PLSC 502: "Statistical Methods for Political Research"

Measures of Association
November 6, 2023

Introduction

Today we'll discuss inference for nominal-level variates. After talking about contingency tables, we'll introduce the chi-square statistic as a general way of conducting inference on nominal-level variates. We'll briefly describe inference for one-way tables, and spend the rest of the time going over chi-square tests for the statistical independence of two nominal-level variates. At the end, we'll very briefly mention a few alternatives to the chi-square approach.

One-Way and Two-Way Crosstabs

Crosstabs (more correctly referred to as *crosstables*, or *contingency tables*) are tabular representations of data. We discussed frequency tables and crosstabs briefly when we reviewed summary statistics, but a review is probably in order.

One-Way Frequency Tables

A *one-way* (or *frequency*) table is simply a table listing the categories of Y and the number of observations in each of those categories. Frequency tables also often contain category (cell) proportions or percentages; if n_y is the observed frequency of observations in Y's category $Y = y$, then the category proportion is just

$$P_y = \frac{n_y}{N}.$$

Thus, for example, in our Africa data, we have:

Category	Frequency	Proportion
No Civil War	30	0.70
Civil War	13	0.30
Total	43	1.00

Two-Way Crosstables

Crosstabs are cross-classified frequency tables. We typically define the main variable of interest Y and place it on the "Y" (vertical) side of the table, while the covariate ("independent," or "X") variable's categories are listed on the horizontal axis of the table. Each cell is then defined as n_{yx}, the number of observations in the data for which $Y = y$ *and* $X = x$.

In addition to the cell frequencies, we can calculate a number of other proportions in two-way tables:

- *Row proportions* (or percentages) are the proportion of observations in that row of the table (that is, with $Y = y$) falling into the column defined by $X = x$. They sum to 1.0 across columns.
- *Column proportions* (or percentages) are the proportion of observations in that column of the table (that is, with $X = x$) falling into the row defined by $Y = y$. They sum to 1.0 down rows.
- *Cell proportions* (or percentages) are the proportion of the total number of observations in that cell of the table. They sum to 1.0 overall columns and rows (cells).

Of these, we are usually most interested in column proportions, since (as we'll discuss) they allow us the most intuitive examination of whether Y and X are *independent*. More on this below.

A typical two-way table might look like this:

	Sub-Saharan?		
Civil War?	**No**	**Yes**	**Total**
No	5	25	30
(Row)	(0.17)	(0.83)	(1.00)
[Column]	[0.83]	[0.68]	[0.70]
{Cell}	{0.12}	{0.58}	{0.70}
Yes	1	12	13
(Row)	(0.08)	(0.92)	(1.00)
[Column]	[0.17]	[0.32]	[0.30]
{Cell}	{0.02}	{0.28}	{0.30}
Total	6	37	43
	(0.14)	(0.86)	(1.00)
	[1.00]	[1.00]	[1.00]
	{0.14}	{0.86}	{1.00}

Note that we'd *never* actually put all three types of proportions in a table – typically we focus on the column proportions. Moreover, many software packages (e.g., Stata) express the cell proportions in terms of percentages rather than proportions.

Statistical Independence

If we believe there is a relationship between two (here, nominal) variables (say, Y and X), the direct implication is that the distribution of Y is different for different values of X. Formally, then, we're interested in $f(Y|X)$, the distribution (say, density, or whatever) of outcomes on Y once we "condition" on values of X. For example, in the table above, the conditional distribution of civil wars given that a country is in sub-Saharan Africa is 12 countries with wars and 25 without them (for a total of 37). This is in contrast to the unconditional distribution of Y, $f(Y)$, which is the

distribution of Y for *all* values of X (above, that's 13 wars and 30 non-wars).

Considered in this way, there's an obvious "null hypothesis": that the distribution of Y is the same across all values of X. We write this:

$$H_0 : f(Y|X) = f(Y). \tag{1}$$

In other words, if two variables are unrelated, then the conditional distribution of each on the other is the same as its unconditional ("marginal") distribution. If H_0 is true, we say X and Y are statistically *independent*.

In practice, and particularly when examining nominal-level variables, there will almost inevitably be *some* departure from strict independence; that is, the conditional distributions of Y will almost never be exactly equal to its marginal distribution. The relevant question is whether those differences are sufficiently small that we can/should attribute them to sampling error, or whether they are reflective of a "real" difference in the population. Knowing something about the sampling error of our variables allows us to make this assessment.

The Chi-Square Statistic

Consider the cells of a one- or two-way frequency table containing a total of N observations on two nominal-level variables Y and X. Define k_Y and k_X as the number of different categories of Y and X, respectively. As above, let n_{yx} be the observed frequency of observations in the cell corresponding to category y of ("dependent") variable Y and, if it is present, category x of ("independent") variable X. The "marginals" of Y and X are defined as

$$R_y = \sum_{k_X} n_{yx}$$

and

$$C_x = \sum_{k_Y} n_{yx},$$

respectively. That is, the total number of observations in category y is the sum of all observations with $Y = y$ across all the categories (columns) of X, and the total number of observations in category x is the (row) sum of all observations with $X = x$.

For a particular cell defined by y (and x), the *expected* number of observations in that cell under the assumption that Y and X are independent is thus equal to:

$$E_{yx} = \frac{R_y \times C_x}{N}. \tag{2}$$

In the case of a one-way table, this reduces to simply $N \times \pi$, where π is $1/k_Y$, the proportion defined as the reciprocal of the number of categories.

Under the "null" hypothesis of the independence of Y and X, we would expect two things:

1. On average $n_{yx} = E_{yx}$. That is, we should expect our cell count to be equal to the expected number of observations in that cell, a defined by the marginals for each variable.

2. The difference between n_{yx} and E_{yx} should be small.

These two points lead directly to the *chi-square test* for the independence of two nominal variates. The test statistic is defined as:

$$\chi^2 = \sum_{k_Y k_x} \frac{(n_{yx} - E_{yx})^2}{E_{yx}}. \tag{3}$$

Under the null hypothesis, this test statistic has a sampling distribution that is chi-squared with degrees of freedom equal to $(k_Y - 1)(k_X - 1)$.[1] The reason this is chi-square is straightforward. By the Law of Large Numbers, we would expect that, under the null hypothesis of independence,

$$n_{yx} - E_{yx} \sim \mathcal{N}(0, \sigma^2_E) \tag{4}$$

where σ^2_E is the sampling variability of the difference between n_{yx} and E_{yx} and is roughly proportional to $\sqrt{E}$. Thus, each squared difference, divided by the expected cell frequency, is χ^2_1, and the sum of $(k_Y - 1) \times (k_X - 1)$ independent χ^2_1 variates is $\chi^2_{(k_Y-1)\times(k_X-1)}$.

If one were to do a chi-square test "by hand," then, it would proceed in four steps:

1. Calculate E_{yx} for each cell,

2. Calculate $\frac{(n_{yx}-E_{yx})^2}{E_{yx}}$ for each cell,

3. Sum these values across all cells, and

4. Compare the resulting statistic to a chi-square distribution with $(k_Y - 1) \times (k_X - 1)$ degrees of freedom

Of course, we have computers for such things these days...

Large values of χ^2 are evidence against the null hypothesis. In addition, the (normed) differences between observed and expected cell counts are often referred to as *Pearson residuals*; these will turn out to be useful later on.

[1] The reason for this number of degrees of freedom is straightforward: it represents the amount of "free" variation in the data once the marginals are accounted for.

A Few Pointers...

1. Note that while the chi-square test as it is implemented in nearly every default I know tests for the independence of two variables (or, alternatively, for the equiprobability of each of the k_Y possible outcomes), it's possible to plug in any values of E_{yx} you care to, so long as they sum to N. This means that one can test for things that are not (necessarily) based upon the data marginals; one common hypothesis, for example, is that all cross-categories are equally likely (that is, that $E_{yx} = \frac{N}{k_Y k_X \, \forall \, x,y}$).

2. A good rule of thumb is that, if your chi-squared statistic is equal to or less than the number of degrees of freedom, you will fail to reject the null at any really viable level of significance.

3. In instances where there are relatively "sparse" data (that is, where there are more than one or two instances where $E_{yx} < 5$), the chi-square test is not recommended; see below.

Other Alternatives: Fisher's "Exact" Test

The chi-square test requires the Law of Large Numbers to be operative in order to work properly. In situations where one or more values of E_{yx} are less than five (and particularly if $E_{yx} < 1$), the chi-square distribution will be a very poor fit to the actual distribution of the test statistic; in those cases, we're not close enough to "asymptopia" . In such circumstances, it is unwise to use a chi-square test, and instead we typically rely on an alternative test based on combinatorics, known as *Fisher's Exact Test*.

The formula for Fisher's test in the case of a $k_y \times k_X$ contingency table is:

$$P = \frac{(R_1!R_2!...R_{k_Y}!)(C_1!C_2!...C_{k_X}!)}{N! \prod_{k_Y,k_X} n_{yx}!}. \tag{5}$$

Here, R, C, and n denote the row, column, and cell frequencies, respectively, and N is once again the total number of observations. The logic behind (5) (which is in fact a multivariate generalization of the probability function for a hypergeometric distribution) is that the denominator represents the possible ways in which one could arrange the data on N observations in a $k_y \times k_X$ contingency table, while the numerator reflects the possible orderings with the marginals determined by the values of R and C (the marginals).

Fisher developed his test for 2×2 tables; it is computationally challenging for larger tables. Moreover, for tables where cell counts are all sufficiently large (that is, with $E_{yx} > 5$ or so for all cells), it is asymptotically the same as a (much easier to calculate) chi-square test. However, it is superior in instances where we have relatively small cell frequencies.

An Example: Feminism as an Insult

In the aforementioned September 1997 CBS/NYT Poll (the one with the questions on the designated hitter, with $N \approx 1000$), the pollsters asked a somewhat different question as well:

Do you consider calling someone a feminist to be a compliment, an insult, or a neutral description?

The first was coded "1," the second "2," and the third "3." The slides illustrate three different examples of using frequency and contingency tables, and chi-square tests, on these data.

Ordinal Variates: Concordance and Discordance

We'll discuss measures of association for ordinal variables today. For notational purposes, we'll refer to a generic crosstable of two variables Y and X, each of which is assumed to consist of three ordinal categories (coded 1,2, and 3). The respective cell frequencies and the row and column marginals are defined as:

		X			
		1	2	3	
	1	n_{11}	n_{12}	n_{13}	n_{1X}
Y	2	n_{21}	n_{22}	n_{23}	n_{2X}
	3	n_{31}	n_{32}	n_{33}	n_{3X}
		n_{Y1}	n_{Y2}	n_{Y3}	N

The central challenge of measuring association between ordinal variates is how to retain the information present in the ordering of the categories, without giving the numerical values assigned to them cardinal content. We do this through the idea of concordant and discordant pairs of observations in the data. Two observations $i = \{1, 2\}$ in a dataset are said to be *concordant* if:

$$\text{sign}(X_2 - X_1) = \text{sign}(Y_2 - Y_1). \tag{6}$$

Similarly, a *discordant* pair exists where:

$$\text{sign}(X_2 - X_1) = -\text{sign}(Y_2 - Y_1). \tag{7}$$

Thus, for an observation in cell (1,1) of the table above, all observations in cells (2,2), (2,3), (3,2), and (3,3) are concordant with it, because in every instance such observations have both a higher value on Y *and* a higher value of X.

The total number of concordant pairs in the 3×3 table above, then, is equal to

$$N_c = n_{11}(n_{22} + n_{23} + n_{32} + n_{33}) + n_{12}(n_{23} + n_{33}) + n_{21}(n_{32} + n_{33}) + n_{22}(n_{33}). \tag{8}$$

Similarly, the number of discordant pairs is

$$N_d = n_{13}(n_{21} + n_{22} + n_{31} + n_{32}) + n_{12}(n_{21} + n_{31}) + n_{23}(n_{31} + n_{32}) + n_{22}(n_{31}). \tag{9}$$

These numbers will be different in larger tables, but the principle by which they are calculated remains the same. N_c and N_d for the basis for our statistics for association among ordinal variables.

Gamma

Gamma (γ) is the normed difference between the number of concordant and discordant pairs in the data:

$$\gamma = \frac{N_c - N_d}{N_c + N_d} \tag{10}$$

It can also be thought of as the difference between the *proportions* of pairs that are concordant versus discordant:

$$\gamma = \frac{N_c}{N_c + N_d} - \frac{N_d}{N_c + N_d} \tag{11}$$

Note that γ does not count "ties" – no data on pairs are used when (say) $Y_2 > Y_1$ and $X_2 = X_1$, nor when $X_2 = X_1$ and $Y_2 = Y_1$.

- $\gamma \in [-1, 1]$.
- $\gamma = 0 \;\leftrightarrow$ no association between X and Y, though it can also happen whenever $N_c = N_d$. That is, $\gamma = 0$ is necessary but not sufficient for statistical independence.
- Higher absolute values of γ correspond to stronger associations between X and Y.
- $\gamma = \pm 1.0$ under conditions of (at least) *weak monotonicity* (γ will equal 1.0 whenever, as X increases, Y either increases or stays the same; it will equal -1.0 whenever, as X increases, Y decreases or stays the same.

Inference on γ

The sampling distribution of $\hat{\gamma}$ is Normal, which means that we can use the "usual" approaches to inference, including creation of $(1-\alpha)$-percent confidence intervals using the normal distribution. We can also test specific hypotheses about the population value of γ by converting our estimate to a z-score:

$$z = (\hat{\gamma} - \gamma)\sqrt{\frac{N_c + N_d}{N(1 - \hat{\gamma}^2)}}. \tag{12}$$

This z-score has a sampling distribution that is $\mathcal{N}(0, 1)$, making inference straightforward.

Kendall's τ

Kendall's τ is similar to γ:

$$\tau = \frac{N_c - N_d}{\frac{1}{2}N(N-1)} \tag{13}$$

You can think of τ as an alternative to γ, one that "norms" the difference between N_c and N_d by a somewhat different number (in this case, the number of all *possible* pairs in the data). As such, the numerator of (13) still signs the statistic, while the denominator scales it.

τ_a, τ_b, and τ_c

τ in (13) is usually called τ_a. Two other variants, τ_b and τ_c, exist in order to "correct" for tied values in the data.

τ_b is generally used for "square" tables; it norms the difference between concordant and discordant pairs to reflect "ties" in the data:

$$\tau_b = \frac{N_c - N_d}{\sqrt{[(N_c + N_d + N_{Y^*})(N_c + N_d + N_{X^*})]}} \tag{14}$$

where N_{Y^*} and N_{X^*} are the number of pairs *not tied* on Y and X, respectively. τ_b is widely used, and is the default in many statistical packages. It's characteristics are:

- $\tau_b \in [-1, 1]$.
- $|\tau_b| = 1.0$ under *strict monotonicity* – that is, if (a) Y increases as X increases (for $\gamma = 1.0$) and (b) there is only one value of Y corresponding to each value of X. Put differently, this means that there are no "ties."
- $\tau_b = 0$ corresponds to no association between X and Y.

τ_c is used for larger, "rectangular" tables.

$$\tau_c = (N_c - N_d) \times \left\{ \frac{2m}{[N^2 2(m-1)]} \right\} \tag{15}$$

where m is the number of rows or columns, whichever is smaller. τ_c is the correct statistic to use when one's table is "rectangular," particularly if one variable has a significantly larger number of categories than the other (e.g., for $2 \times k$ tables where $k \geq 3$).

Because of the way that they norm the differences between N_c and N_d, it will always be the case that

$$\gamma \geq \tau \tag{16}$$

for any of the versions of τ.

The slides contain some examples of the use of γ and τ on some applied data, taken from the September 2008 Big Ten "battleground" poll in Pennsylvania.

Relationships between Interval/Ratio-Level Variates

We'll spend the rest of the day discussing relationships between interval- and/or ratio-level variates.

Linear Relationships

The simplest form of a monotonic relationship between Y and X is a *linear* relationship. We can think of the relationship as one akin to

$$Y = mX + b \tag{17}$$

where (just like in geometry class) m is the "slope" of the line and b is the "intercept." The reason we characterize this as the "simplest" form of relationship is because, for a linear relationship,

$$\frac{\partial Y}{\partial X} = m;$$

that is, the change in Y associated with a one-unit change in X (that is, the slope of the function) is just m, a constant. This is true irrespective of the "location" at which the change in X takes place.

Nonlinearity

Of course, linearity is only one form of a relationship that interval/ratio-level variates might have. Two others are in Figures 1 (logarithmic) and 2 (exponential) in the slides. Note that:

- In a *logarithmic* relationship, we observe *diminishing* returns to Y in X. That is, the change in Y associated with a one-unit change in X is decreasing in X. Formally, this implies that, irrespective of $\frac{\partial Y}{\partial X}$,

$$\frac{\partial^2 Y}{\partial X \partial X} < 0.$$

- In an *exponential* relationship, we observe *increasing* returns to Y in X. That is, the change in Y associated with a one-unit change in X is increasing in X. Formally, this implies that

$$\frac{\partial^2 Y}{\partial X \partial X} > 0.$$

One can also imagine:

- Curvilinear relationships with more "bends" (a la polynomials),
- "Step-functions,"
- "Threshold" effects, and/or
- combinations of these.

All of which counsels that it's always good to "look" at your data.

Pearson's r

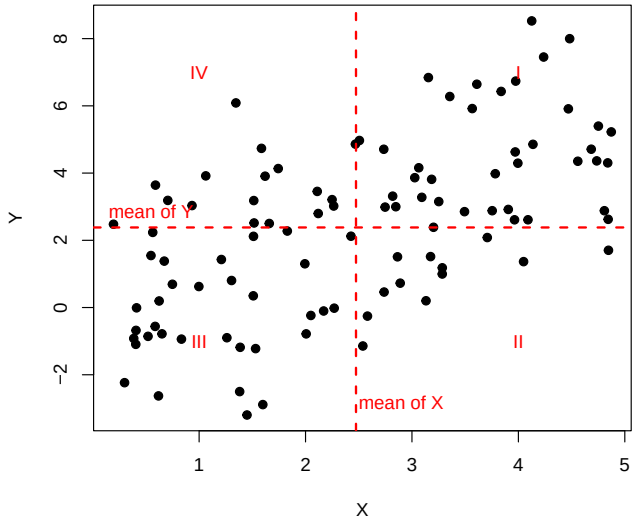
Pearson's product-moment correlation (much better known as *Pearson's* r) is the workhorse of bivariate association measures between two continuous variates. It is a summary measure of the direction and strength of the linear association between two variables. Formally,

$$r = \frac{\sum_{i=1}^{N} \left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)}{N-1} \tag{18}$$

where (as before) s_X and s_Y are the sample standard deviations of X and Y, respectively.

The intuition of Pearson's r is illustrated in Figure 1. The red dashed lines indicate the means of Y and X. Observations in quadrant I have both $X_i - \bar{X}$ and $Y_i - \bar{Y} > 0$; observations in quadrant II have $X_i - \bar{X} > 0$ and $Y_i - \bar{Y} < 0$, and so forth. The product of these two terms "signs" each observation's product of deviations-from-means; summing across observations means that relatively large numbers of observations in quadrants I and III will yield positive values of r, while larger numbers in quadrants II and IV will yield negative values.

Figure 1: Intuition: Pearson's r



Characteristics of r are:

- $r \in [-1, 1]$
- $r = 0 \leftrightarrow$ no association between Y and X.
- The sign of r indicates direction of the (*linear*) relationship; while

- The magnitude of $|r|$ indicates the strength of the (again, *linear*) relationship.
- These are illustrated graphically in the slides.

Note that the fact that r measures linear association means that:

- It cannot tell you anything about nonlinear relationships in the data (as in the figure in the slides).
- It can also be unduly influenced by *outliers* (which one might think of as a form of nonlinearity, depending on the circumstances) (Figure 5).
- Finally, note that the "slope" of the linear relationship has little or no bearing on r; two "perfect" positive, linear relationships between Y_1 and X_1 and Y_2 and X_2, each with different values of m in (17), will nonetheless both have $r = 1.0$. In other words, Pearson's r measures the degree of clustering around a line characterizing the relationship between Y and X, but not the *slope* of that line.

Sampling Distribution of r

The sampling distribution is a bit complicated, since r is necessarily bounded between -1 and 1. In particular, if the population value of r is very high or very low (that is, if $|r| \approx 1.0$), the sampling distribution is *skewed*.

Fisher (yes, *that* Fisher) showed that, even when the sampling distribution of the estimator $\hat{r}$ is skewed,

$$\hat{w} = \frac{1}{2} \ln \left(\frac{1 + \hat{r}}{1 - \hat{r}} \right) \tag{19}$$

is approximately $\mathcal{N}$ormally distributed with a mean of $\frac{1}{2} \ln \left(\frac{1+\hat{r}}{1-\hat{r}} \right)$ and a standard error of $\frac{1}{\sqrt{N-3}}$. This transformation is illustrated in Figure 2; you can see that it looks like a sideways "S-curve," with vertical asymptotes at -1.0 and 1.0. Thus,

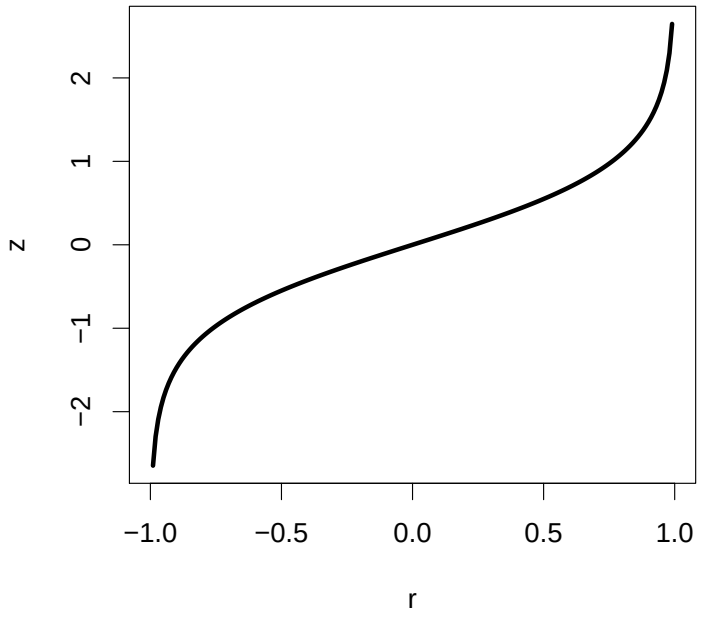
$$z_r = \frac{\frac{1}{2} \ln \left(\frac{1+\hat{r}}{1-\hat{r}} \right) - \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)}{\sqrt{\frac{1}{N-3}}} \sim \mathcal{N}(0, 1)$$

We can also use w to construct confidence intervals around $\hat{r}$, in the usual fashion. An alternative (and more exact) form of the sampling distribution of r – and one that is the default in the `cor.test` routine in R and `pwcorr` in Stata – is:

$$\frac{\hat{r}\sqrt{N-2}}{\sqrt{1-\hat{r}^2}} \sim t_{N-2}. \tag{20}$$

Most software packages will calculate p-values for the usual null hypothesis $r = 0$ automatically.

Figure 2: Fisher's z Transformation for Pearson's r



An Alternative to r: Spearman's ρ

An alternative to Pearson's r is the rank-based test known as *Spearman's ρ*.

Imagine sorting the data on both Y and X, and on the basis of their position on each variable, assigning them a *rank* on each, denoted R_{Y_i} and R_{X_i}, respectively. Thus, the observation in the data with the highest value of Y would be $R_{Y_i} = 1$, the next-highest would be $R_{Y_i} = 2$, and so forth, with a similar procedure for R_{X_i}. Spearman's ρ is then equal to:

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} D_i^2}{N(N^2 - 1)} \tag{21}$$

where D_i is the difference in ranks of observation i between Y and X (that is, $R_{Y_i} - R_{X_i}$).

Characteristics of Spearman's ρ:

- $\rho \in [-1, 1]$
- It has the same interpretation as r.
- ρ is also appropriate for use with ordinal data, since they can also be used to rank observations. However,
- When many "ties" occur, a better alternative is to calculate Pearson's r on the ranks R_{Y_i} and R_{X_i}, and assign "partial" (or "half") ranks to tied individuals.

Summary: Some Measures of Association

		X			
		Nominal	Binary	Ordinal	Interval/Ratio
Y	Nominal	χ^2	χ^2	χ^2	t-test (and η)
	Binary	χ^2	ϕ, Q	γ, τ_c	t-test
	Ordinal	χ^2	γ, τ_c	γ, τ_a, τ_b	Spearman's ρ
	Interval / Ratio	t-test (and η)	t-test	Spearman's ρ	r

Example: Feeling Thermometers

See the slides for some simple examples of how to estimate r (and rho) using R ...