Bivariate 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 measures of association for lower levels of measurement.
2. Give some applied examples in IR and economics.

Measures of Association for Nominal Data

Measures of association for nominal data vary whether the data are 2x2 or not.

Table 1: A Simple 2x2 Contingency Table

	B = 0	**B = 1**
A = 0	a	b
A = 1	c	d

Measures of Association for 2x2 Nominal Comparisons

1. **Yule's Q**: $\frac{(ad-bc)}{(ad+bc)}$
2. **Odds Ratio**: $\frac{ad}{bc}$
3. **Phi**: $\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$

When the number of rows and columns are greater than two:

- Cramer's V
- Contingency coefficient

I'd belabor these more, but both lean on the χ^2 statistic.

The Chi-squared Test

The Chi-squared (χ^2) test is a staple in pedagogical instruction.

- Formally: a test for "independence" of observed and expected counts.
- Scales well no matter the number of rows and columns.

The test communicates whether observed counts for two or more groups are discernibly different than what could be expected by chance. Formally:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

...where *O* is observed counts and *E* is expected counts.

The Null Hypothesis

These types of tests have explicit “null” hypotheses.

- H_0: the two categorical variables are independent (i.e. not associated).
- H_1: the two categorical variables are associated with each other (i.e. not independent).

The tests you do under these conditions compare what you observe with what would be expected if the null were true.

- Greater incompatibility of the test stat with some distribution -> reject the null as unlikely to be true.

By Way of a Blog Post...

Learning About Expected Categorical Relations (Chi-Squared Tests) in R by Way of Arms Races and War

Posted on August 11, 2025 by steve in Political Science Teaching R

An Eritrean soldier stands in front of a destroyed T-55A tank in 1999. This war starts in 1998 but the mutual military build-up for it arguably started in 1996. (Broń Pancerna/Flickr)

This is a post I'm writing just to spam material to my blog, and also to pad material I need to prepare for my IRIII students in their quantitative methods sequence. It's a challenge to teach them stuff that is super basic, but has a real-world application, and in the limited time I have with them. The class had historically been built toward just getting them to do `lm()` in R (or `regress` in Stata) and to be happy if they can do that. However, there's been some shuffling amid hour cuts that incidentally gives me more time to teach them more stuff. But, again, it has to be simple.

Enter the arms race and war debate. I cut my teeth on this debate in graduate school and still like to teach around the basics of this stuff when I can. I still think Lewis Fry Richardson's linear theory of nations gets at the core of how we should conceptualize the arms race, acknowledging it's a glorified port of his training in mathematics to the realm of international relations.[1] Admittedly, it is a bit of a dated topic. Perhaps it's fair to note that the "arms race" of a century or two ago looks nothing like what an arms race would resemble now. It won't be simple expenditures. It won't be manpower. It won't even be boats. But the substance of this debate maps nicely to two thing I want to accomplish in an IR curriculum at the bachelor's level. First, it highlights how a lot of realpolitik conventional wisdom stretches so thin it strains to cover anything in detail. Second, the empirical application of this debate is all chi-squared tests. There are definitely more advanced ways of approaching this, especially with what this means in the 21st century. But, you can learn about the chi-squared test with these things I'd have you read anyway if I could.

Table IX in Sample (1997)

Table IX. Diehl's Index and All Major State Militarized Disputes

	Arms Race	No Arms Race
Escalation	14	17
No Escalation	39	187

$Q = 0.60$	$\chi^2 = 13.0$	$n = 257$
$\text{phi} = 0.23$		$p < 0.001$

Table 2: Table IX in Sample (1997)

	Arms Race	*No Arms Race*
Escalation to War	14	17
No Escalation	39	187

Btw...

```
(14*187 - 39*17)/(14*187 + 39*17) # Yule's Q
#> [1] 0.5958549
(14*187 - 39*17)/(sqrt((14 + 17)*(39 +187)*(14 + 39)*(17+187))) # Phi
#> [1] 0.2246253
```

Table 3: Table IX in Sample (1997), with Row and Column Totals

	Arms Race	*No Arms Race*	**Row Total**
Escalation to War	14	17	**31**
No Escalation	39	187	**226**
Column Total	**53**	**204**	**257**

Table 4: Expected Counts in Table IX in Sample (1997)

	Arms Race	*No Arms Race*
Escalation to War	(31*53)/257 = **6.393**	(31*204)/257 = **24.607**
No Escalation	(226*53)/257 = **46.607**	(226*204)*257 = **179.393**

Table 5: Chi-Squares in Table IX in Sample (1997)

	Arms Race	*No Arms Race*
Escalation to War	(14 - 6.393)^2/(6.393) = **9.051**	(17 - 24.607)^2/(24.607) = **2.351**
No Escalation	(39 - 46.607)^2/(46.607) = **1.241**	(187 - 179.393)^2/(179.393) = **.322**

Sum those up and you get your χ^2 (i.e. 12.967).

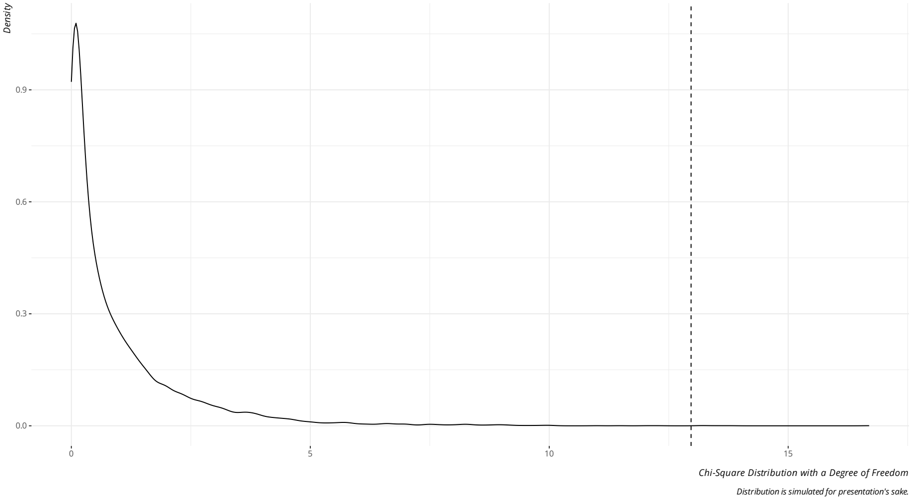
Inference in a Chi-square Test

Inference about the test statistic is compared to its eponymous distribution.

- This is a distribution of squared standard normal variables with just one parameter (k).
- k: the number of squared standard normal variables to summarize.

Comparing Sample's (1997) Chi-square Stat with What Could Be Expected Under No Association

Our test statistic is a near impossibility if there were truly no differences between groups, per the chi-squared test and distribution.



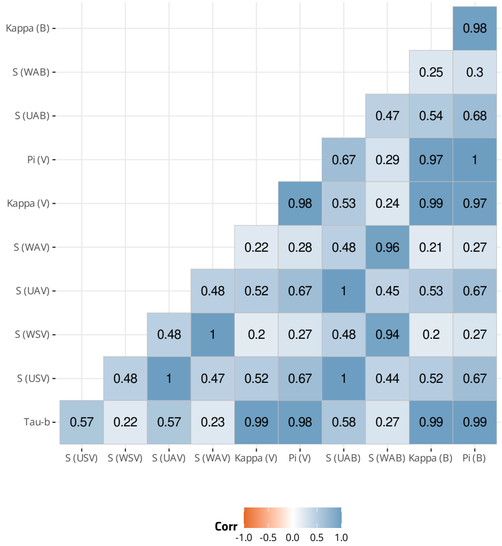
Measures of Association for Ordinal Comparisons

You might still see some of these in the wild, especially for assessing "foreign policy similarity".

- **Spearman's rho**: rank-based analog to Pearson's *r* correlation.
- **Kendall's Tau(-b)**: assesses probability of concordance/discordance.
- **Signorino and Ritter's (1999) S**: measures a kind of "distance" or similarity between pairs.
- **Cohen's (1968) kappa**: measures a "reliability" of coders
  - Older versions of the measure, like **Scott's (1955) pi** work with nominal data too.

A Correlation Matrix of Various Measures of Foreign Policy Similarity in 1816

Notice these measures are not substitutable. More on Pearson's r later.



Data: Häge (2011). Data are for CoW alliances.
U or W determines whether the S stat is weighted by capabilities.
A or S determines whether the S stat is absolute or squared distances.
B or V, for multiple measures, communicates whether the alliances are treated as ordinal or binary.

Measures of Association for Continuous Data

Pearson's correlation coefficient (or **Pearson's *r***) will tell us how strongly two things travel together.

$$\frac{\sum\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right)}{n-1}$$

...where:

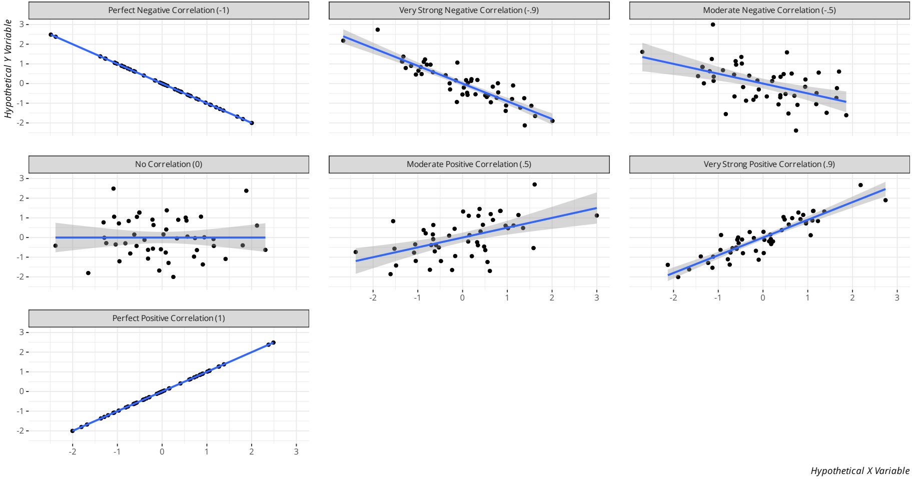
- x_i, y_i = individual observations of *x* or *y*, respectively.
- $\overline{x}$, $\overline{y}$ = means of *x* and *y*, respectively.
- s_x, s_y = standard deviations of *x* and *y*, respectively.
- *n* = number of observations in the sample.

Properties of Pearsons r

1. Pearson's *r* is symmetrical.
2. Pearson's *r* is bound between -1 and 1.
3. Pearson's *r* is standardized.

Various Linear Patterns You Could Deduce from a Scatterplot

Do note: you can describe these correlations however you want. There is no formal metric, beyond direction, perfection, and zero.



Data: Simulated with smvrnorm() in {stevemisc} package.

Comparing Two Means of Two Groups

The (Welch-Satterthwaite) *t*-test (for unequal variances) is ubiquitous as well.

- The experimentalists love to use it when they can (so do clinical researchers).

Assumptions:

- Sample means being compared come from normally distributed population.
- Independence between/within groups being compared.
- Measures are continuous (i.e. you have means).

There are variants of the *t*-test, but we'll generally assume you mean this particular one.

The Formula for Welch-Satterthwaite's t-test

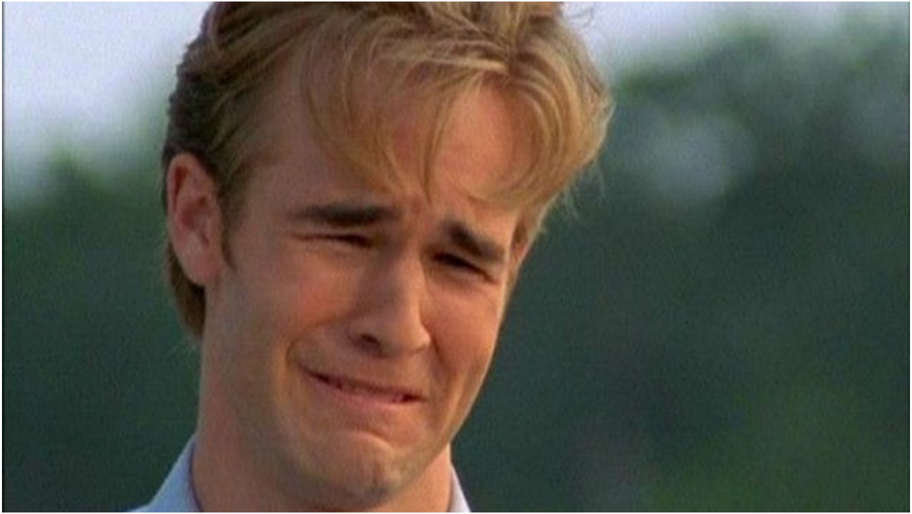
$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2}},$$

where...

$$s_{\bar{X}_i} = \frac{s_i}{\sqrt{N_i}}$$

and...

$$\nu \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2 \nu_1} + \frac{s_2^4}{N_2^2 \nu_2}},$$



IRII.3 Students May Remember This...

	MEAN	STD. DEV.	VARIANCE	N	STD. ERR.
MEN	9023.25	6570.4857	43171282	28	1241.7051
WOMEN	4129.0517	4284.1459	18353906	29	795.54593
DIFFERENCE	-4894.1983				
T-STAT	-3.3187872				
DF (NUM)	4.729E+12				
DF (DENOM)	1.024E+11				
DF	46.207667				
Pr(observing this, if true difference was 0)					
	0.0017695				

By Way of Another Blog Post...

Permutations and Inference with an Application to the Gender Pay Gap in the General Social Survey

Posted on October 24, 2020 by steve in R Political Science

Last updated: 26 October 2024. Some things got lost in transition behind the scenes. These include a move from R version 3.5 to 4, a transition from working directories, and some function changes in `{stevemisc}`.

I'm teaching my quantitative methods class this semester for the first time in three and a half years, an exigency brought on by both a hiring/spending freeze and a faculty departure. The impromptu nature of teaching this class along with the COVID-19 fog that has consumed us all mean I'm basically teaching the same class, with just a few updates, that I last taught three and a half years ago. My quantitative skills have improved greatly since then, as have my computational skills. It's led me to think of ways I can improve what I teach should I have to teach it again soon.

Most analyses show real wage disparities between men and women, even when matching/controlling for important confounders

I stumbled across this tweet from Grant McDermott. It's from last year and I missed it entirely when he first posted it, but it appeared in my timeline again amid some other ongoing conversation. The link is to a keynote speech from John Rauser. Rauser's main point in the keynote, echoed by McDermott, is that computational means to inference are more accessible and better illustrate the underlying point we try to teach students than doing something like calculating standard errors of sample means, calculating *t*-values or *z*-values, and finding the approximate area underneath a Student's *t*-distribution or normal distribution that corresponds with that score. McDermott's comment is that it's only for outdated pedagogy that we don't teach what computational power has made more practical and accessible. I'll stop short of saying that here. After all, bootstrapping was ultimately Bradley Efron's (1979) answer to his own question of what the jackknife was trying to approximate (i.e. a random sampling distribution better done via bootstrap). There is—at least I'm thinking right now—more value in understanding what things like bootstrapping and permutations are trying to approximate before showing how the

The Gender Pay Gap (in the U.S.)

We'll explore the gender pay gap in the U.S. with a simple data set from 2018.

- These are Americans in 2018, between 18-25, who have never been married, have no kids, are not in school (but finished high school).
- Respondent's base income (`realrinc`) is in 1986 USD.

Those of you who remember me from IRII.3 will have seen this before.

- I've uploaded relevant materials to Athena for those that haven't.
- I'd love to get something similar from SOM for Swedes, but I keep getting told no. :(

The Null Hypothesis

Again, this is another “reject the null” type of test.

- H_0: the two means are equal to each other
- H_1: the two means are not equal to each other (or one is greater).

Like the χ^2 test, you compare the test statistic to what could be expected if H_0 is true.

- Greater incompatibility of the test stat with some distribution -> reject the null.
- Assert H_1 is closer to what's true.

```r
t.test(realrinc ~ gender, GSSW)
#>
#>  Welch Two Sample t-test
#>
#> data:  realrinc by gender
#> t = -3.3188, df = 46.208, p-value = 0.00177
#> alternative hypothesis: true difference in means between group Female and group Male is not equal
#> 95 percent confidence interval:
#>  -7862.245 -1926.152
#> sample estimates:
#> mean in group Female   mean in group Male
#>             4129.052             9023.250
```

Inference in a t-test

Like the chi-squared test, the *t*-stat is compared to a hypothetical distribution with some degrees of freedom.

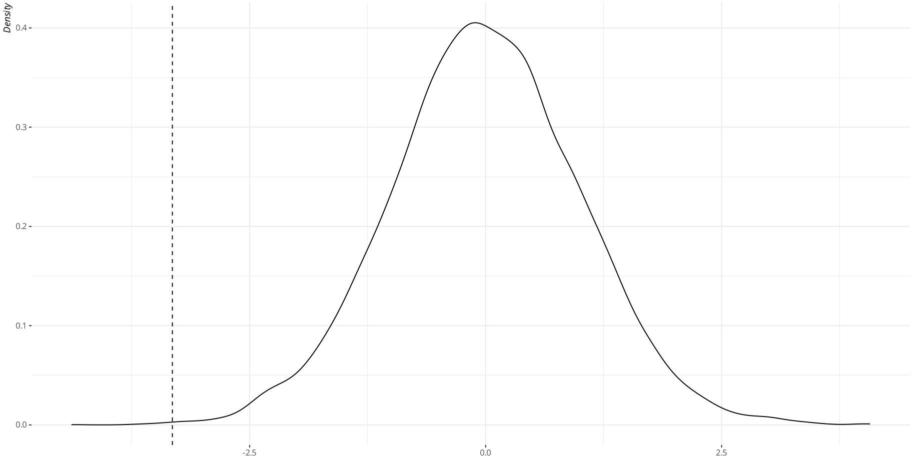
- This is about 46.208, in our case.
- It also has an important mean parameter (0 [for no differences] in this case).

Pertinent features of the *t*-distribution:

- Fewer degrees of freedom -> longer tails.
- More degrees of freedom - > shorter, “normal” tails.
- Rule of thumb: the *t*-distribution approaches the familiar bell curve at 30 degrees of freedom.
- Like the normal distribution, it too is symmetrical around its mean.

Comparing our *t*-stat with What Could Be Expected if There Were No Differences Between Men and Women

Our test statistic is a near impossibility if there were truly no differences between men and women in their wages, per the *t*-distribution.



Student's t Distribution with 46.208 Degrees of Freedom

Distribution is simulated for presentation's sake.

Conclusion

There's more in here that you'll need, but some things to consider:

- Most of your measures of association/correlation are "symmetric".
- Many bivariate measures of association have boutique uses, but you might see some of them in the wild.
  - i.e. dyadic foreign policy similarity has long been measured with them (for better or worse).
- Know the chi-squared distribution; it'll recur in several quadratic-form tests.
  - Prominently: Breusch-Pagan, Lagrange multipliers, Box-Pierce, Hausman, and more.
- Know the t-distribution; it's the one you'll be seeing the most at the intro-level.
  - It's normal-like, but has that important degrees of freedom parameter.
- Notice the inference you're making.
  - i.e. "what is plausible, given some distribution? Is my test statistic consistent with it?"

Table of Contents