**Questions and Answers**

**Introduction to Statistics and Data Analysis – A Case-Based Approach**

Conrad Ziller, University of Duisburg-Essen

Content

[1 – Univariate Statistics 2](#_Toc160628954)

[Q1: 2](#_Toc160628955)

[Q2: 2](#_Toc160628956)

[Q3: 2](#_Toc160628957)

[Q4: 2](#_Toc160628958)

[2 – Bivariate Statistics 3](#_Toc160628959)

[Q1: 3](#_Toc160628960)

[Q2: 3](#_Toc160628961)

[Q3: 3](#_Toc160628962)

[Q4: 4](#_Toc160628963)

[Q5: 4](#_Toc160628964)

[Q6: 4](#_Toc160628965)

[Q7: 4](#_Toc160628966)

[Q8: 4](#_Toc160628967)

[Q9: 4](#_Toc160628968)

[3 – Statistical Inference 5](#_Toc160628969)

[Q1: 5](#_Toc160628970)

[Q2: 5](#_Toc160628971)

[Q3: 5](#_Toc160628972)

[Q4: 5](#_Toc160628973)

[Q5: 5](#_Toc160628974)

[Q6: 6](#_Toc160628975)

[Q7: 6](#_Toc160628976)

[Q8: 6](#_Toc160628977)

[Q9: 6](#_Toc160628978)

[4 – Regression Analysis 7](#_Toc160628979)

[Q1: 7](#_Toc160628980)

[Q2: 7](#_Toc160628981)

[Q3: 7](#_Toc160628982)

[Q4: 7](#_Toc160628983)

[Q5: 8](#_Toc160628984)

[Q6: 8](#_Toc160628985)

[Q7: 8](#_Toc160628986)

[Q8: 9](#_Toc160628987)

[Q9: 9](#_Toc160628988)

# 1 – Univariate Statistics

## Q1:

* What information is conveyed by the map? What information remains implicit?

**Answer:**

* Maps give a good first impression of the spatial distribution of data.
* However, insights depend on further (implicit) information (“What are the specifics of the high-unemployment cities in the middle of the map? Which areas are rural, which are urban?”).
* Difficult to make general statements about maps and to compare maps (especially if the breaks that underly the colorization differ).

## Q2:

* Which characteristics of the data can be identified with histograms? Which aspects remain hidden?

**Answer:**

* What can be seen: The distribution of the data, including the center(s) where most of the observations are located, patterns of skewness, and outliers.
* What remains hidden: The specific estimates of the summary statistics (e.g., mean, variance).

## Q3:

* Interpret two means of your choice. Why is no mean displayed for the variable area?

**Answer:**

* The average unemployment rate across the observed districts is 7.4 percent. The mean of the population variable is 338,218, which means that about 338,218 people live in a region, on average.
* Means can only be calculated for metric or quasi-metric (typically an ordinal variable with five or more categories) variables. area is a nominal variable.

## Q4:

* Can you spot which district has the highest unemployment rate? Which county has the lowest crime rate? Can we infer from the graphs which of the two variables is more dispersed?

**Answer:**

* It is quite hard to match the color of the dots with the ones of the legend. Gelsenkirchen has the highest unemployment rate at 14.9%. The district with the lowest crime rate is Lippe.
* Because of the different scaling of the variables, it is not possible to infer from this plot which variable is more dispersed. To do so, we would need to calculate the so-called coefficient of variation (see below).

# 

# 2 – Bivariate Statistics

## Q1:

* What is the level of measurement (nominal, ordinal, metric) of each variable?

**Answer:**

* vote and trump –> nominal (i.e., information on if something is either the case or not)
* education –> ordinal (i.e., values can be ranked, information on whether a variable value is “higher” or “lower”)
* age –> metric (i.e., interval between ranked variable values are comparable)

## Q2:

* How well did the results from the survey predict the actual election outcome? Can we conclude from that whether or not the sample is representative?

**Answer:**

* In the 2020 US presidential election, Biden won the popular vote against Trump by 51.3% to 46.9%. The survey results indicate that at the time the survey was conducted, 55.5% would have voted for Biden (compared to 44.5% for Trump). While the survey correctly predicted Biden as the winner, the result deviated from the actual election outcome by about 4 percentage points. Inferring from samples to the underlying population is always subject to some uncertainty. We can quantify margins of uncertainty, and we will do so in the case study on statistical inference. Beyond the role of statistical uncertainty, several other factors (e.g., political campaigning or situational circumstances on voting day) might have contributed to the outcome of the election result and could possibly account for the difference between the survey results and the actual vote.
* The representativeness of a survey relates to its features, such as random sampling of respondents and no systematic non-response of those being interviewed. For the American Election Study, we have good reasons to assume that these features are given and the results are representative of the US population.

## Q3:

* Which conclusions can be drawn from a cross tab with row percentages?

**Answer:**

* Row percentages are the number of observations for each cell divided by the total number of observations for each row (adding percentages per row adds up to 100%.). The interpretation would be similar to before, except that we now focus on comparing different categories of the row variable for one selected value of the column variable. Let’s select education = 1: Among those who do not support Trump (= Biden supporters), 15.3% have a low education profile, while among Trump supporters 23.1% have a low education profile. The difference of 7.8 percentage points suggests that Trump support and low education are systematically associated. Hence, the conclusion from the cross tab analysis with row percentages is equivalent to that one from column percentages (given that we adjust the interpretation scheme). Moreover, row percentages in this case also show the descriptive distribution of the education variable within each category of the vote variable. However, it is often not easy to compare the distributions at one glance.

## Q4:

Which conclusions can be drawn from a cross tab with cell percentages?

**Answer:**

* Total percentages are the number of observations for each cell divided by the total number of observations. This shows the relative frequency of each combination of values of the two variables. The goal is a descriptive understanding of the distribution of cases (Note: This is rarely used in applied research.).

## Q5:

* What can be inferred from the graphs?

**Answer:**

* The age distribution of the respondents supporting Biden (left-hand) is a bit more even than the distribution of those supporting Trump (right-hand side). The latter is also more skewed to the left and its center is more on the right, which indicates that respondents who support Trump are older, on average, than those who support Biden.

## Q6:

* Interpret the cross tab. What can be concluded? Why don’t we always go straight to categorizing metric variables and using the recoded variable in a cross tab?

**Answer:**

## Looking at the category of young respondents (age\_cat = 1), 36.9% prefer Trump as president, whereas among old respondents (age\_cat = 3), 47.6% opt for Trump. This difference of 10.7 percentage points indicates a systematic (positive) empirical relationship between age and Trump support: Older respondents tend to prefer Trump more than younger voters.

## The reason why it is not always recommended to categorize metric variables is that the process of categorization involves a loss of information. Usually, more information is better as it produces more detailed results. Note that there are exceptions to this such as striving for a simplified description of empirical relationships.

## Q7:

What can be infered from the figure?

**Answer:**

* The pattern indicates a negative relationship between the two variables. If the proportion of highly educated people in a region increases (moving to the right on the x-axis), we see that this is related to lower support for Trump (lower scores on the y-axis).

## Q8:

* How can Cramér’s V be interpreted in this case?

**Answer:**

* Interpretation: There is an association between the variables education and trump. In our case, the association is negative (the higher the education, the lower the approval of Trump, on average). Given the rule of thumb for small tables, the association of 0.17 is about weak to moderate.

## Q9:

* Why didn’t we calculate the Pearson’s correlation coefficient instead?

**Answer:**

* Pearson’s correlation coefficient requires (quasi-)metric variables with several categories. For ordinal variables, we have tools such as Kendall’s Tau or Spearman’s correlation coefficient (see here for more on different correlation types: https://ademos.people.uic.edu/Chapter22.html). If nominal variables are involved, we use coefficients such as Cramér’s V.

# 

# 3 – Statistical Inference

## Q1:

* Interpret the confidence interval.

**Answer:**

* **Practical interpretation:** Likely values for the population mean are within the limits of 4.2 and 4.4.

## Q2:

* Interpret the confidence interval.

**Answer:**

* **Practical interpretation:** Likely values for the population proportion are within the limits of 0.29 and 0.33.

## Q3:

* Interpret the confidence interval.

**Answer:**

* **Interpretation:** Likely values for correlation in the population are within the limits of -0.003 and 0.08.

## Q4:

* The shown figure displays 95% confidence intervals. Would 90% confidence intervals be narrower or wider?

**Answer:**

* The 90% confidence intervals would be narrower. Going back to the probability-world interpretation: If we only want to capture the true population parameter 90 percent of the time (instead of 95), this means that we could narrow down the range of values in which the parameter will fall. In contrast, if we decrease our readiness to err to 1% (i.e., a 99% confidence interval), the range of values to fulfill this assumption would increase - the confidence interval would get wider.

## Q5:

* Formulate the null and alternative hypotheses if we assume that persons with a high-income are more satisfied with the government.

**Answer:**

Typically, it is easier to start with formulating the alternative hypothesis.

* : High-income individuals are more satisfied with government than low-income individuals (i.e., positive relationship between income and satisfaction with government).
* Formal: or (equivalent)
* : High-income individuals are equally or less satisfied with government than low-income individuals (i.e., null or negative relationship between income and satisfaction with government).
* Formal: or (equivalent)
* Note that it is important that all possible outcomes are covered by both hypotheses. This means that when formulating a directed hypothesis, the null hypothesis must include the case of no relationship (i.e., “” instead of just “”).

## Q6:

* How would the rejection area change if we test one-sided? What would change if we set a 1% probability of error (two-sided test)?

**Answer:**

* Assuming a positive mean difference or correlation, the rejection area would be located only on the right side of the distribution. Since we still use an error probability of 5%, this area would be larger than in the two-sided case. Consequently, the critical value c thus moves a bit in the direction of the center of the distribution. For the z-distribution, this value is +1.645 for a positive alternative hypothesis (and -1.645 for a negative hypothesis).
* If we change the probability of error to 1% and test two-sided, we have a rejection range that is smaller compared to the 5% case. Our estimate and the corresponding z-statistic need to be larger to reject the null hypothesis compared to the 5%-case. In the 1% case, the critical values are -2.326 and +2.326.
* Critical values at which one can reject the null hypothesis can be obtained from here: <https://www.criticalvaluecalculator.com/>

## Q7:

* Interpret the result of the statistical test in detail with reference to the specific p-value.

**Answer:**

* The probability of obtaining such a result from the sample, although 𝐻00 is true in the population, is 0.11 or 11%. Since this exceeds the threshold of our willingness to err of 5%, we cannot reject the null hypothesis (and thus find no evidence in favor of 𝐻𝐴��).
* Even if we take into account that we formulated a directed hypothesis and rely on a one-sided test (we can thus divide the p-value by 2), the p-value still exceeds our assumed probability of error (p-value of 0.055 > 𝛼� of 0.05). Hence, even with a one-sided test, we are not able to reject the null hypothesis.

## Q8:

* Assume that higher income is either positively or negatively related to satisfaction with government. Formulate the null and alternative hypotheses (undirected).

**Answer:**

* 𝐻𝐴: Verbal: Income is either positively or negatively related to satisfaction with government.
* Formal: 𝑟≠0≠0
* 𝐻0: Verbal: Income is unrelated to satisfaction with government.
* Formal: 𝑟=0

## Q9:

* Interpret the result of the statistical test (two-sided test) with reference to the t-value and p-value.

**Answer:**

* The probability of obtaining such a result from the sample, although 𝐻0 is true in the population, is 0.07 or 7%. Since this exceeds the threshold of our willingness to err of 5%, the result is not statistically significant and we cannot reject the null hypothesis (and thus find no evidence in favor of 𝐻𝐴).

# 4 – Regression Analysis

## Q1:

* Could we have recoded the variables differently? What would change? And what would happen if, for example, we recoded only two of the three variables and then built an index with all three variables?

**Answer:**

* Yes, if we were to recode all items, that would be just as possible. We would then have to label an index as “less justice” or “inequality,” since high values represent acceptance of social inequality. Recoding only a few variables is inadmissible, the index would no longer be valid.

## Q2:

* What is the value of the constant in the unemployment-crime example (Model 1)? What information content does it carry?

**Answer:**

* The value of the constant in Model 1 is 2399. This indicates the average crime rate (Y) in a municipality with 0% unemployment. In our case, the constant is uninformative because we don’t have observations with 0% unemployment in our sample.

## Q3:

* Would it also make sense to formulate a different hypothesis 𝐻𝐴?

**Answer:**

* Yes: For example, the higher the income, the higher the preference for social justice (since wealthier people can afford it). Which hypothesis makes sense depends on theory and previous research.

## Q4:

* Interpret the output from the regression model using the following heuristics.
* “The empirical relationship between income and attitudes toward justice is positive/negative.”
* “The relationship is/is not statistically significant.”
* Use the technically correct interpretation of p-values: “The probability of finding such a relationship in the population, …”
* “Thus, the null hypothesis can be rejected/not be rejected.”
* The detailed interpretation of the regression coefficient for income is: “…”
* Interpret additionally the constant and R2 from the model output. What is the value of the constant in the unemployment-crime example (Model 1)? What information content does it carry?

**Answer:**

* The empirical relationship between income and attitudes toward justice is negative.
* The found relationship is statistically significant (given a probability of error of 0.01 or 1% due to \*\*\*).
* The probability of finding such a relationship in the population, even though the null hypothesis is true, is less than 1%. Thus, the null hypothesis can be rejected.
* The detailed interpretation of the regression coefficient for income (b=-0.018) is as follows: If income increases by one unit (income category), then (variant 1:) attitudes toward justice decrease (on average) by 0.018 units (scale points) / (variant 2:) changes by -0.018 units.
* Constant: An interpretation is tedious because the income variable we used has no zero point. (If we would center the income variable at the mean value, then 0 indicated an average income, and the constant was the predicted level of justice attitudes for individuals with an average income.)
* R2 is a measure of model fit and indicates how well the included independent variables explain the outcome. It can be interpreted as the ratio of explained variance by the model to the total variance: R2=0.016 -> 1.6% of the total variance of the outcome variable can be explained by the predictor variables in the model. If R2 = 0.3, then the variance explained by the model would be 30%.
* Note: The so-called adjusted R2 is used to compare models with a different number of predictor variables. While it can no longer be interpreted as % of variance explained, it can be compared across models (better model with higher Adj-R2).

## Q5:

* How relate past or current unemployment and living in East Germany to both variables of interest income and attitudes toward justice?

**Answer:**

* Unemployment: If someone was/is unemployed, he/she is expected to advocate social justice more strongly (e.g. due to self-interest) than someone lacking such an experience. If someone is unemployed, he/she has a lower income, on average, compared to someone who is employed.
* East Germany: East Germans are expected to advocate social justice more strongly than West Germans because of a greater sense of disadvantage or their socialization in the socialist GDR. At the same time, residents of East Germany have a higher probability to earn less money in comparable jobs due to a less developed economic situation.

## Q6:

* Formulate the null and alternative hypotheses for each independent variable.

**Answer:**

* Income 𝐻𝐴: 𝛽1<0 (Income is negatively related to attitudes toward justice)

𝐻0: 𝛽1≥0 (Income is unrelated or positively related to attitudes toward justice)

* Female 𝐻𝐴: 𝛽2>0 (Being female is positively related to attitudes toward justice, which is the same as women advocate social justice more strongly than men)

𝐻0: 𝛽2≤0 (Being female is unrelated or negatively related to attitudes toward justice)

* Unemployed 𝐻𝐴: 𝛽3>0 (Unemployment is positively related to attitudes toward justice, which is the same as people who experienced unemployment advocate social justice more strongly than those without such an experience)

𝐻0�0: 𝛽3≤0 (Unemployment is unrelated or negatively related to attitudes toward justice)

* East Germany 𝐻𝐴: 𝛽4>0 (Living in East Germany is positively related to attitudes toward justice, which is the same as people from East Germany advocate social justice more strongly than those from West Germany)

𝐻0�0: 𝛽4≤0 (Living in East Germany is unrelated or negatively related to attitudes toward justice)

## Q7:

* Interpret the coefficients from the multiple regression model in terms of (a) direction of association, (b) significance, and (c) effect size.

**Answer:**

* Income

*Negative relationship between income and attitudes toward justice, which is statistically significant (p < 0.01). An increase in income by one unit is associated with a****decrease****in attitudes toward justice by 0.011 scale units (while keeping the influence of the other variables in the model constant). In terms of effect size, moving from the lowest to the highest income category would be associated with a decrease in attitudes toward justice by 0.286 scale units (26 categories \* 0.011 for the coefficient = 0.286, which corresponds to the effect of a min-max normalized income variable).*

* Female

*Positive relationship between being female and attitudes toward justice, which is statistically significant (p < 0.05). An increase in the variable female by one unit (= the difference between women and men) is associated with an****increase****in attitudes toward justice by 0.128 scale units (while keeping the influence of the other variables in the model constant).*

* Unemployed

*Positive relationship between being unemployed in the last 10 years and attitudes toward justice, which is statistically significant (p < 0.05). An increase in the variable unemployment by one unit (= the difference between those who were unemployed and those who were not) is associated with an****increase****in attitudes toward justice by 0.066 scale units (while keeping the influence of the other variables in the model constant).*

* East Germany

*Positive relationship between living in East Germany and attitudes toward justice, which is statistically significant (p < 0.05). An increase in the variable east by one unit (= the difference between people from East vs. West Germany) is associated with an****increase****in attitudes toward justice by 0.075 scale units (while keeping the influence of the other variables in the model constant).*

* In terms of effect comparison, the variables female, unemployed, and east are measured on a comparable scale (0/1). The min-max effect of income is also comparable as this corresponds to recoding the variable to range between 0 and 1. Hence, income is comparatively the strongest predictor in the multiple regression model.

## Q8:

* What changed between the bivariate and the multiple model?

**Answer:**

* In the multiple model (compared to the bivariate model), the coefficient of income (or the relationship between income and attitudes toward justice) is smaller. The underlying reason is that in the bivariate model, the coefficient partly represented unobserved factors, which is corrected in the multiple regression model (at least with respect to the variables included in the model).
* In terms of model fit, the multiple model has a better fit to the data (since adj. R2 is higher).

## Q9:

How would the effect be interpreted if both signs were negative?

**Answer:**

* The effect of income on attitudes toward inequality would be negative and decrease for high values of income.