Univariate Analysis

IRIII.2 – Quantitative Methods in the Study of International Relations

Steven V. Miller

Department of Economic History and International Relations

Stockholm University

Goal(s) for Today

1. Introduce basic structure of the course
2. Emphasize the qualitative component of international relations.
3. Discuss some basic univariate methods/statistics you should know.

Course Information

Hej! Jag heter Steve och jag talar inte mycket bra svenska. (Förlåt!)

- Mi español es aceptable como turista.
- I'm also learning Korean.
- (How about we do this in English instead...)

Format: three lectures; seven labs

- See course description for more information about assignments
- You can do these in Swedish (though I prefer English).

When in doubt, read the course description (it's 14 pages!)

Qualitative and Quantitative: What Are These Terms?

Qualitative: the analysis of *non-numerical* data to understand social phenomena

Quantitative: the analysis of *numerical* data to understand social phenomena

Stages of Research

0. Perspective
1. Causal theory
2. Hypothesis
3. Empirical test
4. Evaluation of hypothesis
5. Evaluation of causal theory
6. Advance in scientific knowledge

Perspective Matters

A perspective is a general orientation to the world. They're untestable because:

1. They're too broad. Empirical support will never be total.
2. Perspectives are slippery and contextual.
    - e.g. "People are rational."
3. Any empirical data observed can be interpreted to fit the perspective.

We start with perspectives because we're not blank slates.

Perspective Matters

A perspective is a general orientation to the world. They're untestable because:

1. They're too broad. Empirical support will never be total.
2. Perspectives are slippery and contextual.
    - e.g. "Government should stay out of our lives."
3. Any empirical data observed can be interpreted to fit the perspective.

We start with perspectives because we're not blank slates.

- Rationality, for example, informs our general theories and the data-collection process.

Measuring a Conceptual World

Our world is fundamentally conceptual (qualitative)

- We start with an interest in concepts we observe (e.g. “political tolerance”, “corruption”, “war”)
- We devise a conceptual definition of what that thing is.
- We *operationalize* a definition of the thing to measure it.

From this, we get an empirical *measure* of the concept.

- This allows us to proceed with political *science*.

Concept and Measure

We seek to devise the best measure that best captures the "true" concept.

- However, there's always some slippage between concept and measure.
- We do our best to eliminate as much error as we can.

There are two types of measurement error.

1. **Systematic measurement error**: the chronic, consistent distortion of a measure, leading to a *mis*measure of the concept in question.
2. **Random measurement error**: haphazard, chaotic distortion of a measure, leading to an inconsistent operational reading of the concept.

Systematic Measurement Error Cases

Systematic measurement error is always the bigger concern of the two because it can confound inference. Examples:

- Measuring ideology by party support.
- Measuring corruption by mass-level perception of corruption (or by indicators like arrests).
- Measuring human rights records with U.S. state department reports.

Detecting this is not always easy. You have to use your head.

“Political Tolerance” as a Classic Case

Figure 1: Scenes from the Tenth Communist Party USA convention in Chicago (May 1938)

Validity and Reliability

This distinction maps nicely onto conversations about reliability and validity.

- **Validity**: i.e. am I measuring what I want without picking up anything else?
- **Reliability**: am I consistently measuring what I want to measure?

Likewise, validity is the greater concern of the two.

Types of Variables

Assume you have a measure of your concept, you can summarize it any number of ways.

- However, it's contingent on what information you're measure can communicate.

The Classic Typology, for Better or Worse

The classic typology, a la Stanley Smith Stevens.

1. Nominal (i.e. unordered-categorical)
2. Ordinal (i.e. ordered-categorical)
3. Interval (i.e. continuous)
4. Ratio (i.e. continuous, but with meaningful zero as a kind of bound)

There are important wrinkles to this, so let's get some examples.

Table 1: GATT Members in Kono (2006)

In GATT?	Countries
0	Albania, Algeria, Belarus, Bhutan, Estonia, Ethiopia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Madagascar, Moldova, Nepal, Oman, PRC, Russia, Saudi Arabia
1	Argentina, Australia, Austria, Bangladesh, Bolivia, Brazil, Cameroon, Canada, Cent. Af. Rep., Chad, Chile, Costa Rica, Czech Rep., Ecuador, Egypt, El Salvador, Finland, Ghana, Guatemala, Honduras, Hungary, Iceland, India, Indonesia, Japan, Kenya, Malawi, Malaysia, Mauritius, Mexico, Morocco, Mozambique, New Zealand, Nicaragua, Nigeria, Norway, P. N. Guinea, Paraguay, Philippines, Poland, ROK, S. Africa, Singapore, Slovenia, Sri Lanka, Sweden, Switzerland, Tanzania, Thailand, Trinidad-Tobago, Tunisia, Turkey, USA, Uganda, Uruguay, Venezuela, Zambia, Zimbabwe

Note:

FYI: You will see these data again in one of your problem sets.

Table 2: Groups in the EU Parliament from a 2024 Vote on Ukraine

Group Label	No. of MEPs
European Conservatives and Reformists	68
European People's Party	177
Greens/European Free Alliance	72
Identity and Democracy	59
Non-attached Members	50
Progressive Alliance of Socialists and Democrats	140
Renew Europe	102
The Left in the European Parliament – GUE/NGL	37

Note:

Data: ?eu_ua_fta24 in {stevedata}

Unordered-Categorical Data

Unordered-categorical data takes on a few forms.

- *"Dummy" variable*: has just two values.
- *Nominal variable*: has multiple values where one category is not the other.

Don't be fooled by order/information you perceive in a dummy variable.

- They're just a special case of a nominal variable.

Table 3: Financial Satisfaction in South Korea, 2023

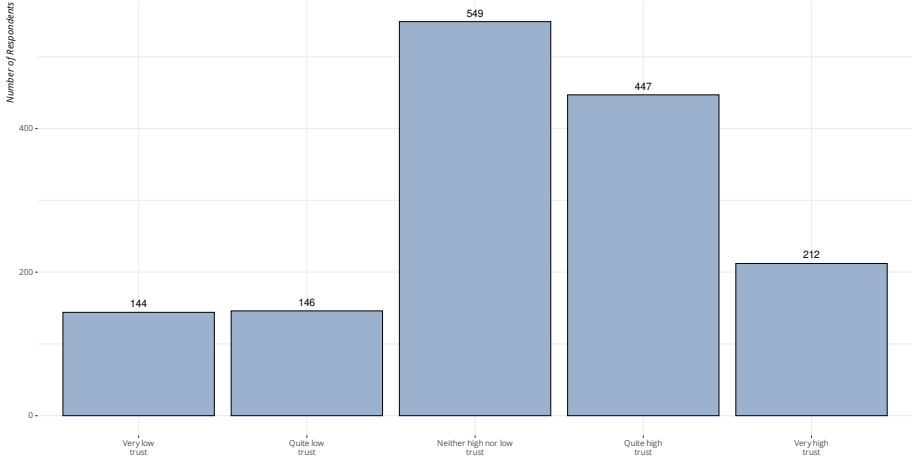
Financial Satisfaction	No.	Cum. Sum	%
Very Dissatisfied	59	59	5.25%
Somewhat Dissatisfied	271	330	29.39%
Neither Satisfied nor Dissatisfied	472	802	71.42%
Somewhat Satisfied	291	1093	97.33%
Very Satisfied	30	1123	100%

Note:

Data: Korean General Social Survey, 2023 (by way of ?kgss_sample in {simqi}).

Trust in the Royal Family in Sweden, 2020

Relatively few Swedes say they have low trust in the royal family. Notice, though, the available responses that Swedish respondents could select.



Data: SOM, 2020 (by way of ?som_sample in {simoji}).

Ordered-Categorical Data

Ordered-categorical data have an order/rank, but:

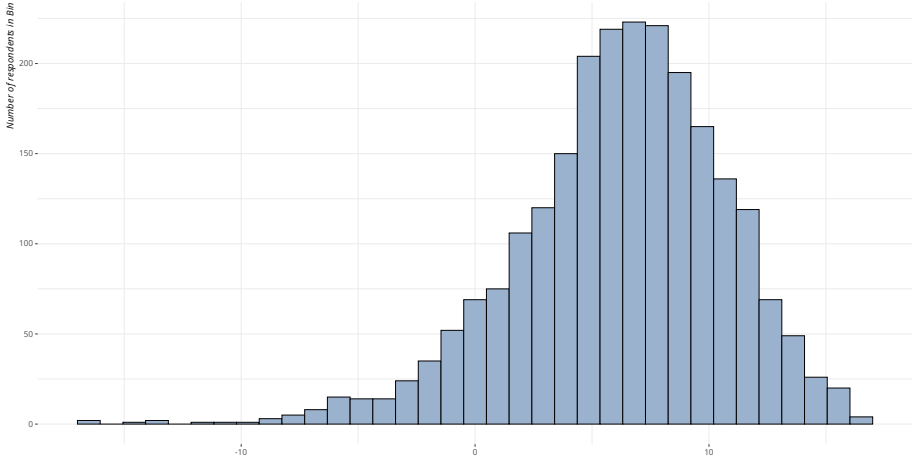
- A finite set of available responses
- No consistent difference between categories.

You'll see these kind of data often on:

- Questions of political support/trust/confidence
- Likert items (with 5- or 7-point agree/disagree scale)
- Assessments of political/financial satisfaction

American Attitudes toward Government Spending, 2018

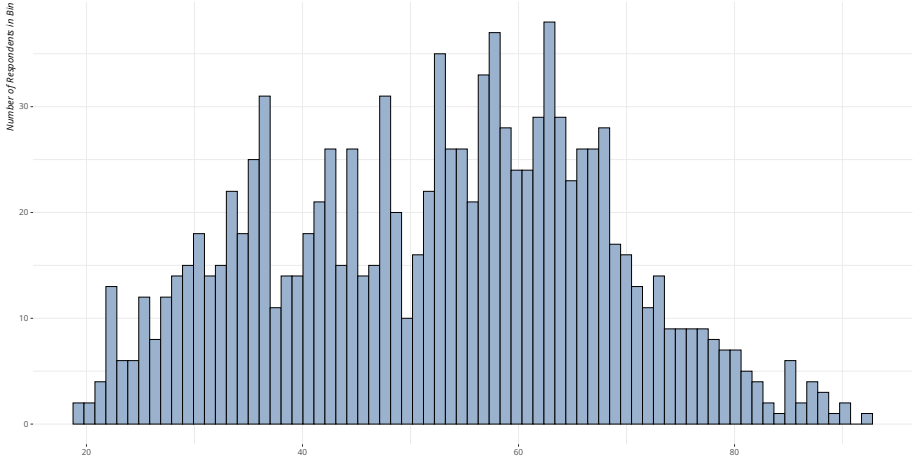
Attitudes about whether the U.S. is spending too much, too little, or not enough on various programs approximates a normal distribution.



Data: General Social Survey, 2018 (by way of ?gss_spending in {stevedata}). Histograms do have arbitrary bins that you should consider. The U.S. definitely has anti-spending weirdos, but they're rare (if overpowered, unfortunately).

The Distribution of Age Among KGSS Respondents, 2023

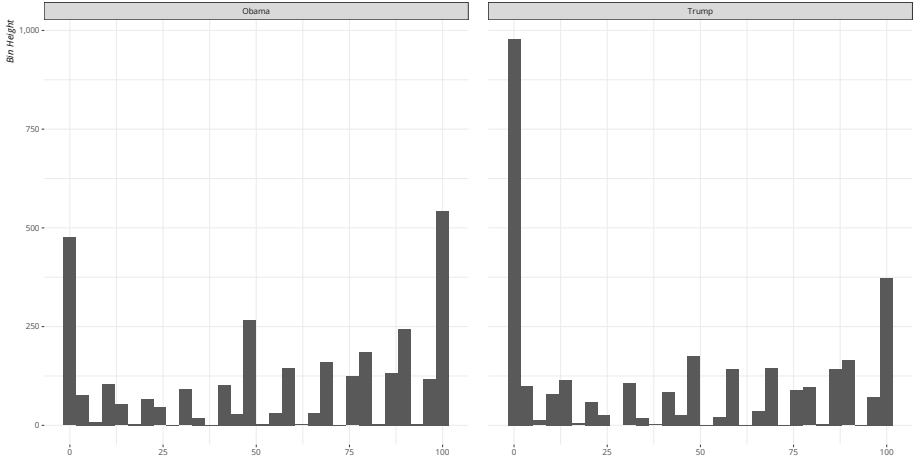
By and large, a distribution of age in a survey will follow a normal distribution (with obvious left truncation and a small right tail).



Data: Korean General Social Survey, 2023 (by way of ?kgss_sample in {simqi}). Histograms do have arbitrary bins that you should consider. Again, I'm deliberately cheesing this for presentation.

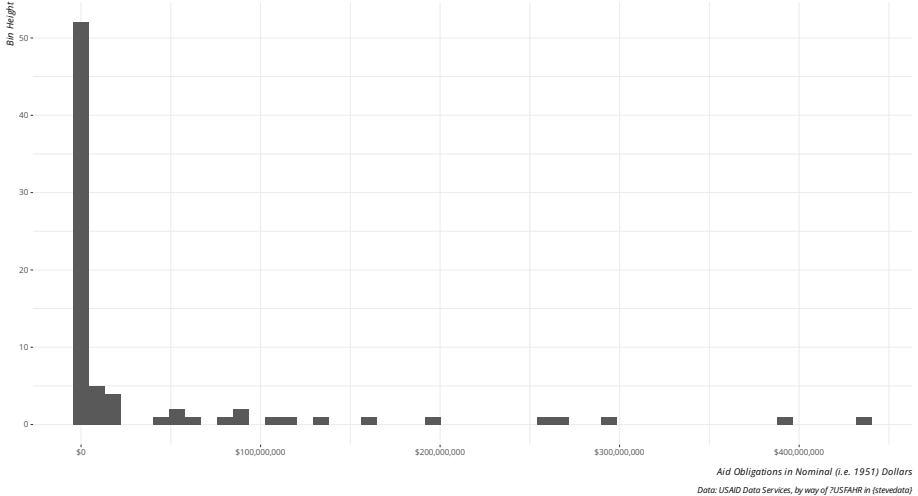
Thermometer Ratings Toward Donald Trump and Barack Obama (April 2020)

This would be 'interval' in the classic scale, but has left and right bounds (and a hideous distribution to boot).



The Distribution of U.S. Foreign Aid in 1951 (under Harry Truman)

You could argue this is 'ratio' in the classic sense, but that hides important peculiarities of this data-generating process.



Interval, Ratio, and Yadda Yadda Yadda

Both “interval” and “ratio” have granular (practically infinite) possible values.

- In the classic typology, they are distinguished by what 0 means in the measure.

Rather than split these hairs, I’d encourage you to think of these as “continuous”.

- i.e. the values are infinitely (or practically) granular.
- An arithmetic mean may not be faithful, but would make sense.
    - This even applies to some integers, like age and income.

Summarizing Variables

Type	Central Tendency	Dispersion
Unordered-Categorical	Mode	(Some you'll use at an advanced level)
Ordered-Categorical	Median	IQR, MAD (Better to eyeball it)
'Continuous'	Mean	Standard Deviation

Identifying Measures of Central Tendency

Table 5: County/Counties of Residence for Respondents in SOM (2019, 2020)

County	No.	County	No.
Blekinge	31	Stockholm	637
Dalarna	92	Södermanland	73
Gävleborg	75	Uppsala	88
Halland	95	Värmland	77
Jämtland	45	Västerbotten	88
Jönköping	91	Västernorrland	61
Kalmar/Gotlands	94	Västmanland	72
Kronoberg	69	Västra Götaland	493
Norrbotten	63	Örebro	83
Skåne	383	Östergötland	131

Note:

Data: SOM (2019 and 2020, by way of ?som_sample in {simqi}).

Identifying Measures of Central Tendency

Table 6: The Justifiability of Abortion in the U.S. in 2011

Values	No.	Cum. Sum	Cum. Sum (%)
Never Justifiable	497	497	22.93%
2	161	658	30.36%
3	129	787	36.32%
4	96	883	40.75%
5	505	1388	64.05%
6	154	1542	71.16%
7	146	1688	77.9%
8	176	1864	86.02%
9	81	1945	89.76%
Always Justifiable	222	2167	100%

Note:

Data: World Values Survey, 2011 (by way of ?wvs_usa_abortion in {stevedata}.

Mean

The arithmetic **mean** is used only for continuous variables.

- This is to what we refer when we say “average”.

Formally, *i* through *n*:

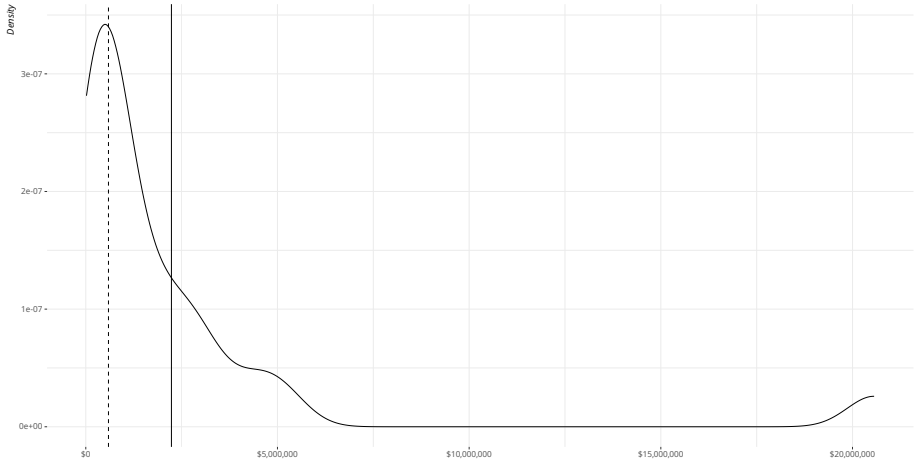
$$\frac{1}{n}\Sigma x_i \tag{1}$$

We can always describe continuous variables with the median.

- We cannot do the same for ordinal or nominal with the mean.
- For really granular data, there is likely no real proper “mode” to report.

Real GDP for 22 Select (OECD?) Countries

The median is the difference between Belgium and Switzerland. The mean wants to hang out with the U.S. (the largest economy on the planet).



Data: Penn World Table (10.0), by way of ?pwt_sample in {stevedata}. Density plots are smoothed histograms and give a better assessment of the overall shape of the data and is less sensitive to arbitrary bin selection.

A Comment on Dummy Variables

Dummy variables behave curiously in measures of central tendency.

- Mode: most frequently occurring value (as it is nominal).
- Median: also the mode.
- Mean: the proportion of 1s.

Dispersion

We also need to know variables by reference to its **dispersion**.

- i.e. "how average is 'average'?"
- How far do variables deviate from the typical value?
- If they do, measures of central tendency can be misleading.

In a lot of applications, you can just visualize this or look for a table.

- If you have continuous data, you can get a precise measure: the **standard deviation**.
  - i.e. the square root of the sum of squared deviations for each observation from the mean.
  - There is a standard deviation for dummy variables, but it's different: $\sqrt{p(1-p)}$
- For less precise data: just eye-ball it.
  - You could ask for an inter-quartile range or MAD, but, again, eye-ball it.

How to Calculate a Standard Deviation

Table 7: Calculating the Mean and Standard Deviation of GDP per Capita in the Ten ASEAN Countries (2019)

isocode	rgdppc	mean	dvtn	sum_dvtn	dvtn2	sum_dvtn2	variance	sd
BRN	73249.186	24322.66	48926.528	0	2393805147	7550598490	838955388	28964.73
IDN	11595.102	24322.66	-12727.556	0	161990680	7550598490	838955388	28964.73
KHM	4500.053	24322.66	-19822.605	0	392935676	7550598490	838955388	28964.73
LAO	7585.554	24322.66	-16737.104	0	280130653	7550598490	838955388	28964.73
MMR	5153.375	24322.66	-19169.283	0	367461401	7550598490	838955388	28964.73
MYS	25735.309	24322.66	1412.651	0	1995583	7550598490	838955388	28964.73
PHL	8448.533	24322.66	-15874.124	0	251987828	7550598490	838955388	28964.73
SGP	82336.342	24322.66	58013.684	0	3365587542	7550598490	838955388	28964.73
THA	17116.308	24322.66	-7206.350	0	51931482	7550598490	838955388	28964.73
VNM	7506.817	24322.66	-16815.841	0	282772498	7550598490	838955388	28964.73

Note:

Data: Penn World Table (10.01)

Alternatively:

```
sd(asean_rgdppc)
#> [1] 28964.73
```

Please Install RStudio

IRIII.2 LECTURES R/RSTUDIO LAB SCRIPTS PROBLEM SETS ATHENA

R/RStudio

The lab sessions themselves will take place in rooms with computers in them for the student to use, though there is a (reasonable, implicit) assumption that the student has a personal computer. Tablets are not advised for these purposes as it is difficult to install the required third-party software needed for this course, which would be a major issue no matter the third-party software the student used for statistical analysis. Tablets will also typically lack the kind of memory and processing power for computational uses like this.

RStudio

Lab sessions and problem sets (more in the next section) will all be done in the **R** programming language. Students should download this free software programming language at cran.r-project.org and install it on their personal computer. Binaries are available for Windows and Mac (even Linux, if that is the weapon of choice for the student). The instructor will be teaching around R version 4.1.2. It's fine if you have a more current version that you install. If you have an older version than this, you should really upgrade to a more current version (just in case).

The **R** scripts I provide are designed to work on the student's computer with minimal maintenance. This should be close in

Conclusion

Welcome to IRIII.2 (the dungeon mini-boss of our program).

- All social science research is qualitative; some of it is quantitative.
- Know your perspectives (i.e. you have them, do trust).
- There's always slippage (ideally random) between concept and measure.
- Know your variable types and what information they communicate.

Table of Contents