STATS 10x Revision

CONTENT COVERED: CHAPTERS 7 - 9

Chapter 7: Data on Continuous Variables

ONE SAMPLE

TWO+ INDEPENDENT SAMPLES

PAIRED DATA

PARAMETRIC VS NON-PARAMETRIC

The t-Test

- We can use the **t-Test** when dealing with one or two independent samples (one or two means).
- A t-Test can be used on the following:
  - **Single mean**
  - **Paired data**
  - **Two independent means**
- For more than two samples, we will use the **F-Test for One-way ANOVA** (discussed later).

T-test Procedure: Assumptions

- **INDEPENDENCE (critical)**:
  - for single mean, observations within a sample must be independent.
  - for two means, observations within and between samples must be independent.
- **NORMALITY ASSUMPTION**: the underlying distribution of samples is the normal distribution. The data should be unimodal and have no clusters.
- **15 – 40 GUIDE**: depending on the total group size (n or $n_1 + n_2$), allowances can be made.
  - The greater the size, the more allowances can be made.

SMALL: n or $n_1 + n_2 \leq 15$	MEDIUM: $15 < n$ or $n_1 + n_2 < 40$	LARGE: n or $n_1 + n_2 \geq 40$
No outliers	No outliers	No gross outliers
Slight skewness at most	Not strongly skewed	May be strongly skewed

t-Test for a Single Mean

Clipping from Coursebook Chapter 7, pg 2.

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Concentration	10	5.0410	.16003	.05061

One-Sample Test

	Test Value = 4.92					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Concentration	2.391	9	.040	.12100	.0065	.2355

In a one-sample case, add the test value (eg. 4.92) onto these CI values to get the CI estimate.

Remember to **halve the p-value** if you are doing a one-tailed test.

t-Test for Paired Data

- You would approach the t-Test for paired data similarly to your single mean.
- Usually for paired data you analyse **differences** within each unit's measurements.

Clipping from Coursebook Chapter 7, pg 7.

T-Test

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Cardboard - Metal	1.61111	2.14583	.50578	.54402	2.67821	3.185	17	.005

$\bar{x}_{Difference}$

This is your confidence interval of **difference**.

This is the p-value for a 2-tail test. Halve it if you are doing a 1-tail test.

Remember the output will be in terms of **differences between factor 1 and factor 2.**

MAKE SURE YOU USE **MEDIAN** NOT MEAN FOR NON-PARAMETRIC TEST HYPOTHESES!! >> $\tilde{\mu}$ <<

Non-Parametric Paired Data Testing

- Non-parametric tests don't have an **underlying distribution assumption** (whereas t-Tests have the normality assumption).
- The non-parametric equivalent to a one-sample t-Test is a **sign test**.
- Parametric tests are **superior to non-parametric tests**, but take the same **independence assumptions** (slide 4).

Clipping from Coursebook Chapter 7, pg 10.

Subject	Weight before	Weight after	Difference	Sign
1	65	62	3	+
2	76	75	1	+
3	56	59	−3	−
4	68	66	2	+
5	48	48	0	=
6	59	58	1	+
7	63	61	2	+
8	72	71	1	+
9	60	54	6	+

Assign + - or = to values in respect to the hypothesised value, eg. 0.

SPSS output:

Frequencies

		N
Before - After	Negative Differences [a]	1
	Positive Differences [b]	7
	Ties[c]	1
	Total	9

a. Before < After
b. Before > After
c. Before = After

Make your interpretation from the +/- balance.

Test Statistics[b]

	After - Before
Exact Sig. (2-tailed)	.070[a]

a. Binomial distribution used.
b. Sign Test

This is your p-value. You will need to halve it if you are doing a 1-sided non-parametric test.

t-Test for Two Means from Two Independent Samples

- Both must be random samples and have the same underlying normal distribution. (slide 4)

Group Statistics

	Sexuality	N	Mean	Std. Deviation	Std. Error Mean
Androsterone	Het	11	3.5182	.72086	.21735
	Hom	15	2.5000	.92273	.23825

Clipping from Coursebook Chapter 7, pg 16.

$\mu_{Het} - \mu_{Hom}$

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Androsterone	Equal variances assumed	.564	.460	3.037	24	.006	1.01818	.33523	.32630	1.71007
	Equal variances not assumed			3.157	23.862	.004	1.01818	.32249	.35239	1.68398

IGNORE THIS ROW AND ONLY LOOK AT THE BOTTOM ROW ON SPSS OUTPUTS!

t-test statistic: $t_0 = 1.01818 / 0.32249$

P-value = $\text{pr}(T \geq 3.157) + \text{pr}(T \leq -3.157)$ where $T \sim \text{Student}(df = 23.862)$

$\bar{x}_{Het} - \bar{x}_{Hom} = 3.5182 - 2.5000$

$\text{se}(\bar{x}_{Het} - \bar{x}_{Hom})$

F-test for One-way ANOVA: Assumptions

- **INDEPENDENCE (critical)**: observations between and within the samples are random.
- **NORMALITY ASSUMPTION**: the underlying distribution of all samples is the normal distribution. The data should be unimodal and have no clusters. Plots should **not be strongly skewed**.
- **STANDARD DEVIATIONS**: the standard deviations of the underlying distributions are all equal.
  - As a guide:

You can find the two standard deviations on your SPSS Oneway output. ☺

$$\frac{largest\ standard\ deviation}{smallest\ standard\ deviation} < 2$$

- the F-test is robust against departures from the normal distribution.

F-test for One-way ANOVA

- To calculate the f-Test statistic, use the formula:

$$f_0 = \frac{S^2{}_B}{S^2{}_W}$$

these two values can be found in the 'mean square' column on the ANOVA SPSS output. ☺

Clipping from Coursebook Chapter 7, pg 23.

ANOVA

Increase in Reading Age

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	27.062	3	9.021	4.445	.008
Within Groups	93.351	46	2.029		
Total	120.412	49			

F-test for One-way ANOVA (cont.)

Oneway

Descriptives

Test mark (out of 22)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Yellow	617	13.77	3.978	.160	13.45	14.08	0	22
Pink	624	13.35	3.963	.159	13.04	13.66	3	22
Green	613	13.28	4.123	.167	12.96	13.61	2	22
Blue	623	13.47	4.047	.162	13.15	13.78	2	22
Total	2477	13.47	4.030	.081	13.31	13.63	0	22

These are your smallest and largest standard deviations. Use them in the equation from the previous slide.
eg. 4.123/3.963 = 1.040
1.040 < 2, so your F-test is valid ☺

Clipping from Coursebook Chapter 7, pg 24.

on box plots: horizontal variability

A measure of the variability **between** the four sample means, s_B^2

on box plots: vertical variability

A measure of the variability **within** the four samples, s_w^2

The *F*-test statistic, $f_0 = 28.163/16.226$

ANOVA

Test mark (out of 22)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	84.490	3	28.163	1.736	.158
Within Groups	40126.012	2473	16.226		
Total	40210.502	2476			

don't halve p-values for F-tests

$df_2 = n_{tot} - k$

$df_1 = k - 1$

P-value = pr($F \geq 1.736$) where $F \sim F(df_1 = 3, df_2 = 2473)$

F-test for One-way ANOVA (cont.)

If you see that all the intervals include 0, then it means that there might not actually be any underlying difference in means. There's no point doing the Tukey analysis in this case.

For multiple samples, use Tukey *glubglubglub*

Post Hoc Tests

Multiple Comparisons

Test mark (out of 22)
Tukey HSD

(I) Version	(J) Version	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Yellow	Pink	.414	.229	.269	-.17	1.00
	Green	.484	.230	.151	-.11	1.07
	Blue	.301	.229	.553	-.29	.89
Pink	Yellow	-.414	.229	.269	-1.00	.17
	Green	.070	.229	.990	-.52	.66
	Bliue	-.113	.228	.960	-.70	.47
Green	Yellow	-.484	.230	.151	-1.07	.11
	Pink	-.070	.229	.990	-.66	.52
	Blue	-.183	.229	.855	-.77	.41
Blue	Yellow	-.301	.229	.553	-.89	.29
	Pink	.113	.228	.960	-.47	.70
	Green	.183	.229	.855	-.41	.77

However, if there are CIs where 0 is **not** included, make sure you take note of them and compare them.

eg. page 23: MapScan/Neither

Clipping from Coursebook Chapter 7, pg 24.

Chapter 8: Data on Qualitative Variables

CHI-SQUARE TESTS

One-way vs Two-way Tables of Counts

- **ONE-WAY TABLES OF COUNTS** indicate the test will be for **goodness of fit** between observed and expected values.
  - To write hypotheses for one-way:

  H_0 : The data **comes** from the specified distribution.

  H_1 : The data **did not come** from the specified distribution.

  eg. equal chance every day of the week. (Coursebook, Chapter 8, page 3)

- **TWO-WAY TABLES OF COUNTS** indicate the test will be for **independence** of multiple categories or factors, between observed and expected values.
  - To write hypotheses for two-way:

  H_0 : The two variables **are** independent.

  H_1 : The two variables **are not** independent.

  eg. place of occurrence and type of cancer. (Coursebook, Chapter 8, page 11)

Individual Cell Contributions

- To calculate this, use the formula (part of a larger formula found on the formula sheet):

the value that is actually recorded

the figure you expect to get if the null hypothesis was true (eg. % x total)

$$\frac{(Observed - Expected)^2}{Expected}$$

Degrees of Freedom

- For **one-way tables of counts**, your *df* can be calculated as:

$$No.of\ categories\ -1$$

- For **two-way tables of counts**, your *df* can be calculated as:

$$(Rows\ -1)\times(Columns\ -1)$$

- You can find these formulas on the formula sheet. ☺ Except rows and columns are replaced with variables *i* and *j*.

Chi-Square Test Statistic

- This can also be found on the formula sheet.

$$x_0{}^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

This means ‘the sum of’. Basically, you are adding up all your individual cell contributions that you may have calculated previously. ☺

- The **higher the Chi-square Test Statistic, the more significant the results are**.
- Use the Chi-square Test Statistic to calculate your **P-value**.

Chi-Square Test: P-value

- This will most likely be given to you in the form of an SPSS or Excel output.

It's also the Pearson Chi-Square Sig. value ☺

Chi-Square Tests

	Value	df	Sig.
Pearson Chi-Square	18.922[a]	4	.001
Likelihood Ratio	18.812	4	.001
Linear-by-Linear Association	8.257	1	.004
N of Valid Cases	190		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 9.84.

- You would interpret the P-value as you would with the *t*-Test, *F*-test, etc.
- **NEVER HALVE THE CHI-SQUARE P-VALUE** – you are finding the probability in the right-tail at all times. (Theory is complicated, don't question.)

FAQ: Calculator Skills (graphics only)

- **HELP I DON'T KNOW HOW TO WORK OUT THE P-VALUE USING MY CALCULATOR!**
- You would find the P-value as you normally would for any other probability.
  - From **MAIN MENU**,
  - **> STATS**
  - **> DIST**
  - **> CHI**
  - **> Ccd**: DO NOT choose Cpd. Ccd is the cumulative probability, which you want.
  - **> Lower**: enter in your Chi-Square test statistic here.
  - **> Upper**: enter some random large number such as 9999999999 here.
  - **> *df***: your degree of freedom which you may have calculated previously.
  - **> EXE**
- This is for just in case they are mean and don't give you the SPSS output with the p-value.

Chi-Square Test Validity

- A Chi-Square test won't work unless there is a large number of sample observations.
- We can judge the validity of a Chi-Square test by seeing if the **expected counts** meet the criteria.
  - At least **80% of expected counts** must be **≥ 5;**

AND

  - Each **expected count must be > 1.**

Chapter 9: Regression & Correlation

REGRESSION

EQUATION OF THE LINE

LEAST SQUARES REGRESSION

SAMPLE CORRELATION CO-EFFICIENT (R)

Scatter Plots: Revisited

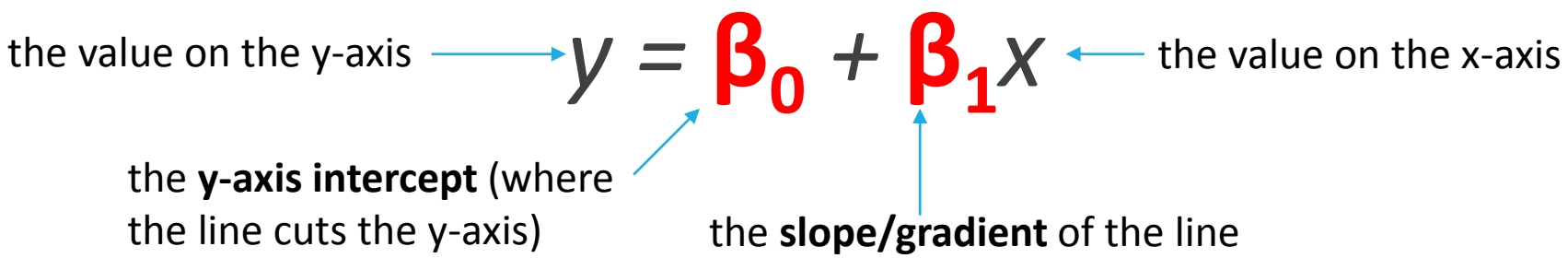
- From my previous Powerpoint slides:
  - **SCATTER PLOT**: you can observe
    - **Trend** – linear vs non-linear
    - **Scatter** – constant vs non-constant
    - **Outliers**
    - **Relationship** – strong vs weak
    - **Association** – positive vs negative
    - **Groupings**
  - Be careful of **subgroups** and **scales of axes**.
- This chapter is all about the scatter plot.

Simple Linear Regression

- The **variables** on a scatter plot need to be carefully identified:
  - The variable along the **x axis** is the independent or explanatory variable. ("the thing affecting")
  - The variable along the **y axis** is the dependent or response variable. ("the thing being affected")
- **INDIVIDUAL POINT DEVIATION**: this can be found by

$$Residual = Observed - Predicted = y - \hat{y}$$

- A **linear regression model** or equation is the equation of a line that is of **best fit** to the plotted data. It takes the form of a normal line equation. You can use it to **predict** values.



Finding Values for the Linear Equation

Clipping from Coursebook Chapter 9, pg 3.

The Regression Equation (the equation of the *fitted* line)

y-intercept

slope/gradient

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.881	3.470		2.271	0.028
	Actual weight (kg)	0.781	.056	0.895	13.866	0.000

a Dependent Variable: Ideal weight (kg)

** Because this slope/gradient is POSITIVE, that means there is a positive association in the scatter plot. The line goes upwards.

LOOK AT THE CORRECT ROW

← About the y-intercept

← About the slope/gradient

ignore this column

The equation for this line would be **y = 07.881 + .781x** ☺

Least Squares Regression Line

- This line is the one with the **smallest sum of squared residuals**.

$$Minimise \sum (residuals)^2$$

- There is only ever ONE least squares regression line for every linear regression!
- The form of the least squares regression line is the same as a standard line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

(it's just laid out the same as slide 23, the symbols correspond ☺)

Things to Be Careful About

- **LINEAR RELATIONSHIP**: do not fit a line if the trend is not clearly linear!
- **OUTLIERS**: outliers can lift the regression line, causing the slope/gradient to be higher than it really is; therefore making predictions less reliable.
- **EXTRAPOLATING**: making predictions beyond the given data set may not be reliable! You don't know what really happens after the data set, observed values may actually drop.
- **SUBGROUPS**: these should be analysed separately as conclusions might not be validly applied to all groups. eg. males vs females.

Sample Correlation Co-efficient (r)

- The **sample correlation co-efficient** is a **value between -1 and 1**. It does **not have units**.
- It measures the strength of the linear association between the x and y variables.
- It measures how closely the points fall on a straight line (the linear regression model).
- r can be obtained by using a calculator or in SPSS outputs.

The Sample Correlation Coefficient, *r*

A correlation co-efficient close to -1 or 1 indicates that the relationship between the two variables are very strong.

Closer to -1 means a negative associated change.
Closer to 1 means a positive associated change.

Correlations

		Actual weight (kg)	Ideal weight (kg)
Actual weight (kg)	Pearson Correlation	1	.895**
	Sig. (2-tailed)		.000
	N	50	50
Ideal weight (kg)	Pearson Correlation	.895**	1
	Sig. (2-tailed)	.000	
	N	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

This is your correlation co-efficient ☺

Testing for No Linear Relationship

- You can test for no linear relationship between x and y variables by testing:
  - $\beta_1 = 0$ as the **null hypothesis** (no relationship; the pattern we saw was due to **chance**)
  - $\beta_1 \neq 0$ as the **alternative hypothesis**.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	7.881	3.470		2.271	.028	.903	14.858
	Actual weight (kg)	.781	.056	.895	13.866	.000	.667	.894

a. Dependent Variable: Ideal weight (kg)

Very strong evidence against the null. There is strong evidence of a positive association between x and y.

Clipping from Coursebook Chapter 9, pg 14.

Making Predictions: Confidence Intervals

From **SPSS** Data Editor

speed	life	LMCI_1	UMCI_1	LICI_1	UICI_1
100	22	22.09571	27.43286	13.90624	35.62233

Lower/Upper **MEAN CI**
This is for estimating a mean y value for a specified x value for a **group or population**.

Lower/Upper **INDIVIDUAL CI**
This is for estimating a y value for a specified x value for an **individual**. (PREDICTION INTERVAL)

Clipping from Coursebook Chapter 9, pg 16.

We use **confidence/prediction intervals** to provide estimates because **point estimates do not account for variability between samples/or between individuals**.

Things to Be Careful About Predicting

- **EXTRAPOLATING**: making predictions beyond the given data set may not be reliable! You don't know what really happens after the data set, observed values may actually drop. You also don't know factors that may occur in the future! :O
- **WEAK RELATIONSHIPS**: if your correlation co-efficient is weak, and there is lots of scatter about your linear regression line, then the predictions may not be very accurate. You might end up with very wide confidence or prediction intervals.