Sociology 213 – Statistics II

November 30, 2018

Multiple regression (continued)

Multiple correlation

Review for final exam

Previous class

- Correlation with Pearson's *r*
  - Preferred measure of association for two interval-ratio variables.
- Coefficient of determination: r^2
- Testing Pearson's *r* for significance
- Partial correlation
- Multiple regression
  - Assessing the effects of the independent variables
  - Using the multiple regression line to predict *Y'*

In this class

- Multiple regression
  - Using the multiple regression line to predict *Y'*
- Standardized partial slopes
- Multiple Correlation R^2
- Review for final exam

Multiple Regression: An Example

- The **zero-order** correlations (correlation coefficients calculated for bivariate relationships are often referred to as "zero-order" correlations as opposed to **partial order** or "first-order") indicate that:
  - husband's contribution to housework is positively related to number of children (r_{y1}=0.50)
  - husbands in higher-SES families tend to do less housework (r_{y2}=-0.30)
  - higher-SES families have fewer children (r_{12} =-0.47)

Multiple Regression: An Example

- The partial slope for the first independent variable (number of children), X_1 is:

$$b_1 = \left(\frac{s_y}{s_1}\right)\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$

$$b_1 = \left(\frac{2.1}{1.5}\right)\left(\frac{0.50 - (-0.30)(-0.47)}{1 - (-0.47)^2}\right)$$

$$b_1 = (1.4)\left(\frac{0.50 - 0.14}{1 - 0.22}\right)$$

$$b_1 = (1.4)\left(\frac{0.36}{0.78}\right)$$

$$b_1 = (1.4)(0.46)$$

$$b_1 = 0.65$$

- A slope of .65 means that the amount of time the husband contributes to housekeeping chores increases by .65 hours per week for each additional child in the family, while controlling for the effects of SES.

Multiple Regression: An Example

- The partial slope for the second independent variable (SES), X_2 is:

$$b_2 = \left(\frac{s_y}{s_2}\right)\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$

$$b_2 = \left(\frac{2.1}{2.6}\right)\left(\frac{-0.30 - (-0.24)}{1 - 0.22}\right)$$

$$b_2 = (0.81)\left(\frac{-0.30 + 0.24}{0.78}\right)$$

$$b_2 = (0.81)\left(\frac{-0.06}{0.78}\right)$$

$$b_2 = (0.81)(-0.08)$$

$$b_2 = -0.07$$

- A slope of -.07 means that the amount of time the husband contributes to housekeeping chores decreases by .07 hours per week for each additional year of education completed by the husband, while controlling for the effects of number of children.

Multiple Regression: An Example

- With the partial slopes, we can now find the *Y* intercept (*a*):

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$
$$a = 3.3 - (0.65)(2.7) - (-0.07)(13.7)$$
$$a = 3.3 - (1.8) - (-1.0)$$
$$a = 3.3 - 1.8 + 1.0$$
$$a = 2.5$$

and the least-squares multiple regression equation:

$$Y = a + b_1X_1 + b_2X_2$$
$$Y = 2.5 + (0.65)X_1 + (-0.07)X_2$$

- As was the case with the bivariate regression line, this formula can be used to **predict** scores on the dependent variable from scores on the independent variable**s**.

Multiple Regression: An Example

- If there are 7 children in the household and the husband has 20 years of education, how many hours per week will he spend on household chores, based on the following least squares multiple regression equation?

$$Y = a + b_1X_1 + b_2X_2$$
$$Y = 2.5 + (0.65)X_1 + (-0.07)X_2$$

Y = 2.5 + (0.65) x 7 + (-0.07) x 20

Y = 2.5 + 4.55 + (-1.4)

Y = 2.5 + 3.15

Y = 5.65

- What about 7 children and 11 years of education?

Y = 2.5 + (0.65) x 7 + (-0.07) x 11

Y = 2.5 + 4.55 + (-0.77)

Y = 6.28

Standardized Partial Slopes (beta-weights)

- Partial slopes (b_1 and b_2) are in the original units of the independent variables.
- To compare the ***relative*** effects of the independent variables when they are based on **different units of measurement**, compute beta-weights (b^*).
- Beta-weights show the amount of change in the *standardized* scores of *Y* for a one-unit change in the *standardized* scores of each independent variable, while controlling for the effects of all other independent variables.

Standardized Partial Slopes (beta-weights)

- Use **Formula 14.7** to calculate the beta-weight for X_1

$$b_1^* = b_1\left(\frac{s_1}{s_y}\right)$$

- Use **Formula 14.8** to calculate the beta-weight for X_2

$$b_2^* = b_2\left(\frac{s_2}{s_y}\right)$$

Beta-weights: An Example

- Beta-weight for the first independent variable, number of children (X_1):

$$b_1^* = b_1\left(\frac{s_1}{s_y}\right)$$

$$b_1^* = (0.65)\left(\frac{1.5}{2.1}\right)$$

$$b_1^* = (0.65)(0.71)$$

$$b_1^* = 0.46$$

- Beta-weight for the second independent variable, SES (X_2):

$$b_2^* = b_2\left(\frac{s_2}{s_y}\right)$$

$$b_2^* = (-0.07)\left(\frac{2.6}{2.1}\right)$$

$$b_2^* = (-0.07)(1.24)$$

$$b_2^* = -0.09$$

Husband's Housework	Number of Children	SES
$\overline{Y} = 3.3$	$\overline{X}_1 = 2.7$	$\overline{X}_2 = 13.7$
$s_y = 2.1$	$s_1 = 1.5$	$s_2 = 2.6$

Beta-weights: An Example

- The beta-weights show that number of **children** (X_1) (**0.46**) relative to **SES** (X_2) **(-0.09**) has a much **larger** effect on husband's contribution to housekeeping chores.

Standardized least-squares regression line

- When using standardized z scores, the least-squares regression equation is defined as:

Formula 14.9 $Z_y = a_z + b^*_1 Z_1 + b^*_2 Z_2$

- Recalling the formula for the *Y* intercept (*a*)

$$a = \overline{Y} - b\overline{X}$$

Standardized least-squares regression line

- Since the mean of any standardized distribution of scores is zero, the mean of the standardized *Y* scores will be zero and the *Y* intercept will also be zero
- The formula can be more simply stated as:

  Formula 14.10 $Z_y = b^*_1 Z_1 + b^*_2 Z_2$
- The standardized regression line with beta weights for the **number of children/SES** example would be:

$$Z_y = (0.46)Z_1 + (-0.09)Z_2$$

Multiple Correlation (R^2)

- The multiple correlation coefficient (R^2) shows the **combined** effects of all independent variables on the dependent variable.

Multiple Correlation (R^2)

- Formula 14.11 allows X_1 to explain as much of the variation in Y as it can, and then adds in the effect of X_2 after X_1 is controlled.
- Formula 14.11 eliminates the overlap in the explained variation between X_1 and X_2.

Formula 14.11

$$R^2 = r_{y1}^2 + r_{y2.1}^2(1 - r_{y1}^2)$$

where R^2 = the multiple correlation coefficient

r_{y1}^2 = the zero-order correlation between Y and X_1, the quantity squared

$r_{y2.1}^2$ = the partial correlation of Y and X_2, while controlling for X_1, the quantity squared

Multiple Correlation (R^2)

Husband's Housework	Number of Children	SES
$\overline{Y} = 3.3$	$\overline{X}_1 = 2.7$	$\overline{X}_2 = 13.7$
$s_y = 2.1$	$s_1 = 1.5$	$s_2 = 2.6$

Zero-order correlations

$r_{y1} = 0.50$

$r_{y2} = -0.30$

$r_{12} = -0.47$

$$r_{y2.1} = \frac{r_{y2} - (r_{y1})(r_{12})}{\sqrt{1 - r_{y1}^2}\sqrt{1 - r_{12}^2}}$$

$$r_{y2.1} = \frac{(-0.30) - (0.50)(-0.47)}{\sqrt{1 - (0.50)^2}\sqrt{1 - (-0.47)^2}}$$

$$r_{y2.1} = \frac{(-0.30) - (-0.24)}{\sqrt{0.75}\sqrt{0.78}}$$

$$r_{y2.1} = \frac{-0.06}{0.77}$$

$$r_{y2.1} = -0.08$$

$$R^2 = r_{y1}^2 + r_{y2.1}^2(1 - r_{y1}^2)$$

$$R^2 = (0.50)^2 + (-0.08)^2(1 - 0.50^2)$$

$$R^2 = 0.25 + (0.006)(1 - 0.25)$$

$$R^2 = 0.25 + 0.005$$

$$R^2 = 0.255$$

Multiple Correlation (R^2)

- $R^2 = 0.255$
- In combination, the two independent variables explain a total of 25.5% of the variation in the dependent variable.
  - Note, since number of children (X_1) explains 25% (r^2_{y1}=.25) of the variance by itself, SES (X_2) adds just .5% to R^2 (.25 +.5= 0.255).
  - There still remains unexplained variation that can be attributed to other variables, random chance, and measurement error.

Final exam review

Final exam topics

1. Hypothesis testing:
   - Definition
   - The notion of null and alternative hypothesis (H_0 and H_1)
   - The logic of rejecting H_0 or failing to reject H_0
     - Critical region
     - Comparison of critical and obtained values
     - Difference between one-tailed and two-tailed distributions
     - The factors determining the rejection of H_0

Final exam topics

2. You will be asked to assess three or four hypothesis testing distributions using the 5-step model (you will be asked to provide steps 2 to 5):
   - One-sample hypothesis testing *t* or *z* distribution (**calculation**)
   - Two-sample hypothesis testing : *t* distribution (**calculation**)
   - Two-variables independence hypothesis testing : Chi-square distribution (**calculation**)
   - Analysis of variance (ANOVA) : *F* distribution (**calculation**)
     - SPSS output (**interpretation**)
     - Decision for rejecting or not rejecting H_0 (**interpretation**)
     - Probability of error = alpha level = significance (**interpretation**)

Final exam topics

3. Association between variables:
   - Information provided by a scattergram (**interpretation**)
   - Interpretation of correlations (**interpretation**)
   - Interpretation of Coefficient of determination (**interpretation**)
   - Predicting dependent variable scores from the least-squares regression equation (simple or **multiple**)

Question format

- Part 1
  - Multiple choice questions (30)
- Part 2
  - Hypothesis testing
    - Calculation problems (4)
- Part 3
  - Analysis of Variance (ANOVA) SPSS output
    - Interpretation questions (3 – 5)

Exam time and office hours

- Final exam December 13 2:00 pm to 5:00 pm
  - Bring a (not connected to the internet) calculator
  - All formulae will be provided
- Wednesday December 5, 2:00 pm to 3:00 pm or by appointment
- Celine will have usual office hours on Tuesday December, 4 (2:45 to 4:00) and she will announce extended office hours as well.

Sections to study

- Hypothesis testing: general considerations
- Hypothesis testing I: The one-sample case
- Hypothesis testing II: The two-sample case
- Hypothesis testing III: The analysis of variance
- Hypothesis testing IV: Chi square
- Bivariate associations:
  - Scattergrams
  - Correlations
  - Coefficient of determination
- Multivariate associations
  - Multiple regression
  - Multiple correlation

Hypothesis testing - differences

Type of comparison	Hypotheses	Sampling distribution	Statistical test
One-sample mean comparison Comparing one group to population	H_0: µgroup = : µpopulation	t distribution	t test $t = \frac{\bar{X}_i - \mu}{s/\sqrt{N-1}}$ df - N -1
Two-sample means comparison	H_0: µgroup 1= : µgroup2	t distribution	t test $t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{x}-\bar{x}}}$ $\sigma_{\bar{x}-\bar{x}} = \sqrt{\left(\frac{s_1^2}{N_1-1} + \frac{s_2^2}{N_2-1}\right)}$ df = N1 + N2 -2
ANALYSIS OF VARIANCE (ANOVA) **Three or more samples means** comparison $H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$		F distribution	F test $F = MSB/MSW$ $dfw = (N - k) = 33$ $dfb = k - 1 = 2$

Hypothesis testing - Associations

Type of comparison	Hypotheses	Sampling distribution	Statistical test
Nominal-level variables	H_0: variables are independent $H_0: f_o = f_e$	$\chi 2$ distribution	$\chi 2$ test $\chi^2\ (obtained) = \Sigma[(f_o - f_e)^2/f_e]$ $df = (r-1)(c-1) = 1$

Summary table

Type of variables	Measures of association	Calculating statistics	Interpretation
Interval-ratio variables	**Relationship between I/R variables**	**The value of the coefficients ranges between -1 and +1**	
	1. Scattergram : visual inspection	Place cases in a two-dimensional space - Vertical axis (Y-axis) : dependent variable - Horizontal axis (X-axis) : independent variable	Pattern of the dots should indicate : - whether an association exists between the variables - the direction of the association (+ or -) - Strength of the association
	2. Correlation	Pearson r $r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$	The Pearson r indicates: -If variables are associated positively (higher values on one variable are associated with higher values on the other variable) or negatively (higher values on one variable are associated with lower values on the other variable) - The strength of the association Use table 14.2 in text to assess strength of association
	3. Coefficient of determination	$r^2 = \frac{\Sigma(Y'-\bar{Y})^2}{\Sigma(Y-\bar{Y})^2}$	The proportional reduction of error in predicting the outcome when the independent variable is considered. The proportion of the variation of the dependent variable (Y) that is explained by the independent variable (X)

Summary table

Type of variables	Measures of association	Calculating statistics	Interpretation
Interval-ratio variables	**Relationship between I/R variables**		
	4. Regression line (Ordinary Least Squares Line – OLS)	$Y = a + bX$	
		a = intercept	The value on the Y axis where the regression line meets the Y axis The value of Y when X = 0
		b = slope	the amount of change in Y for every change of one unit in X Gives the inclination of the line and its direction (positive or negative)
	5. Multiple regression line	$Y = a + b_1X_1 + b_2X_2$	
			SHORT SUMMARY We use the observed data to find the regression line The regression line equation is a link function between X and Y Using this link function, we can calculate the predicted value of Y (Y') for any value of X

Hypothesis testing

Hypothesis testing

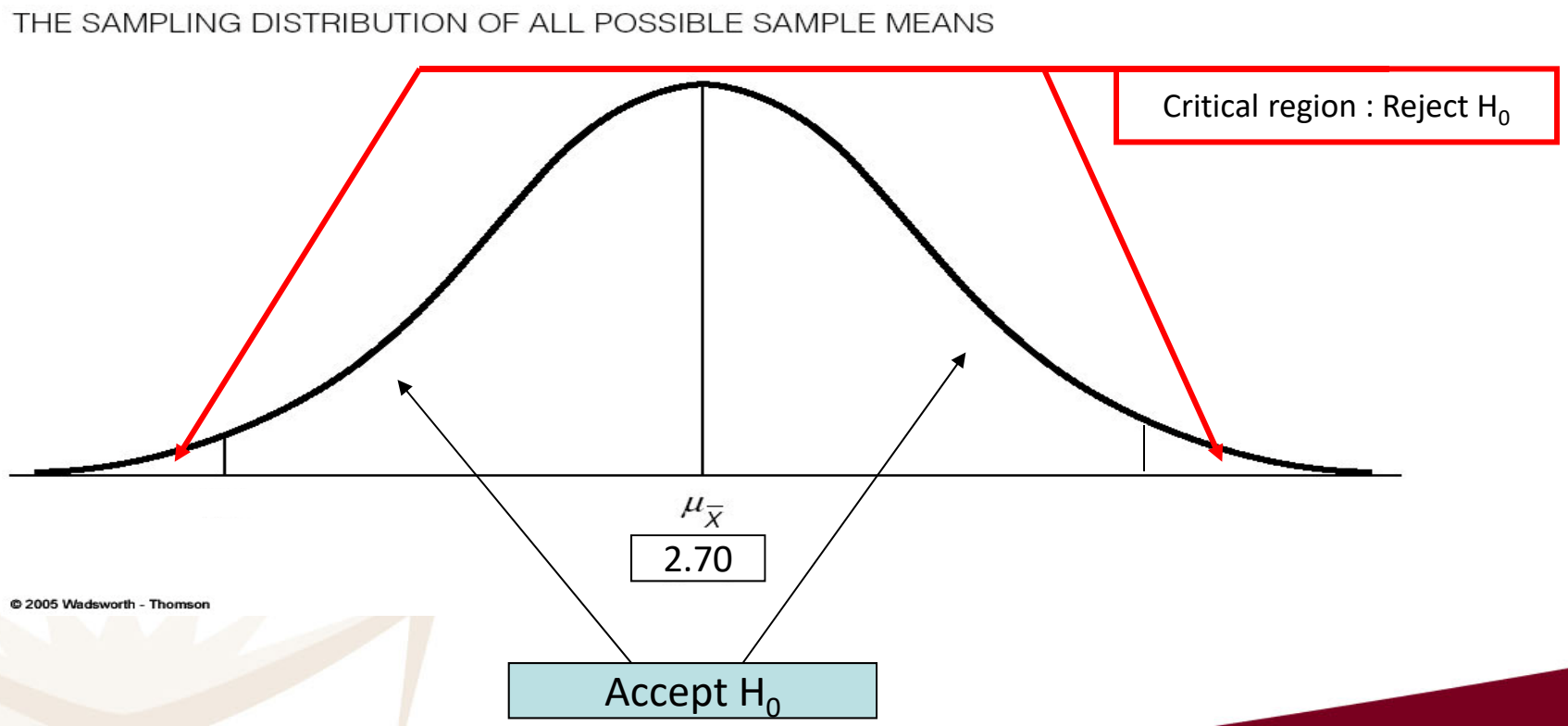
- In all situations:
  - We have a randomly selected SAMPLE to represent the general population
  - We aim to generalize (or make inferences) from the sample to the population

Hypotheses

1. Null Hypothesis (H_0)
   - The H_0 always states there is “no significant difference.”
   - The difference is observed by chance
2. Alternative hypothesis (H_1)
   - “The difference is real”.
   - (H_1) always contradicts the H_0.

The logic of rejecting H_0

1. The critical region



The logic of rejecting H_0

- Comparing the obtained value on the test to the critical value:
  - Value (obtained) < value (critical)
    - ✓ Fail to reject H_0
  - Value (obtained) > value (critical)
    - ✓ Reject H_0

The logic of rejecting H_0

- Comparing the obtained value on the test to the critical value:
  - Value (obtained) : calculate in step 4
  - Value (critical) : find in the appropriate table

The logic of rejecting H_0

- Alternatively, examining the alpha level, p-value, level of **sig**nificance from SPSS output:
  - P-value should be <= 0.05

One-Sample Test

	Test Value = 18.15					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
In general, would you say your health is	-1659.695	13045	.000	-15.711	-15.73	-15.69

The logic of rejecting H_0

Comparing obtained to critical values

- Assuming that the critical value for $\chi^2 = 3.84$, (alpha .05) which of the following results should make us reject the null hypothesis:

  a. Obtained $\chi^2 = 1.95$

  b. Obtained $\chi^2 = 3.89$

  c. Obtained $\chi^2 = 13.05$

  d. Obtained $\chi^2 = 32.13$

Examining the alpha level (p-value)

- Link the below listed p-values to the corresponding χ^2 value:

  a. P-value = 0.001

  b. P-value = 0.75

  c. P-value = 0.000

  d. P-value = 0.048

The logic of rejecting H_0

Comparing obtained to critical values

- Assuming that the critical value for $\chi^2 = 3.84$, (alpha .05) which of the following results should make us reject the null hypothesis:

  a. Obtained $\chi^2 = 1.95$

  b. Obtained $\chi^2 = 3.89$

  c. Obtained $\chi^2 = 13.05$

  d. Obtained $\chi^2 = 32.13$

Examining the alpha level (p-value)

- Link the below listed p-values to the corresponding χ^2 value:

  a. P-value = 0.001

  b. P-value = 0.75

  c. P-value = 0.000

  d. P-value = 0.048

The logic of rejecting H_0

- What are the determining factors?:
  - The size of the difference (test)
  - The sample size (N)
  - Alpha level
  - One or two tailed test

Testing Hypotheses
The Five Step Model

1. Make assumptions and meet test requirements.
2. State the null hypothesis H_0.
3. Select the sampling distribution and determine the critical region.
4. Calculate the test statistic.
5. Make a decision and interpret results.

The five-step model: one-sample versus two-sample testing

- **Step 2** – Stating the null hypothesis
  - Statement of "no difference"
    - H_0: $\mu_1 = \mu_2$
      - The Null asserts there is no significant difference between the populations.
    - H_1: $\mu_1 \neq \mu_2$
      - The research hypothesis contradicts the H_0 and asserts there is a significant difference between the populations.
  - There are no differences between group A and group B on the variable being measured

$$H_0: \mu_{males} = \mu_{females}$$

The five-step model: one-sample versus two-sample testing

- **Step 3** – Select sampling distribution and establish critical region
  - Sampling Distribution = *t* distribution
  - Alpha (α) = 0.05
  - *t(*critical*) or* Z (critical) = $\pm$ 1.96 for large samples
  - The sample outcome is the **DIFFERENCE BETWEEN THE SAMPLE STATISTICS**

The five-step model: one-sample versus two-sample testing

- **Step 4** – Calculating the test statistic
  - Calculating *Z*
  - *Z*(OBTAINED) =

$$\frac{(\bar{X}_{MALES} - \bar{X}_{FEMALES}) - (\mu_{MALES} - \mu_{FEMALES})}{\sigma_{\bar{x}-\bar{x}}}$$

The five-step model: one-sample versus two-sample testing

- **Step 4** – Calculating the test statistic
  - Calculating *Z*
  - *Z*(obtained) =

Formula 11.2

$$\frac{(\bar{X}_{MALES} - \bar{X}_{FEMALES})}{\sigma_{\bar{x}-\bar{x}}}$$

?

  - pooled estimate of the standard error, formula 11.3 or second formula on the next slide

The five-step model: one-sample versus two-sample testing

- **Step 4** – Calculating the test statistic
  - Use formula 11.3 when the population standard deviations are known

Formula 11.3

$$\sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

  - Use formula below when the population standard deviations are unknown

$$\sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$$

The five-step model: one-sample versus two-sample testing

- **Step 5** – Make a decision and interpret results
  - If the test statistic is in the Critical Region(α=.05, beyond critical values):
    - *Reject* the H_0. The difference is significant.
  - If the test statistic is not in the Critical Region (at α=.05, between critical values):
    - *Fail to reject* the H_0. The difference is not significant.

The 5 step model
Multiple samples test (ANOVA)

Basic Logic

- Can think of ANOVA as extension of *t* test for more than two groups.
- ANOVA asks “are the differences between the samples large enough to reject the null hypothesis and justify the conclusion that the populations represented by the samples are different?”

Basic Logic

- If the H_0 is true, the sample means should be about the same value.
  - If the H_0 is true, there will be little difference between sample means.
- If the H_0 is false, there should be substantial differences *between* categories, combined with relatively little difference *within* categories.
  - The sample standard deviations should be low in value.
  - If the H_0 is false, there will be big difference between sample means combined with small values for *s*.

Basic Logic

- The larger the differences *between* the sample means, the more likely the H_0 is false, particularly when there is little difference *within* categories.
- When we reject the H_0, we are saying there are differences between the *populations* represented by the sample.

Step 3: Select Sampling Distribution and Establish the Critical Region

- Sampling Distribution = *F* distribution
- Alpha = 0.05
- dfw = $(N - k) = 33$
- dfb = $k - 1 = 2$
- *F*(critical) = 3.32

  (Note, the exact dfw (33) is not in the table but dfw = 30 and dfw = 40 are. Choose the *larger F* ratio as *F* critical).

Step 4: Calculating F

- Use this formula to find SST.

Shortcut **Formula 12.10**

$$SST = \sum X^2 - N\bar{X}^2$$

$$SST = \sum X^2 - N\bar{X}^2$$

= (20 213 + 27 607 + 45 253) – (36)(47.86)^2

$$SST = 93073 - (36)(47.86)^2$$

$$SST = 93073 - (82460.87)$$

$$SST = 10612.13$$

Example of Computation of ANOVA

Voter	Municipal		Provincial		National	
	X	X^2	X	X^2	X	X^2
1	33	1089	35	1225	42	1764
2	78	6084	56	3136	40	1600
3	32	1024	35	1225	52	2704
4	28	784	40	1600	66	4356
5	10	100	45	2025	78	6084
6	12	144	42	1764	62	3844
7	61	3721	65	4225	57	3249
8	28	784	62	3844	75	5625
9	29	841	25	625	72	5184
10	45	2025	47	2209	51	2601
11	44	1936	52	2704	69	4761
12	41	1681	55	3025	59	3481
$\Sigma X =$	441		559		723	
$\Sigma X^2 =$		20213		27607		45253
$\overline{X_k}$	(441/12) = 36.75		(559/12) = 46.58		(723/12) = 60.25	
$\overline{X} = (441+559+723)/36 = 47.86$						

Step 4: Calculating F

Use Formula 12.4 to find SSB

$$\Sigma N_k(\bar{X}_k - \bar{X})^2$$

$= 12(36.75\text{-}47.86)^2 + 12(46.58\text{-}47.86)^2 + 12(60.25\text{-}47.86)^2$

$= 12(123.43) + 12(1.64) + 12(153.51)$

$= 1481.16 + 19.68 + 1842.12$

$= 3342.96$

Step 4: Calculating F

- Find SSW by subtraction (Formula 12.11)
  - SSW = SST – SSB
  - SSW = 10612.13 - 3342.96
  - SSW= 7269.17

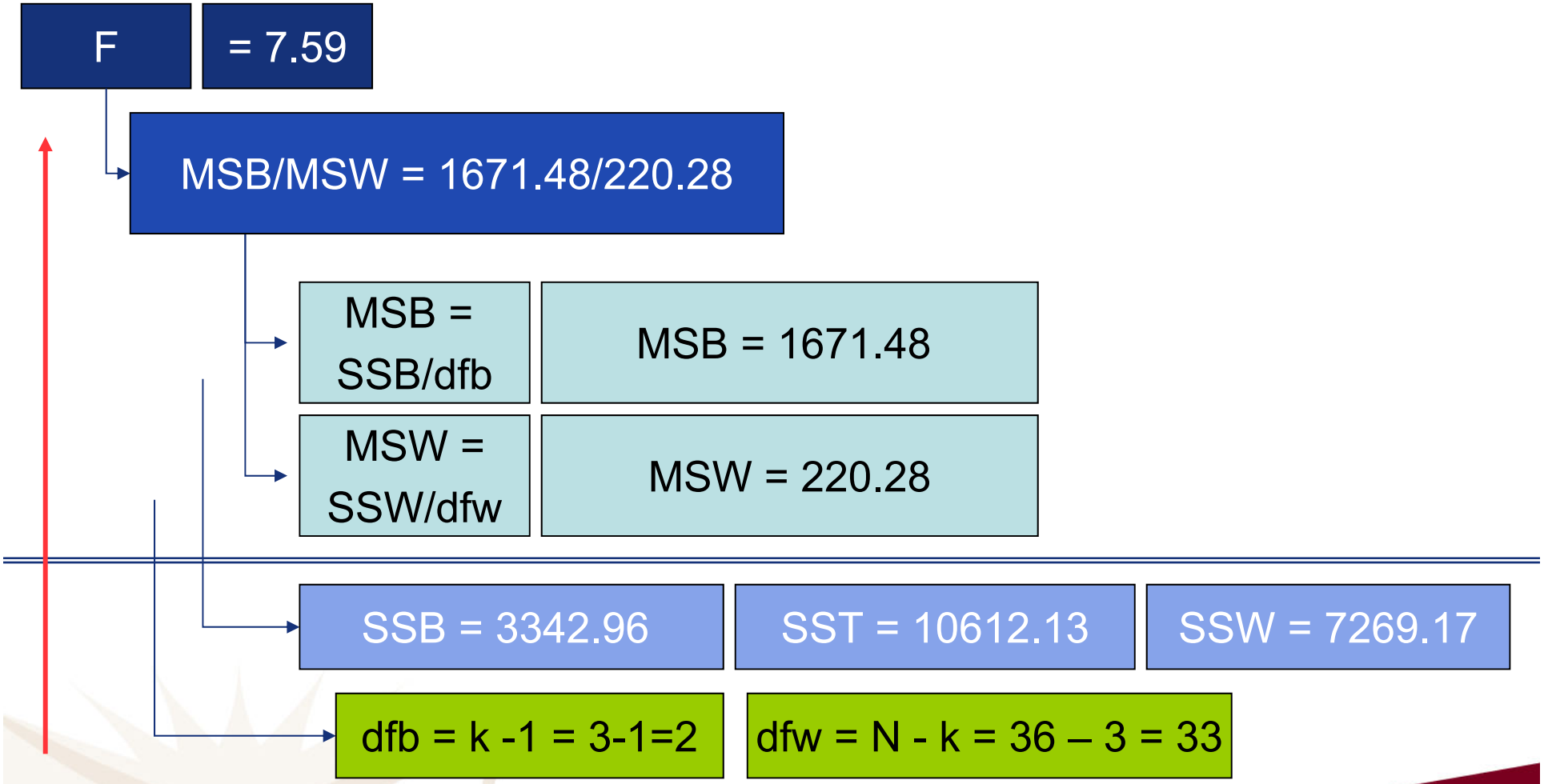
Step 4: Calculating F

- Use Formulas 12.7 and 12.8 to find the Mean Square Estimates (estimates of the population variance:

  - MSW = SSW/dfw
  - MSW =7269.17/33
  - MSW = 220.28

  - MSB = SSB/dfb
  - MSB = 3342.96/2
  - MSB = 1671.48

Step 4: Calculating F



Interpreting **SPSS** output

Descriptives

Self-perceived physical health

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
MARRIED	26969	2.41	1.009	.006	2.39	2.42	1	5
COMMON-LAW	5308	2.35	.950	.013	2.32	2.37	1	5
WIDOW/SEP/DIV	13225	2.68	1.098	.010	2.66	2.70	1	5
SINGLE/NEVER MAR	17787	2.33	.973	.007	2.32	2.35	1	5
Total	63289	2.44	1.022	.004	2.43	2.45	1	5

ANOVA

Self-perceived physical health

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1055.924	3	351.975	342.768	.000
Within Groups	64984.828	63285	1.027		
Total	66040.752	63288			

The 5-step model

Relationship between nominal-level variables

Test of independence

Bivariate Tables

- **Columns** are scores of the independent variable.
  - There will be as many columns as there are scores on the independent variable.
- **Rows** are scores of the dependent variable.
  - There will be as many rows as there are scores on the dependent variable.

The 5-step model

- **Step 1** Make Assumptions and Meet Test Requirements
- **Step 2** State the Null Hypothesis
- **Step 3** Select the sampling distribution and Establish the critical region.
- **Step 4** Calculate the test statistic
- **Step 5** Make a Decision and Interpret the Results of the Test

Step 2: State the Null Hypothesis

- H_0: The variables are independent
  - Another way to state the H_0, more consistently with previous tests:
    - H_0: $f_o = f_e$
- H_1: The variables are dependent
  - Another way to state the H_1:
    - H_1: $f_o \neq f_e$

Step 3: Select the Sampling Distribution and establish the Critical Region

- Sampling Distribution = χ^2
- Alpha = .05
- df = (r-1)(c-1) = 1
- χ^2 (critical) = 3.841

Step 4: Calculate the Test Statistic

Formula 7.1

$$\chi^2(\text{obtained}) = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o = the cell frequencies observed in the bivariate table
f_e = the cell frequencies that would be expected if the variables were independent

Formula 7.2

$$f_e = \frac{\text{Row marginal} \times \text{Column marginal}}{N}$$

Step 5: Make a Decision and Interpret the Results of the Test

- Compare χ^2 (critical) with χ^2 (obtained)
  - If the test statistic is in the critical region. we reject the H_0.
  - Rejecting the H_0 means the variables are dependent and related
  - Failing to reject the H_0 means the variables are independent

Measures of Association

Association and Bivariate Tables

- Association between two variables (bivariate associations):
  - Are best illustrated with bivariate tables
  - Can be investigated by finding answers to three questions:
    - Does an association **exist**?
    - How **strong** is the association?
    - What is the **pattern or direction** of the association?

Scattergrams

- **Inspection of the scattergram** should always be the first step in assessing the correlation between two interval-ratio variables
- The scattergram provides a first impression about :
  - The existence of a relationship
  - The linearity of a relationship (minimal condition for linear regression)
  - The strength of a relationship
  - The direction of a relationship

Regression analysis

- Assesses:
  - The existence of a relationship (via the slope)
  - The strength of a relationship (moderately via the slope)
  - The direction of the relationship
- Predicts an **outcome** (dependent variable) from a independent variable using:
  - Least-squares regression line

$$Y = a + bX$$

Correlation

- Pearson's *r*, known as the correlation coefficient is the preferred measure of association for interval-ratio variables.

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

- Use the guidelines stated in Table 9.3 as a guide to interpret the strength of Pearson's *r*.

Table 9.3 **The Relationship Between the Value of Ordinal-Level Measures of Association and the Strength of the Relationship**

Value	Strength
If the value is	*The strength of the relationship is*
between 0.00 and 0.10	weak
between 0.11 and 0.30	moderate
greater than 0.30	strong

Coefficient of determination

- Coefficient of determination represents the proportion of variation in a dependent variable that is explained by the independent variable.
- It can also be interpreted in terms of the **proportional reduction of error** in predicting a dependent variable (Y)
- Provides a more direct and less arbitrary interpretation than *r*.
- The coefficient of determination – *explained variation* divided by *total variation* (square of the correlation coefficient (r^2))

Multiple Regression

- Least-squares multiple regression equation:

$$Y = a + b_1X_1 + b_2X_2$$

where,

a = the Y intercept (Formula 14.6)

b_1 = the partial slope of the first independent variable (X_1) on Y (Formula 14.4)

b_2 = the partial slope of the second independent variable (X_2) on Y (Formula 14.5)

Multiple Regression

- This least-squares regression equation involves at least two independent variables, but could be modified to include any number of independent variables.
- As is the case with the bivariate regression line, this formula can be used to predict scores on the dependent variable from scores on the independent variables.

Sample Questions

Question

1. The critical region is:
   a) The area between the mean and the obtained value
   b) The area between the mean and the critical value
   c) The area beyond the critical value
   d) The area given in appendix B

The area beyond the critical value

Question

2. The statement in which you say there is a difference between the sample and the population is called the
   a) Null hypothesis
   b) Research hypothesis
   c) Assumption statement
   d) Critical region

Research hypothesis

Question

3. If you state that the sample mean is hypothesized to be less than the population mean, then the critical region should be:
   a) On both sides of the mean
   b) On the right side of the mean
   c) On the left side of the mean
   d) In the centre of the distribution

On the left side of the mean

Question

4. In order to test for significance between the two populations, we assume:
   a) Simple random samples
   b) Independent random samples
   c) Sampling distribution to be random
   d) Sampling distribution to be positive

Independent random samples

Question

5. The null hypothesis states:
   a) No difference between samples
   b) No difference between populations
   c) The two samples are different
   d) The two populations are different

No difference between populations

Question

6. The research hypothesis states:
   a) No difference between samples
   b) No difference between populations
   c) The two samples are different
   d) The two populations are different

The two populations are different

Question

7. ANOVA uses the _____ distribution
   a) A
   b) F
   c) t
   d) Z

F

Question

8. The test statistic for ANOVA is equal to:
   a) dfb/dfw
   b) dfw/dfb
   c) Mean square within/mean square between
   d) Mean square between/mean square within

Mean square between/mean square within

Question

9. Observed frequencies are:
   a) Cell frequencies expected if the two variables are independent
   b) Total frequencies observed
   c) Total frequencies expected
   d) Cell frequencies observed from the raw data

Cell frequencies observed from the raw data

Question

10. Expected frequencies are:
    a) Cell frequencies expected if the two variables are independent
    b) Total frequencies observed
    c) Total frequencies expected
    d) Cell frequencies observed from the raw data

Cell frequencies expected if the two variables are independent

Question

11. The chi-square test is affected by:
    a) Row marginals
    b) Sample size
    c) Cell frequencies
    d) All of the above

All of the above

Question

12. In order to reject a null hypothesis, the obtained value must fall_____ the critical value.
    a) Within
    b) To the left
    c) Beyond
    d) Nowhere near

Beyond

Statistics means never having to say you're certain!