The Data Driven Manager

2

Correlation and Association

Learning Objectives

- Differentiate between correlation and association
- Calculate the Pearson product moment correlation coefficient
- Create and interpret scatter plots
- Interpret the strength of the relationship between two continuous variables

Learning Objectives

- Explain the coefficient of determination
- Test for significance of the Pearson product-moment correlation coefficient when the null hypothesis is zero and when the null hypothesis is a number greater than zero

Learning Objectives

- Calculate, interpret and test the Spearman rank-order correlation coefficient for significance
- Create a contingency table
- Compute the Pearson Phi, Cramer's V and Chi-Square statistics

One Sample Tests for Correlation & Association

Measures of Relationship

- Correlation and association are measures of the strength of a relationship between two variables.

Measures of Relationship

Before we calculate statistics related to relationship, we must first properly classify each variable.

- Nominal
- Ordinal
- Continuous

Correlation and Association

- Where both variables are **continuous**, the statistic employed to measure the relationship may be referred to as a coefficient of **Correlation**

- Where both variables are **nominal**, the statistic employed to measure the relationship may be referred to as a coefficient of **Association**

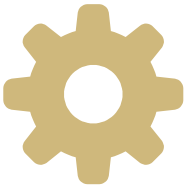
Correlation and Association

Coefficients of Correlation and Association can vary given all possible combinations of nominal, ordinal, and continuous data that can occur.

	One-Sample Association or Correlation		
	Interval or Ratio	Ordinal	Nominal
Interval or Ratio	t for Pearson's r, Fisher's Z for r	↓	t for r_{pbi} (point biserial)
Ordinal	→	t for Spearman's r_s	↓
Nominal	t for r_{pbi} (point biserial)	→	χ^2 for C, V, ϕ

Correlation

What is Correlation?



- **Correlation** is the relationship between two variables which may be described with a correlation coefficient

Coefficient of Correlation

- The most frequently used coefficient of correlation is the Pearson Product-Moment Coefficient of Correlation

- Symbols
  - Population: ρ_{xy}
  - Sample: r_{xy}

Product-Moment Coefficient

- Correlation coefficient has two components:
  - Sign (+ or -)
  - Numeric value

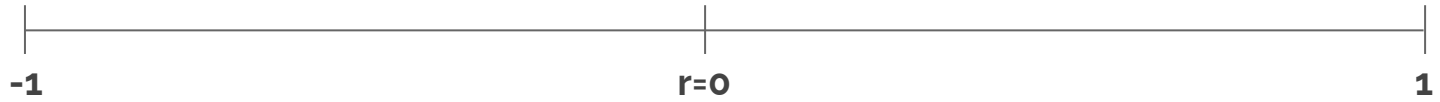
$$r_{XY} = \frac{\sum(X - \bar{X})\,(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

Product-Moment Coefficient

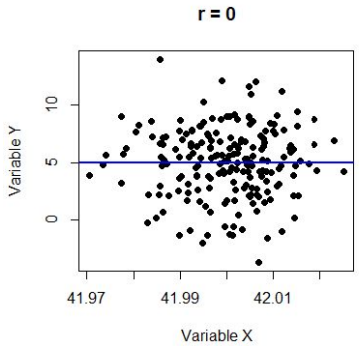
- Sign: (+ or -) gives the direction of the relationship
    - Positive: As one variable increases in magnitude, the other variable **increases**
    - Negative: As one variable increases in magnitude, the other variable **decreases**

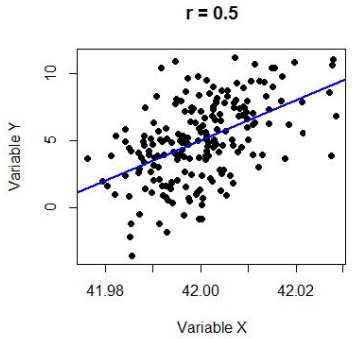
Product-Moment Coefficient

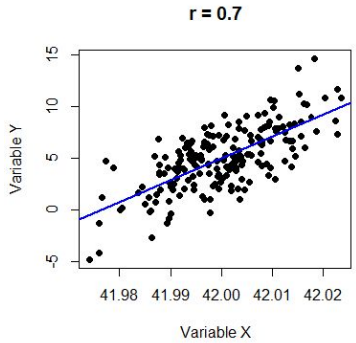
- Numeric value: Strength of the relationship
- Strength can be “seen” on a scatter plot
  - 0 : no relationship, circular shape
  - ~ 0.4 - 0.5 : shape of a “football”
  - ~ 0.7 - 0.8 : shape of a Zeppelin
  - ~ 0.9 : shape of a cigar
  - 1.0 : perfect line

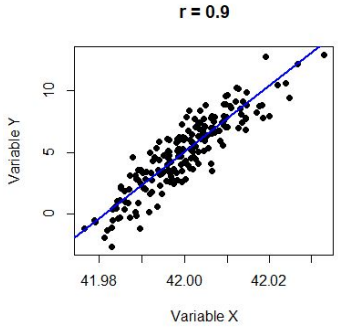


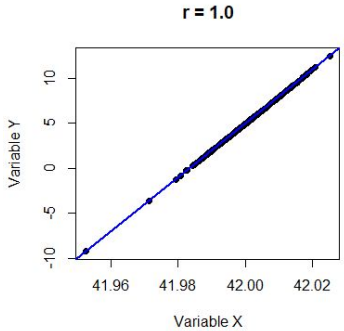
Scatter Plots

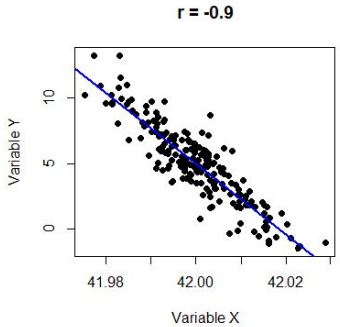












Product Moment Coefficient Example

- An engineering manager was studying a riveting process that assembles a sprocket on the nose-end of a chain saw bar.

- They reviewed the control chart for one of the major product characteristics, rivet height, which corresponds to the height the rivet protrudes above the bar surface.

Product Moment Coefficient Example

- Based upon a possible upward trend on the range chart, they became concerned about a possible increase in the variability of the process. They called a meeting with the project team to discuss the issue.

- One team member suggested that the rivet height variation might be attributed, in part, to the paint thickness around the rivet holes.

Product Moment Coefficient Example

- The paint is applied prior to the rivet crushing operation after which rivet height is measured.

- The team decided to design and run an appropriate study to determine if there was a relationship (correlation) between the two variables. The data for this example are in the file named **Rivet.txt**.

The One Sample t Test for Correlation

- Testing $\rho = 0$

- Underlying Assumptions
  - The pairs are randomly drawn from the population, assures independence of the Pairs of X,Y scores (Critical)
  - Populations are normally distributed

The One Sample t Test for Correlation

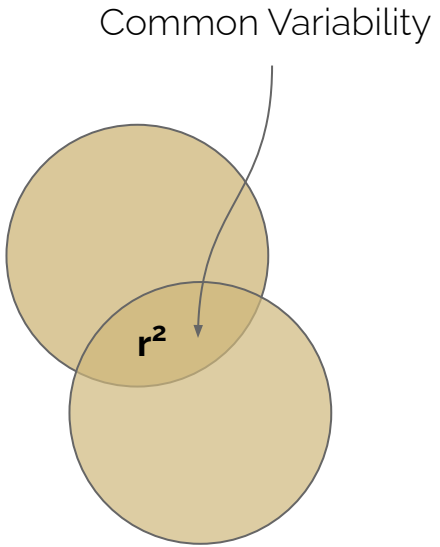
- Additional points concerning Pearson's r
  - Assesses **only** linear relationship
  - r estimates ρ
  - Statistical significance allows us to infer that $\rho \neq 0.0$
  - Statistical importance (r^2) is not equivalent to practical importance
  - Larger n does not guarantee larger values for r
  - Correlation is **not** causation!

The One Sample t Test for Correlation

- Testing the Hypotheses:
  - H_0: $\rho = 0$
  - H_1: $\rho \neq 0$
- The test statistic is a t
- The t has n - 2 degrees of freedom
- Another test exists if the hypothesized correlation is not zero

Coefficient of Determination, r^2

- Interpretation:
  - r^2 is an indication of the strength of the relationship
  - r^2 is the percent of variability the two variables have in common
  - r^2 is the amount of variability in Y that can be explained by the variability in X and the linear relationship between X and Y



The One Sample t Test for Correlation Example

- A manager wishes to determine whether there is a relationship between the viscosity of a cutting tool coolant and the amount of tool wear during a concurrent period.

- Calculating the Pearson r for data from **70** randomly selected periods of time, they finds a value of **r = 0.30**.

The One Sample t Test for Correlation Example

- Could this value have occurred due to chance, or is it reasonable to assume that a true relationship exists between these two variables?

- If you reject the hypothesis of independence, interpret the results of the study. What do you conclude? What action, if any, should be taken?

Step 1

1

- State the Research Question:
  Is there a relationship between the viscosity of a cutting tool coolant and the amount of tool wear?

Step 2

2

- Dependent Variable: ***Viscosity, Tool Wear***
- Criterion Measure: ***Viscosity Measurement***
  Wear Measurement
- Level of Data: ***Ratio, Ratio***
- Performance Criteria: ***Target is Best***, ***Smaller is Better***

Step 3

3

- State the Statistical Hypotheses.
  - H_o: $\rho_{xy} = 0$
  - H_1: $\rho_{xy} \neq 0$

Hypothesized Value

Non-Directional

Step 4

4

- Select the Statistical Test and Identify its RSD when H_0 is true
  - ***One sample test of the correlation using Pearson's r***
  - ***t ~ t (n-2 df) when H_0 is true***

Step 5

5

- Select the Type I and a Type II Error Rates and Decision for reject H_o
  - Type I Error and Its Consequence:
    If the null hypothesis is falsely rejected we would conclude that there is a relationship between coolant viscosity and tool wear when there is not. This may cause us to improperly change or try to control coolant viscosity when there is no underlying relationship to justify such activity.

Step 5

5

- Type II Error and Its Consequence:
  If we do not reject H0 when we should we would miss out on the opportunity to further investigate and possible exploit the relationship between these two variables.

Step 5

5

- Type I Error Rate: $\alpha = 0.05$
- Decision Rule for Rejecting H_0: $p \leq \alpha$

Step 6

6

- Validate the Underlying Assumptions
  - ***Random Sample, Independent specimens***
  - ***Normality of X and Y scores***

Step 6

6

- Perform a Basic Descriptive Analysis
  - Graphic - ***Scatterplot***
  - Numeric

    r = 0.30

    n = 70

Step 7

7

- Perform the Statistical Test and Obtain Its Probability (p-value)

Step 7

7

- Calculate the value of the test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.3}{\sqrt{\frac{1-0.3^2}{70-2}}} = 2.5933$$

Test Statistic

In RStudio

```
cor.pearson.r.onesample.simple( )
```

Levels of Importance

r^2 value (%)	Importance Level
70 - 100%	High Importance
50 – 69%	Moderate Importance
25 – 49%	Low Importance
< 25%	Unimportant

Step 8

8

- State the Statistical Conclusion with Regard to the Null Hypothesis, H_o. Provide Appropriate Estimates and Compute Power if needed.
  - ***Reject H_o***
  - Report the p value(s) (for each Hypothesis):
    p = 0.01163

Step 8

8

- WHSSETIT:
  We have sufficient statistical evidence to infer that a relationship exists.
- Importance: ***r^2 = 9%***
- Interval Estimate:
  The 95% confidence interval for r is 0.070 to 0.500.

Step 9

9

- Interpretation of the Results in Terms of the Research Question.
  A relationship exists between coolant viscosity and tool wear, but the variance of tool wear explained by the change in viscosity is only 9%

Fisher's Z_F Test for Correlation

$$Z_F = \frac{r_{tr} - \rho_{tr}}{\sqrt{\frac{1}{n-3}}}$$

- Testing $\rho = \rho_0$
- Where r_{tr} and ρ_{tr} are transformed values
  - $r_{tr} = 0.5 * \log((1+r)/(1-r))$
    - Where log is the R function for the base e natural log and
  - $r_{tr} = \text{atanh}(r)$
    - Where atanh is an R function

Fisher's Z_F Test for Correlation

Underlying Assumptions

- The n pairs of X,Y scores are independent of one another.
- Populations are normally distributed
- This test assumes that the null hypothesis to be tested is H0: $\rho_{XY} = \rho_0$, where ρ_0 is some non-zero value.

Fisher's Z_F Test for Correlation Example

- As a design engineer you "know" from examining past data that there is correlation between the "solderability" of the components on a ceramic substrate and the alumina content of the substrate which has been determined to be **0.62**.

- You are also aware of the fact that the higher the correlation coefficient, the lower the process defective rate at the end line.

Fisher's Z_F Test for Correlation Example

- In an effort to improve the process, you make a change, hoping to increase the strength of this correlation.

- Randomly selecting and testing a sample of **40** components from the changed process, the resulting Pearson correlation coefficient (r) is **0.75**.

Fisher's Z_F Test for Correlation Example

- Has the process been improved? That is, has the correlation been strengthened from its previous value of 0.62?

Step 1

1

- State the Research Question:
  Have the changes in the process increased the strength of the correlation between solderability and alumina content of the substrate which in the past has been documented to be $\rho = 0.62$.

Step 2

2

- Dependent Variable: ***Solderability, Substrate Quality***
- Criterion Measure: ***Solder Strength***
  Alumina Content
- Level of Data: ***Ratio, Ratio***
- Performance Criteria: ***N/A***

Step 3

3

- State the Statistical Hypotheses.
  - H_o: $\rho_{xy} \leq 0.62$
  - H_1: $\rho_{xy} > 0.62$

Hypothesized Value

Directional

Step 4

4

- Select the Statistical Test and Identify its RSD when H_0 is true
  - ***One sample test of the correlation using Pearson's r***
  - ***Z_F ~ approx. N(0,1/(n-3)) when H_0 is true***

Step 5

5

- Select the Type I and a Type II Error Rates and Decision for reject H_o
  - Type I Error and Its Consequence:
    By Rejecting H0 when it is True, one concludes that the correlation has changed, when in fact it has not. This could easily lead the researcher to incorrectly take action based on the presumed relationship. This could lead to making changes or promises based on the presumed relationship which would not materialize in the process over time.

Step 5

5

- Type II Error and Its Consequence:
  By Failing to Reject H0 when it is False, one concludes that the correlation has not changed, when in fact it has. Knowledge of the specific effect on the correlation of the process changes would be lost. This could easily lead to continuing efforts to impact the correlation which would result in a loss of time, effort and money.

Step 5

5

- Type I Error Rate: $\alpha = 0.05$
- Decision Rule for Rejecting H_0: $p \leq \alpha$

Step 6

6

- Validate the Underlying Assumptions
  - ***Independence of the pairs of scores***
  - ***Normality of X and Y scores***

Step 6

6

- Perform a Basic Descriptive Analysis
  - Graphic - ***Scatterplot***
  - Numeric

r = 0.75

n = 40

Sample Statistic

Step 7

7

- Perform the Statistical Test and Obtain Its Probability (p-value)

Step 7

7

- Calculate the value of the test statistic

$$Z_F = \frac{r_{tr} - \rho_{tr}}{\sqrt{\frac{1}{n-3}}} = 1.5082$$

Test Statistic

In RStudio

```
cor.pearson.r.onesample.simple( )
```

Step 8

8

- State the Statistical Conclusion with Regard to the Null Hypothesis, H_o. Provide Appropriate Estimates and Compute Power if needed.
  - ***Fail to Reject H_o***
  - Report the p value(s) (for each Hypothesis):
    p = 0.0657

Step 8

8

- WHSSETIT:
  We have sufficient statistical evidence to infer the population correlation between solderability and the alumina content of the substrate has NOT changed from its previous value.

Step 9

9

- Interpretation of the Results in Terms of the Research Question.
  We conclude that the relationship between Solderability and the Alumina Content of the substrate has not changed. As a result, we should not take action based on the value of this correlation coefficient.

Correlation for Ordinal Data

- Spearman's rank-order correlation, r_s
  - Scores are first converted to ranks
  - Ties are given the average rank
  - The product moment correlation coefficient is then calculated on the ranks
  - This may be calculated using a Pearson r on the ranks or as follows:

$$r_S = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Spearman's r_s

Example

A surface grinding operation is under study by an Engineering Manager. The objective is to determine whether or not there is a relationship between the number of finish passes we take on a particular component and the number of surface cracks revealed by dye penetrant testing after the grinding operation.

Does an association exist?

Spearman's r_s

Example

Spearman

	Cracks	Passes	Rank.Cracks	Rank.Passes
1	45	3	1	1.5
2	46	3	2	1.5
3	54	4	4	3.5
4	51	5	3	5.0
5	57	6	5	6.0
6	61	4	6	3.5
7	72	8	7	7.5
8	77	8	8	7.5
9	79	9	9	9.0

Testing Significance of Spearman's r_s

- When n is >= 20
  - Use the product moment formula
  - The test is a t with n - 2 degrees of freedom

- When n is < 20 you can use tables for exact critical r_s values

Step 1

1

- State the Research Question:
  Is there a relationship between the number of finish passes and the number of surface cracks revealed in a grinding operation?

Step 2

2

- Dependent Variable: ***Finish Passes, Surface Quality***
- Criterion Measure: ***# of Finish Passes***
  # of Surface Cracks
- Level of Data: ***Discrete Ordinal Count***
- Performance Criteria: ***N/A***

Step 3

3

- State the Statistical Hypotheses.
  - H_0: $\rho_s = 0$
  - H_1: $\rho_s \neq 0$

Hypothesized Value

Non-Directional

Step 4

4

- Select the Statistical Test and Identify its RSD when H_0 is true
  - ***One sample test for the Rank Order Correlation, rs. The test statistic is rs. (When n > 20, the test statistic is t.)***

Step 5

5

- Select the Type I and a Type II Error Rates and Decision for reject H_o
  - Type I Error and Its Consequence:
    Rejecting Ho when it is True would lead us to conclude that there is a relationship that may be beneficially used to exploit or predict a given result for one variable based on the values of the other. This would lead to conclusions that would not be supported by ongoing data.

Step 5

5

- Type II Error and Its Consequence:
  Failing to Reject H0 when it is False, or failing to detect a relationship when it truly exists would lead to the missed opportunity of being able to use information about the number of passes and the prediction of the cracks for improvement purposes.

Step 5

5

- Type I Error Rate: $\alpha = 0.05$
- Decision Rule for Rejecting H_0: **Reject H0 if |rS| is ≥ the rS critical value of 0.700.**

Step 6

6

- Validate the Underlying Assumptions
  - ***Independence of the pairs of scores***
  - ***The data values are measured on at least an ordinal scale***

Step 6

- Perform a Basic Descriptive Analysis
  - Graphic - ***Scatterplot***
  - Numeric

    $r_s = 0.895$

    $n = 9$

Sample Statistic

Step 7

7

- Perform the Statistical Test and Obtain Its Probability (p-value)

Step 7

7

- Calculate the value of the test statistic

```
> cor.spearman.rank(x1 = Spearman$Cracks, x2 = Spearman$Passes)

        Spearman Rank Correlation Coefficient

data:  data
t = 5.2962, null hypothesis rho_sp = 0, p-value = 0.001128
alternative hypothesis: true rho_sp is not equal to 0
95 percent confidence interval:
 0.4602525 0.9833885
sample estimates:
     r_sp        df
0.8945864 7.0000000
```

Test Statistic

In RStudio

```
cor.spearman.rank( )
```

Step 8

8

- State the Statistical Conclusion with Regard to the Null Hypothesis, H_o. Provide Appropriate Estimates and Compute Power if needed.
  - ***Reject H_o***
  - Report the p value(s) (for each Hypothesis):
    p = 0.001128

Step 8

8

- WHSSETIT:
  We have sufficient statistical evidence to infer that the population rank correlation, ρ_s is not 0.0. We estimate that its value is 0.8946.

Step 9

9

- Interpretation of the Results in Terms of the Research Question.
  Based on the data and the sampling, there is a significant relationship between the number of passes in grinding and the number of cracks detected with dye penetrant inspection. This relationship is positive, the more passes, the greater the number of cracks. Additional work should be conducted to discover the causal mechanism for this relationship and to identify possible corrective actions.

The Point-Biserial Correlation Coefficient, r_{pbi}

- Assesses the relationship between a continuous (interval or ratio scale) variable and a two-outcome nominal variable, also called a dichotomous variable.

The Point-Biserial Correlation Coefficient, r_{pbi}

Underlying Assumptions

- The pairs of scores are independent.
- The data values are continuous for one variable and a genuine dichotomy on the other variable.
- The scores on the continuous variable are normally distributed for each level of the dichotomous variable.
- The items within each of the levels of the dichotomous variable are independent of one another.

Point-Biserial Example

- We are manufacturing polymer sheet that has a critical characteristic of thickness (ratio scale data)
- We wish to determine whether the end-of-line sheet thickness is related to production line on which the sheet was produced, either a high-speed line (1) or a low-speed line (2).
- The data have been collected and are in the file named **PolyThk.txt**.We will perform the assessment using a Type I error rate of $\alpha = 0.10$.

Step 1

1

- State the Research Question:
  Is there a relationship between the end-of-line thickness of polymer sheet and the line that produced the material?

Step 2

2

- Dependent Variable: ***Sheet Thickness, Production Line***
- Criterion Measure: ***Thickness, Line Number***
- Level of Data: ***Continuous, Nominal***
- Performance Criteria: ***N/A***

Step 3

3

- State the Statistical Hypotheses.
  - H_o: $\rho_{pbi} = 0$
  - H_1: $\rho_{pbi} \neq 0$

Hypothesized Value

Non-Directional

Step 4

4

- Select the Statistical Test and Identify its RSD when H_0 is true
  - ***One sample test for the Point-Biserial Correlation, ρ_{pbi}***
  - ***$t \sim t$ (n-2 df) when H_0 is true***

Step 5

5

- Select the Type I and a Type II Error Rates and Decision for reject H_o
  - Type I Error and Its Consequence:
    Rejecting H0 when it is True would indicate that there is a relationship between line and sheet thickness when there is not. This could lead to acting on a presumed relationship, but the expected benefit would not materialize over time.

Step 5

5

- Type II Error and Its Consequence:
  Failing to reject H0 when it is False would lead to presuming there is no relationship when there is. This could lead to not exploiting a potentially beneficial relationship.

Step 5

5

- Type I Error Rate: $\alpha = 0.10$
- Decision Rule for Rejecting H_0: $p \leq \alpha$

Step 6

6

- Validate the Underlying Assumptions
  - ***The pairs of scores are independent.***
  - ***The data values are continuous for one variable and a genuine dichotomy for the other variable.***

Step 6

- Perform a Basic Descriptive Analysis
  - Graphic - ***Scatterplot***
  - Numeric

r_{pbi} = ***-0.517***

n = 15

Sample Statistic

Note: must use cor( ) to get sign (+ or -)

Step 7

7

- Perform the Statistical Test and Obtain Its Probability (p-value)

Step 7

7

- Calculate the value of the test statistic

```
> cor.test(x = PolyThk$Line
         ,y = PolyThk$Thick, conf.level = 0.90)
Pearson's product-moment correlation

data:  PolyThk$Line and PolyThk$EOLThick
t = -2.1804093, df = 13, p-value = 0.04819906
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 -0.78091473614 -0.09774098038
sample estimates:
          cor
-0.5174729577
```

Test Statistic

In RStudio

```
cor.test( )
```

Step 8

8

- State the Statistical Conclusion with Regard to the Null Hypothesis, H_o. Provide Appropriate Estimates and Compute Power if needed.
  - ***Reject H_o***
  - Report the p value(s) (for each Hypothesis):
    p = 0.0482

Step 8

8

- WHSSETIT:
  We have sufficient statistical evidence to infer that the population point-biserial correlation, ρ_{pbi} is not 0.0.
- Importance: ***r^2_{pbi}= 26.78%***
- Point Estimate: ***−0.5175***
- Interval Estimate: ***-0.7809, -0.0977***

Step 9

9

- Interpretation of the Results in Terms of the Research Question.
  Based on the data and the sampling, there is a significant and important relationship between the production line and the end of line thickness. The relationship is "negative" indicating that the higher "line number," (the low-speed line), is associated with smaller sheet thickness values.

Association for Nominal Data

- Assesses the degree of relationship between two nominal (aka two categorical) variables
  - Pearson's Phi (for 2x2)
  - Pearson's C (for square tables, larger than 2x2)
  - Cramer's V (for rectangular tables)

Categorical Data Formats in R

- Contingency Table (Cross Tabulation)
- You need to know how to get your data in this format

	K1	K2	K3
J1	2	1	2
J2	1	2	3

Categorical Data Formats in R

- Frequency Format

J	K	Freq
1	1	2
1	2	1
1	3	2
2	1	1
2	2	2
2	3	3

Categorical Data Formats in R

- Individual Format

J	K
1	1
1	1
1	2
1	3
1	3
2	1
2	2
2	2
2	3
2	3
2	3

Measures of Association for Categorical Data

In RStudio

- `cor.pearson.phi(table)`
- `cor.pearson.c(table)`
- `cor.cramer.v(table)`

n = total number of specimens

m is the smaller of the number of Rows and columns in the contingency table when m ≥ 2

$$\text{Pearson's } \phi = \sqrt{\frac{\chi^2}{n}}$$

$$\text{Pearson's } C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

$$\text{Cramér's } V = \sqrt{\frac{\chi^2}{n(m-1)}}$$

χ^2 Test

- Test Statistic

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- Where
  - Oi = Observed frequency (count)
  - Ei = Expected frequency (count)

Association (Correlation) for Nominal Data

- A quality improvement team is attempting to determine whether a relationship exists between ingot cracking and the presence or absence of a filter used during the casting process.
- The data collected are listed below and are contained in the file named **Casting.txt**. We will use an **α** of 0.05.
- Note, for the data below, for the Cracked variable, 1 = Cracked, 2 = Not Cracked. For the Filter variable, 1 = In, 2 = Out.
- The third variable, Count, represents the number of cases in each combination of these two variables.

Association (Correlation) for Nominal Data

- Before we get going, we need to create a contingency table.
- Let's use the following "Standard" format

	Cracked = 1	Not Cracked = 2	Total
Filter = 1	A	B	
No Filter = 2	C	D	
Total			

	Cracked = 1	Not Cracked = 2	Total
Filter = 1	**A**	**B**	
No Filter = 2	**C**	**D**	
Total			

1. What value goes Cell A?
   a. 30
   b. 59
   c. 12
   d. 18

2. What value goes Cell B?
   a. 30
   b. 59
   c. 12
   d. 18

	Cracked = 1	Not Cracked = 2	Total
Filter = 1	**A**	**B**	
No Filter = 2	**C**	**D**	
Total			

3. What value goes Cell C?
   a. 64
   b. 47
   c. 46
   d. 93

4. What value goes Cell D?
   a. 64
   b. 47
   c. 46
   d. 93

Creating the Contingency Table

- Create: ct <- transform.independent.format.to.xt(x_row = Casting$Cracked, x_col = Casting$Filter, weight = Casting$Count) # Converts from Freq format to Table
- Or: casting.table <- matrix(c(12,18, 47,46), nrow = 2) # Fills by Columns
- Or: casting.table <- matrix(c(12, 47, 18, 46), nrow = 2, byrow = TRUE) # Fills by rows
- Now, conduct proper test: Phi, C, or V

Cramer's V Example

	Defect Type				
Customer Type	Blister	Crack	Pit	Scratch	**Total**
Automotive	25	6	12	33	76
Appliance	20	10	10	25	65
Internal	4	7	41	13	65
Total	49	23	63	71	206

Step 1

1

- State the Research Question:
  Is there a relationship between Customer Type and Defect Type?

Step 2

2

- Dependent Variable: ***Customer Type, Defect Type***
- Criterion Measure: ***Customer Category***
  Defect Category
- Level of Data: ***Nominal***
- Performance Criteria: ***N/A***

Step 3

3

- State the Statistical Hypotheses.
  - H_0: $\chi^2 = 0$
  - H_1: $\chi^2 \neq 0$

Hypothesized Value

Non-Directional

Step 4

4

- Select the Statistical Test and Identify its RSD when H_0 is true
  - ***Cramer's V Test for Association, V***
  - ***$\chi^2 \sim \chi^2$ (df = (r-1)(c-1), when Ho is True (where r is the number of rows and c is the number of columns).***

Step 5

5

- Select the Type I and a Type II Error Rates and Decision for reject H_o
  - Type I Error and Its Consequence:
    Rejecting H0 when it is True would indicate that there is a relationship between customer type and defect type when there is not. This could lead to acting on a presumed relationship, but the expected benefit would not materialize over time.

Step 5

5

- Type II Error and Its Consequence:
  Failing to reject H0 when it is False would lead to presuming there is no relationship when there is. This could lead to not investigating a potentially harmful relationship.

Step 5

5

- Type I Error Rate: $\alpha = 0.05$
- Decision Rule for Rejecting H_0: $p \leq \alpha$

Step 6

6

- Validate the Underlying Assumptions
  - ***The pairs of scores are independent.***
  - ***The table is rectangular***

Step 6

6

- Perform a Basic Descriptive Analysis
  - Graphic/Numeric - ***Contingency Table***

Step 7

7

- Perform the Statistical Test and Obtain Its Probability (p-value)

Step 7

7

- Calculate the value of the test statistic

$$\text{Cramér's } V = \sqrt{\frac{\chi^2}{n(m-1)}}$$

```
> cor.cramer.v(cust.def)

        Cramer's V and Pearson Chi-Squared

data:  data
V = 0.35886, null hypothesis V = 0, p-value = 0.000000001142
alternative hypothesis: true V is not equal to 0
sample estimates:
        chi.square       chi.square.p       min.exp.freq
53.057917450751347  0.000000001142199  7.257281553398058
```

Test Statistic

Step 8

8

- State the Statistical Conclusion with Regard to the Null Hypothesis, H_o. Provide Appropriate Estimates and Compute Power if needed.
  - ***Reject H_o***
  - Report the p value(s) (for each Hypothesis):
    p = 0.0000

Step 8

8

- WHSSETIT:
  We have sufficient statistical evidence to infer that a relationship between the categorical variables exists.

Step 9

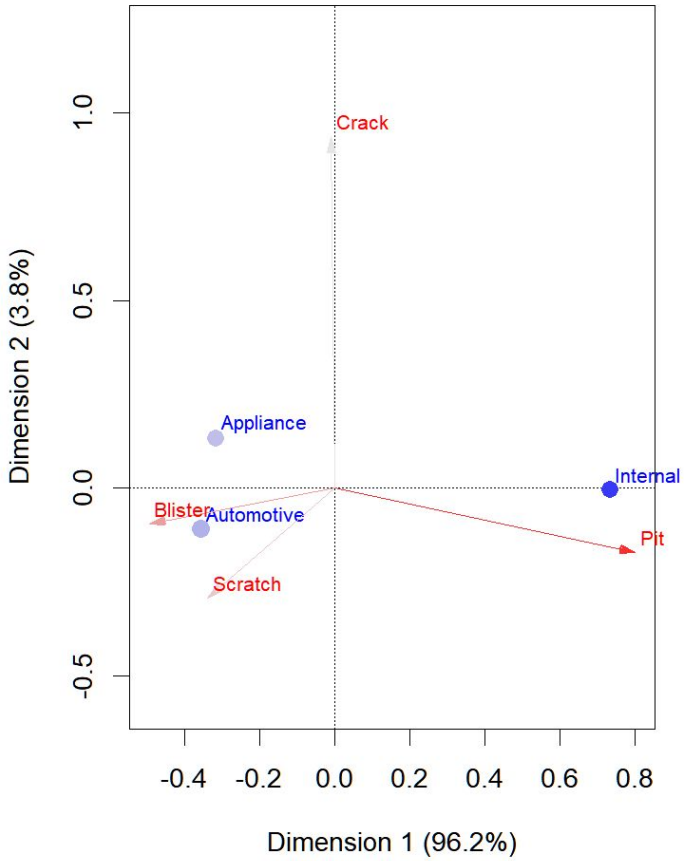
9

- Interpretation of the Results in Terms of the Research Question.
  Based on the data and the sampling, there is a relationship between customer type and defect type.

Step 9

9

- Interpretation of the Results in Terms of the Research Question.
  Based on the data and the sampling, there is a relationship between customer type and defect type.


Crack
Appliance
Internal
Blister
Automotive
Pit
Scratch
1.0
0.5
0.0
-0.5
Dimension 2 (3.8%)
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
Dimension 1 (96.2%)