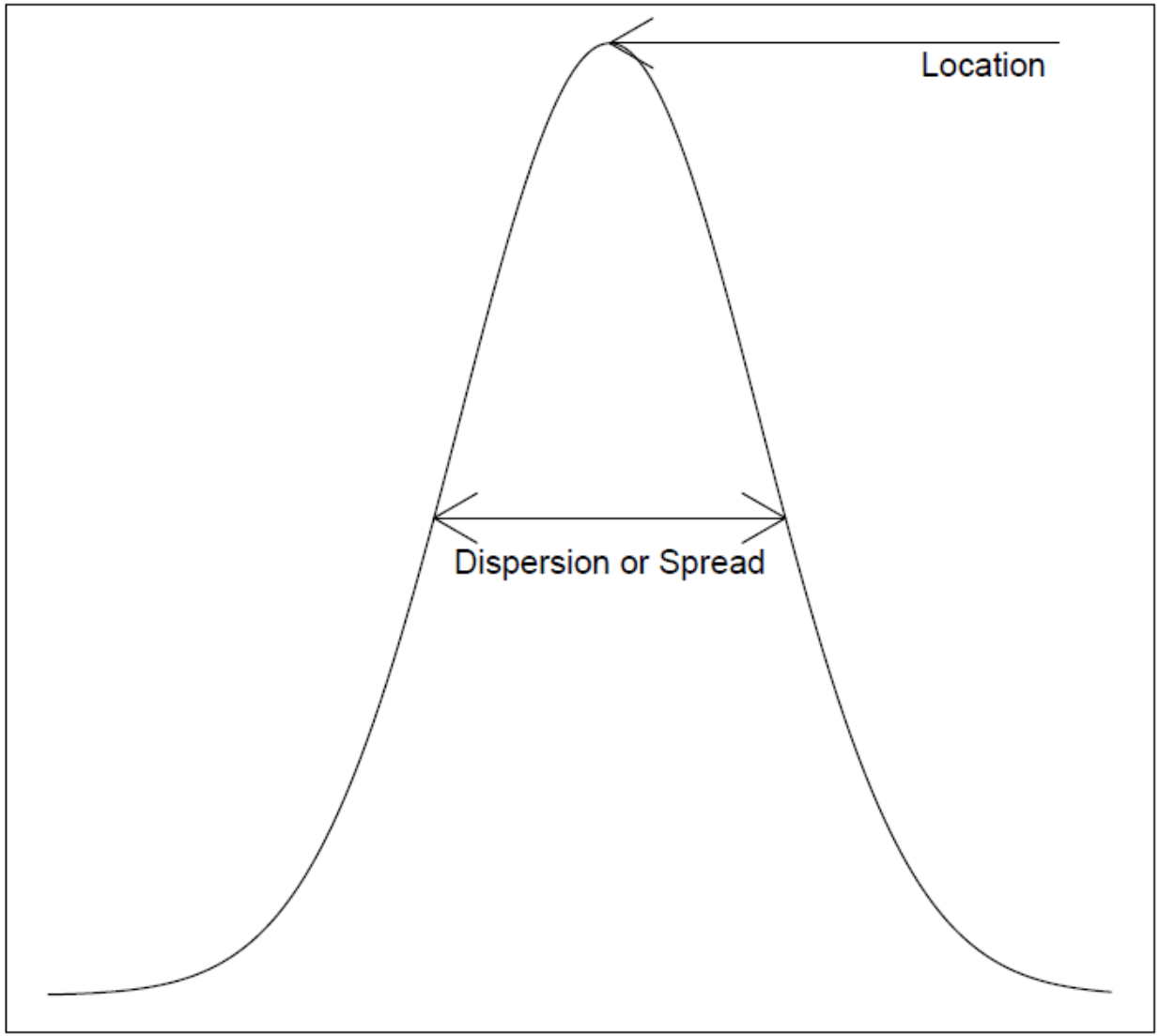
Regression

- What happens to Y as X increases?
  - increases?
  - decreases?
  - nothing?
- **Statistics answers these questions systematically**

Analysing a general dataset

- **Two basic questions**
  1. Location or mean
  2. Spread or variance
- **Answering these questions systematically is vital for research. Statistics enables us to do that.**
  1. One sample and two-sample t-test
  2. Chi-squared test and F-test

Recall the following mental picture...



Outline of the lecture

1. One-sample t-test
2. One-sample variance ratio or χ^2 test – non-examinable
3. Two-sample t-test
4. Two-sample variance ratio or F-test
5. χ^2 test for data on observed counts – non-examinable

1.1 One-sample t-test – illustration by example

- Suppose we have the following data (next slide) on consumer confidence in 2007 and 2009
- We want to see if the crash of 2008 has a lasting impact upon consumer confidence
- **Use a formal statistical test – in this case a t-test to see if there is a detectable difference beyond what would occur by random chance alone**

1.2 Data for illustrative example

Consumer Confidence Index	J	F	M	A	M	J	J	A	S	O	N	D
2007	86	86	88	90	99	97	97	96	99	97	90	90
2009	24	22	21	21	19	18	17	18	21	23	22	21
Difference	62	64	67	69	80	79	80	78	78	74	68	69

1.3 Calculating a t-test

- If calculating a t-test you need the estimated mean and the standard deviation/ estimated standard error (e.s.e)
- In our example the mean difference is

$\bar{x} = \frac{\sum x_i}{n} = \frac{62+64+67+\ldots+68+69}{12} = \frac{868}{12} = 72.33$

- The standard deviation of the differences is

$$s = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}}$$

$$\sum x_i^2 = 62^2 + 64^2 + 67^2 + \ldots + 68^2 + 69^2 = 63260$$

$$s = \sqrt{\frac{63260 - 12\left(\frac{868}{12}\right)^2}{11}} = \sqrt{\frac{474.666}{11}} = 6.5689$$

1.4 t-statistics

- t-statistic

$$t = \frac{\text{estimate} - \text{hypothesised value}}{\text{e.s.e}}$$

- e.s.e. $=$ estimated standard error
- Since the hypothesised value is usually zero, usually

$$t = \frac{\text{estimate}}{\text{e.s.e}}$$

- For a one-sample t-test e.s.e $= s/\sqrt{n}$, n=no. of observations
- For regression problems the e.s.e will usually be given to you or will be calculated for you automatically by the computer software.
- **But the basic procedure and interpretation remains the same!**

1.5 A quick note on p-values

• A p-value is a measure of how strange the data is in relation to the null hypothesis

• **LOW** p-values would cause you to reject the null hypothesis – my high school teacher and lots of students make mistakes here!

• For a two-sided test need values in the 0.025 column of the t-tables [INSERT PICTURE HERE]

- On this course make sure you tell me whether or not $p < 0.05$

- More generally in project work $p < 0.1$ gives some weak or inconclusive evidence against the null hypothesis

1.6 A second note on p-values

- **LOW** p-values are significant
- There is a slightly strange way of setting up the problem
- Easiest to assume the null hypothesis that parameter (in this case the mean difference μ) is equal to zero
- If you reject the null hypothesis this tells you where the true value really is
- (Process is a lot easier than it sounds but may require some practice)

1. **Calculation**
2. **Reference statistical tables or numerical values**
3. **Interpretation**

1.7 Back to the one-sample t-test example

1. **Calculation**

$$t = \frac{\text{estimate} - \text{hypothesised value}}{e.s.e}$$

$$t = \frac{\sqrt{n}(\bar{x} - 0)}{s} = \frac{\sqrt{12}(72.33)}{6.5689} = 38.145$$

2. **Reference the t-tables**

• For 1 sample t-test need t distribution with $n - 1$ degrees of freedom

• For regression t-test need t distribution with $n - p$ degrees of freedom – p is the number of parameters in the model including the constant term

$$t_{0.025}(11) = 2.201, |t| = 38.145 > 2.201, \quad \text{therefore} \quad p < 0.05$$

1.8 One sample t-test example (continued...)

3. **Interpretation**

• The mean difference is positive ($\bar{x} = 72.33$) and statistically significant $p < 0.05$

• Consumer confidence is higher in 2007 than in 2009 – i.e. the 2008 crash has reduced consumer confidence

1.9 Analysis in R

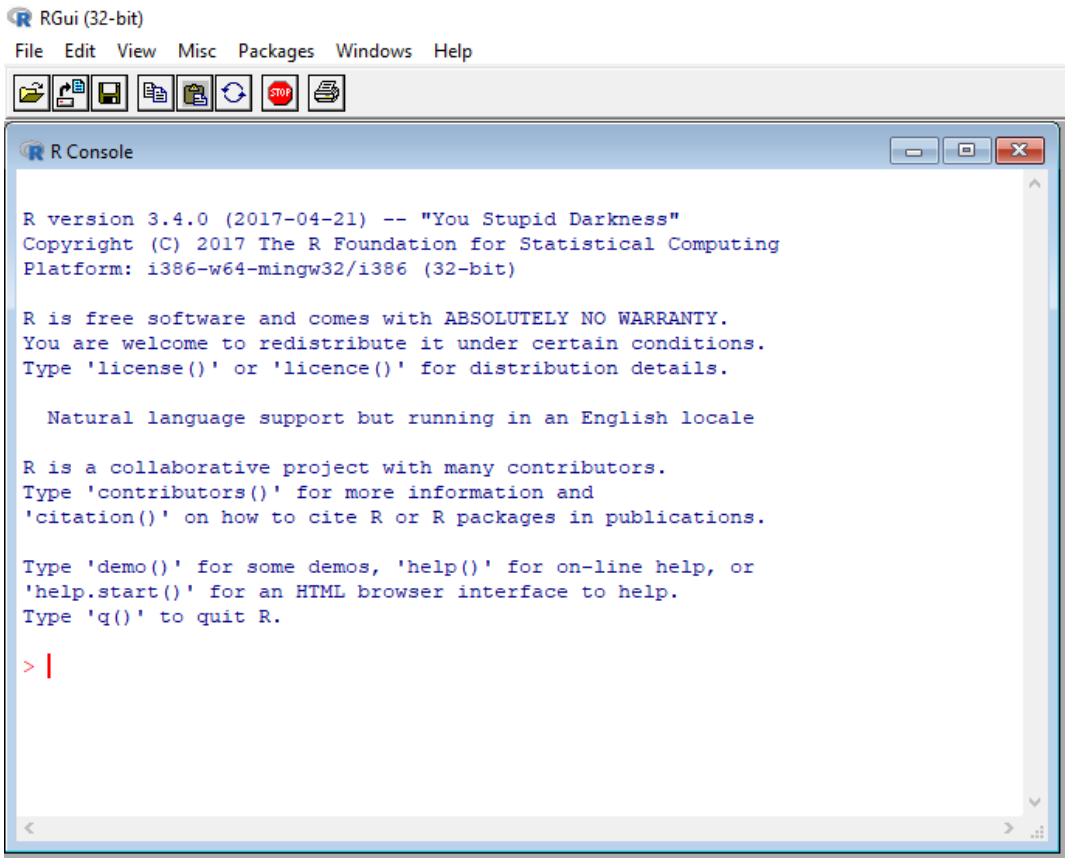
- **Just think of R or any other software package as a computer that's too big to fit in your pocket!**
- Two basic ways of reading data into R
  1. Directly via the command line for trivial lecture examples
  2. Read in using the `read.table` command via the command line from a .txt file in notepad saved to your USB

     - Need to be able to write down the filepath. This is usually shortest and easiest if saved to a USB stick

     - The .txt file that includes the data has to contain equal numbers of evenly spaced columns. This can get a bit fiddly in large problems.

1.10 Downloading R and what R looks like

- In Windows download R on to your computer by googling "R download CRAN" and following the on-screen instructions
- To run R double-click on the R icon on your desktop. This should bring up the command window which on my laptop looks like this



1.11 R solution I (easy)

• Either in the command window in R (pressing enter after each line) or write in notepad (see below) and copy into R from there

```
2007score<-c(86, 86, 88, 90, 99, 97, 97, 96, 99, 97,
90, 90)
2009score<-c(24, 22, 21, 21, 19, 18, 17, 18, 21, 23,
22, 21)
difference<-2007score-2009score
t.test(difference)
```

1.12 R solution II (practical way with decent-sized datasets)

- Access notepad. On my computer `Windows Accessories`$\longrightarrow$`Notepad`
- Copy the file `L2data.txt` from Canvas on to your USB stick. On my computer this is drive E but might be different for your computer
- Read the data in using

```
scores<-read.table(‘‘E:L2data.txt")
score2007<-scores[,1]
score2009<-scores[,2]
```

- The previous analysis can then be repeated using

```
difference<-2007score-2009score
t.test(difference)
```

1.13 Solution in R

- Analysis in R gives

```
data:  difference
t = 38.144, df = 11, p-value = 4.861e-13
alternative hypothesis:  true mean is not equal to 0
95 percent confidence interval:
68.15960 76.50706
sample estimates:
mean of x
72.33333
```

- **This inevitably gives a lot of redundant information, interpretation is deceptively simple.**
- **"Results give significant evidence $p < 0.05$ that the mean difference is not equal to zero. The mean difference is positive and significantly different from zero suggesting that consumer confidence is higher in 2007 than in 2009."**

2.1 One-sample variance-ratio test – non-examinable

- **Two central questions (1) the mean, (2) the variance**
  - 1. What is my best guess of the value?
  - 2. How far away from the true value am I likely to be?
- The one-sample *t*-test answers question 1.
- The one sample variance-ratio test answers question 2.
  - Uses a different chi-squared (χ^2) distribution
  - Not as natural to use as the one-sample *t*-test but may still be useful for statistical work on MSc dissertations.

2.2 One-sample variance ratio test – example

• We have the following data on consumer expectations and consumer spending in 2011

	J	F	M	A	M	J	J	A	S	O	N	D
Cons. Expect.	66	53	62	61	78	72	65	64	61	50	55	51
Cons. Spend.	72	55	69	65	82	77	72	78	77	75	77	77
Diff.	-6	-2	-7	-4	-4	-5	-7	-14	-16	-25	-22	-26

• Want to test the null hypothesis that the standard deviation is equal to 1

• If $\sigma = 1$ then with probability 0.95 the consumer spending index will be within $\pm 2\sigma$ of the consumer expectations index

2.3 One-sample variance-ratio test in broad outline

- Need to calculate the mean
- Need to calculate the variance
- Need to calculate the χ^2 statistic $\frac{(n-1)s^2}{\sigma^2}$ and compare against χ^2_{n-1} – the chi-squared distribution with $n-1$ degrees of freedom where n is the number of observations

2.4 One-sample variance-ratio test – background calculations

- **Calculating the mean**

$$\bar{x} = \frac{-6-2-7-4-\ldots-22-26}{12} = -\frac{138}{12} = -11.5$$

- **Calculating the variance**

$$\begin{aligned} s^2 &= \frac{\sum x_i^2 - n\bar{x}^2}{n-1} \\ \sum x_i^2 &= (-6)^2 + (-2)^2 + (-7)^2 + \ldots + (-22)^2 + (-26)^2 = 2432 \\ s^2 &= \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{2432 - 12(-11.5)^2}{11} = \frac{845}{11} = 76.8182 \end{aligned}$$

2.5 Solving the example

1. **Calculation**
   - Using the results on the previous slide the χ^2 statistic becomes $\frac{(n-1)s^2}{\sigma^2} = \frac{11\left(\frac{845}{11}\right)}{1} = 845$
2. **Reference statistical tables or numerical values**
   - From the tables $\chi^2_{11}(0.025) = 2.201 < 845$
   - Therefore we reject the null hypothesis $\sigma^2 = 1$ $(p < 0.05)$
3. **Interpretation**
   - Reject the null hypothesis $\sigma^2 = 1$ $(p < 0.05)$ – the true standard deviation appears much larger
   - A better estimate for σ would appear to be $s = \sqrt{\frac{845}{11}} = 8.765$. This in turn suggests that with probability 0.95 the consumer spending index will be within $2 \times 8.765 = 17.53\%$ of the consumer expectations index.

2.6 Solution in R

1. Calculate the number in the sample
   `n<-length(difference)`
2. Calculate the chi-squared statistic $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$
   `chisquare<-(n-1)*var(difference)/1`
3. Calculate the p-value
   `1-pchisq(chisquare, n-1)`
   0
4. Interpretation
   Results give significant evidence $p = 0.000 < 0.05$ that the variance is not equal to 1.

3.1 Two-sample t-test

- Want to compare the mean of two independent samples
- t-statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

- n_1=No. in the first sample, n_2=No. in the second sample.
- $\bar{x}_1$=Mean of the first sample, $\bar{x}_2$=Mean of the second sample
- s_1^2=Variance of the first sample, s_2^2=Variance of the second sample

3.2 Two-sample t-test: Example

- Wage data on 10 Advertising Professionals and 13 Accountants
- Want to see if there is any evidence for differences in average pay

Advertising Professionals	36, 40, 46, 54, 57, 58, 59, 60, 62, 63
Accountants	37, 37, 42, 44, 46, 48, 54, 56, 59, 60, 60, 64, 64

3.3 First sample

- **Number** n_1=No. of advertising professionals=10
- **Mean**

$$\bar{x}_1 = \frac{36 + 40 + 46 + \ldots + 62 + 63}{10} = \frac{535}{10} = 53.5$$

- **Variance**

$$\begin{aligned} s_1^2 &= \frac{\sum x_{1,i}^2 - n\bar{x}_1^2}{n_1 - 1} \\ \sum x_{1,i}^2 &= 36^2 + 40^2 + 46^2 + \ldots + 62^2 + 63^2 = 29435 \\ s_1^2 &= \frac{29435 - 10(53.5)^2}{9} = 90.2778 \end{aligned}$$

3.4 Second sample

- **Number** n_2=No. of accountants=13
- **Mean**

$$\bar{x}_2 = \frac{37 + 37 + 42 + \ldots + 64 + 64}{13} = \frac{671}{13} = 51.6154$$

- **Variance**

$$\begin{aligned} s_2^2 &= \frac{\sum x_{2,i}^2 - n\bar{x}_2^2}{n_2 - 1} \\ \sum x_{2,i}^2 &= 37^2 + 37^2 + 42^2 + \ldots + 64^2 + 64^2 = 35783 \\ s_2^2 &= \frac{35783 - 13(51.6154)^2}{12} = 95.7547 \end{aligned}$$

1. Calculating the t-statistic

$$
\begin{aligned}
t &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \\
&= \frac{53.5 - 51.6154}{\sqrt{\frac{90.2778(9)+95.7547(12)}{21}}\sqrt{\frac{1}{10}+\frac{1}{13}}} \\
&= \frac{1.8846}{9.6648 \times 0.4206} = 0.464 \ (3 \ d.p.)
\end{aligned}
$$

2. Referencing the t-tables

• Degrees of freedom $= n_1 + n_2 - 2 = 10 + 13 - 2 = 21$, $t_{21}(0.025) = 2.08$

• So $|t| < t_{21}(0.025) = 2.08$, $p > 0.05$

3. Interpretation

• No evidence ($p > 0.05$) for differences in the average pay amongst advertising professionals and accountants.

• The average pay levels seem to be roughly the same for both professions.

3.7 Two-sample t-test in R

- **Standard statistical tests should be easy in R**
- Two-sample t-test can be conducted in two basic steps
  1. Read in the data
     advertisers<-c(36, 40, 46, 54, 57, 58, 59, 60, 62, 63)
     accountants<-c(37, 37, 42, 44, 46, 48, 54, 56, 59, 60, 60, 64, 64)
  2. The basic command is then `t.test` applied to (first variable, second variable)

```
t.test(advertisers, accountants)
data:  advertisers and accountants
t = 0.46546, df = 19.795, p-value = 0.6467
alternative hypothesis:  true difference in means
is not equal to 0
```

4.1 Two-sample F-test (variance-ratio test)

- **Two basic problems in statistics**
  - Mean/location. Where does a typical observation lie?
  - Variance/dispersion. How "spread out" is the data?
- Two-sample F-test
  - Given two independent samples, is one more spread out than the other?
- Two-sample F statistic

$$F = \frac{s_1^2}{s_2^2} \text{ should be compared against } F_{n_1-1,n_2-1}$$

- s_1^2 **refers to the sample with the largest variance**

4.2 F-test example

- Wage data on 10 Advertising Professionals and 13 Accountants
- Want to see if there are any differences in the spread of the two samples

Advertising Professionals	36, 40, 46, 54, 58, 59, 57, 60, 62, 63
Accountants	37, 37, 42, 44, 46, 48, 54, 56, 59, 60, 60, 64, 64

4.3 Repeat the previous calculations for the advertising professionals...

- **Number** n=No. of advertising professionals=10
- **Mean**

$$\bar{x} = \frac{36 + 40 + 46 + \ldots + 62 + 63}{10} = \frac{535}{10} = 53.5$$

- **Variance**

$$\begin{aligned} s^2 &= \frac{\sum x_i^2 - n\bar{x}^2}{n-1} \\ \sum x_i^2 &= 36^2 + 40^2 + 46^2 + \ldots + 62^2 + 63^2 = 29435 \\ s^2 &= \frac{29435 - 10(53.5)^2}{9} = 90.2778 \end{aligned}$$

4.4 Repeat the previous calculations for the accountants

- **Number** n=No. of accountants=13
- **Mean**

$$\bar{x} = \frac{37 + 37 + 42 + \ldots + 64 + 64}{13} = \frac{671}{13} = 51.6154$$

- **Variance**

$$\begin{aligned} s^2 &= \frac{\sum x_i^2 - n\bar{x}^2}{n-1} \\ \sum x_i^2 &= 37^2 + 37^2 + 42^2 + \ldots + 64^2 + 64^2 = 35783 \\ s_2^2 &= \frac{35783 - 13(51.6154)^2}{12} = 95.7547 \end{aligned}$$

4.5 Calculating the F-statistic

1. Calculating the F-statistic

- We have that $s_1^2 = 95.7547$, $s_2^2 = 90.2778$

$$F = \frac{95.7547}{90.2778} = 1.061 \text{ (3 d.p.)}$$

2. Reference the F-tables

- Degrees of freedom $= n_1 - 1, n_2 - 1 = 13 - 1, 10 - 1 = 12, 9$
- From the F-tables $F_{12,9}(0.05) = 3.07$
- $F = 1.061 < F_{12,9}(0.05) = 3.07$, $p > 0.05$

3. Interpretation

- No evidence ($p > 0.05$) for differences in the variances of each group
- Wage levels for the two groups appear to be equally well spread out

4.6 F-tables: Calculating missing values by interpolation

• **The MOST important thing is that you WRITE DOWN the numbers that you obtain from the F-table**

- Can give you marks for this in the exam – even if you then make mistakes subsequently

- Remember, I am more interested in what you CAN do than what you can't!

- **Also make sure that you WRITE DOWN if $p < 0.05$ or $p > 0.05$**

- Also remember to include a sentence or to in plain English about **interpretation**.

4.7 Interpolation of F-tables: Example 1

- The F-tables used in this course unfortunately contain a lot of gaps
- **Suppose that you need to find a value for** $F_{13,13}(0.05)$
  - Only have instead $F_{12,13}(0.05) = 2.60$, $F_{15,13}(0.05) = 2.53$

Since

$$13 = \frac{2}{3}(12) + \frac{1}{3}(15)$$

Use

$$\begin{aligned} F_{13,13}(0.05) &= \frac{2}{3}F_{12,13}(0.05) + \frac{1}{3}F_{15,13}(0.05) \\ &= \frac{2}{3}(2.60) + \frac{1}{3}(2.53) = 2.53 \text{ (2 d.p.)} \end{aligned}$$

4.8 Interpolation of F-tables: Example 2

- **Suppose that you need to find a value for** $F_{19,8}(0.05)$
  - Only have instead $F_{15,8}(0.05) = 3.22$, $F_{20,8}(0.05) = 3.15$

Since

$$19 = \frac{4}{5}(20) + \frac{1}{5}(15)$$

Use

$$\begin{aligned} F_{19,8}(0.05) &= \frac{4}{5}F_{20,8}(0.05) + \frac{1}{5}F_{15,8}(0.05) \\ &= \frac{4}{5}(3.15) + \frac{1}{5}(3.22) = 3.16 \text{ (2 d.p.)} \end{aligned}$$

4.9 Simple two-sample F-test in R

- As this is a slightly non-standard test this is slightly fiddlier than other examples

1. Read in the data

```
advertisers<-c(36, 40, 46, 54, 57, 58, 59, 60,
62, 63)
accountants<-c(37, 37, 42, 44, 46, 48, 54, 56,
59, 60, 60, 64, 64)
```

2. Calculate the sample values

```
n1<-length(advertisers)
n2<-length(accountants)
F<-var(advertisers)/var(accountants)
```

3. Calculate two-sided p-values (we don't know a priori if the two variances are different which sample is bigger)
   - If $F > 1$ use `2*(1-pf(F, n1-1, n2-1))`
   - If $F < 1$ use `2*pf(F, n1-1, n2-1)`

5.1 χ^2 test for count data

- Chi-squared test is commonly used for count data and is simple and useful
- Examples in Cortinhas and Black (2012), Chapter 16

Cortinhas, C. and Black, K. (2012) *Statistics for business and economics.* Wiley.

- Illustration by example – **will not be on the assessment for this module but may be of some help to dissertation students**

5.2 χ^2 test example

- Data on house residential ownership status by region
- Want to see if residential status depends on region
- Null hypothesis is that there is no association between residential status and region

Region	Owner Occupied	Rented	Total
North West	2180	871	3051
London	1820	1400	3220
South West	1703	614	2317
Total	5703	2885	8588

5.3 How the χ^2 test works

- Need to calculate the expected number for each square in the table E_i
- Compare with the observed number in each square O_i

$$E_i = \frac{\text{Row Total} \times \text{Column Total}}{\text{Total Number of Observations}}$$

Use

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \text{ compare with } \chi^2_{(r-1)(c-1)}$$

- r=No. of rows (going across)=3
- c=No. of columns (going down)=2

5.4 Calculate expected values

Owner Occupied	Rented	Total
$\frac{3051 \times 5703}{8588} = 2026.066$	$\frac{3051 \times 2885}{8588} = 1024.934$	3051
$\frac{3220 \times 5703}{8588} = 2138.293$	$\frac{3220 \times 2885}{8588} = 1081.707$	3220
$\frac{2317 \times 5703}{8588} = 1538.641$	$\frac{2317 \times 1885}{8588} = 778.359$	2317
5703	2885	8588

5.5 Calculate the chi-squared statistic

$$
\begin{aligned}
\chi^2 &= \frac{(O_i - E_i)^2}{E_i} \\
\chi^2 &= \frac{(2180 - 2026.066)^2}{2026.066} + \frac{(871 - 1024.934)^2}{1024.934} \\
&+ \frac{(1820 - 2138.293)^2}{2138.293} + \frac{(1400 - 1081.707)^2}{1081.707} \\
&+ \frac{(1703 - 1538.641)^2}{1538.641} + \frac{(614 - 778.359)^2}{778.359}
\end{aligned}
$$

5.6 Chi-squared Calculations (Continued...)

$$\begin{aligned}
\chi^2 &= 11.695 + 23.119 \\
&= 47.379 + 93.658 \\
&= 17.557 + 34.707 \\
\chi^2 &= 228.12 \text{ (2 d. p.)}
\end{aligned}$$

5.7 Referencing the tables and interpretation

- From tables $\chi^2_2(0.05) = 5.99 < 228.12$
- Therefore there is evidence ($p < 0.05$) that residential status depends on area
- **Interpretation.** Evidence from the table of percentages below suggests that in London fewer homes are owner occupied and more homes are rented

Region	Owner Occupied	Rented
North West	71.5%	28.5%
London	56.5%	43.5%
South West	73.5%	26.5%

5.8 Chi-squared test in R

- It easy to run standard tests like the chi-squared test in R. This is a simple two-step process

1. Need to enter the data as a matrix with the observations in the right order. Here, enter the data first then specify how the table is laid out. Printing the data just by stating the word `count` shows you how the variable `count` has now been constructed:

```
count<-c(2180, 871, 1820, 1400, 1703, 614)
count<-matrix(count, ncol=2, byrow=T)
count
```

2. The basic command to run the test is `chisq.test`

```
chisq.test(count)
```

- **Gives same results as before albeit presented slightly differently**

```
data:  count
X-squared = 228.11, df = 2, p-value < 2.2e-16
```