Bivariate Relationships

David A. Hughes, Ph.D.

Associate Professor
Auburn University at Montgomery
david.hughes@aum.edu

Reviewing hypothesis-testing

- What are the steps that go into conducting a hypothesis test?
- What is the sampling distribution?
- What is a z-score?
- What is a p-value?
- What is an α-value?

Reviewing difference of mean testing (one sample)

- Suppose you take a random sample of 100 baseball players.
- You find that the average player has a batting average of $\bar{x} = 0.275$ and $\sigma = 0.100$.
- You hypothesize that $\mu >= 0.250$. Assuming $\alpha = 0.05$ (one-tailed), can you reject H_0?

Reviewing difference of means testing (two samples)

- Suppose you gather a sample of 100 male and 100 female black and tan coonhounds.
- You find that the average weight of the males (in pounds) is $\bar{x}_m = 71$ while $\bar{x}_f = 65$ with variance parameters $\sigma_m^2 = 15$ and $\sigma_f^2 = 10$.
- Assuming $\alpha = 0.05$ (two-tailed), can we conclude with confidence that $\mu_m > \mu_f$?

The *t*-distribution

- The *t*-distribution has *a lot* in common with the *z*-distribution.
- Like *z*, a *t*-score measures how far an observation is from its expected outcome, and every *t*-score has a *p*-value affiliated with it.
- The major difference between the two has to do with a concept known as "degrees of freedom."
- The *t*-distribution is more conservative than the *z*, and its density is a function of a sample size.

An example using the t-distribution

- You gather data from 10 Alabama counties, 5 in the black belt, 5 not ($n = 10$).
- You hypothesize the counties not in the black belt are more Republican.
- Therefore, you're conducting a difference of means test.
- Because we're comparing *two means*, we have two constraints ($k = 2$).
- Therefore, $df = 10 - 2 = 8$.

Conducting a t-test

- To perform a difference-of-means test using t, we'll use similar procedures as with a $z-$test.
- Suppose we find that $\bar{X}_{bb} = 31$ and $\bar{X}_{-bb} = 69$.
- Furthermore, $\sigma^2_{bb} = 213$ and $\sigma^2_{-bb} = 91$.
- Then,

$$t = \frac{|\bar{X}_{bb} - \bar{X}_{-bb}|}{\sqrt{\frac{\sigma^2_{bb}}{n_{bb}} + \frac{\sigma^2_{-bb}}{n_{-bb}}}} = \frac{|31 - 69|}{\sqrt{\frac{213}{5} + \frac{91}{5}}} = \frac{38}{\sqrt{43 + 18}} = \frac{38}{8} = 4.75$$

- Now we can go and check a t-table to determine statistical significance.

What about other types of relationships?

- So far, we've simply been examining differences in means. Hence, the independent variable is dichotomous, and the dependent variable is continuous.
- Oftentimes, however, we're interested in more complex relationships.
- What do we do when we have variables measured at other levels?

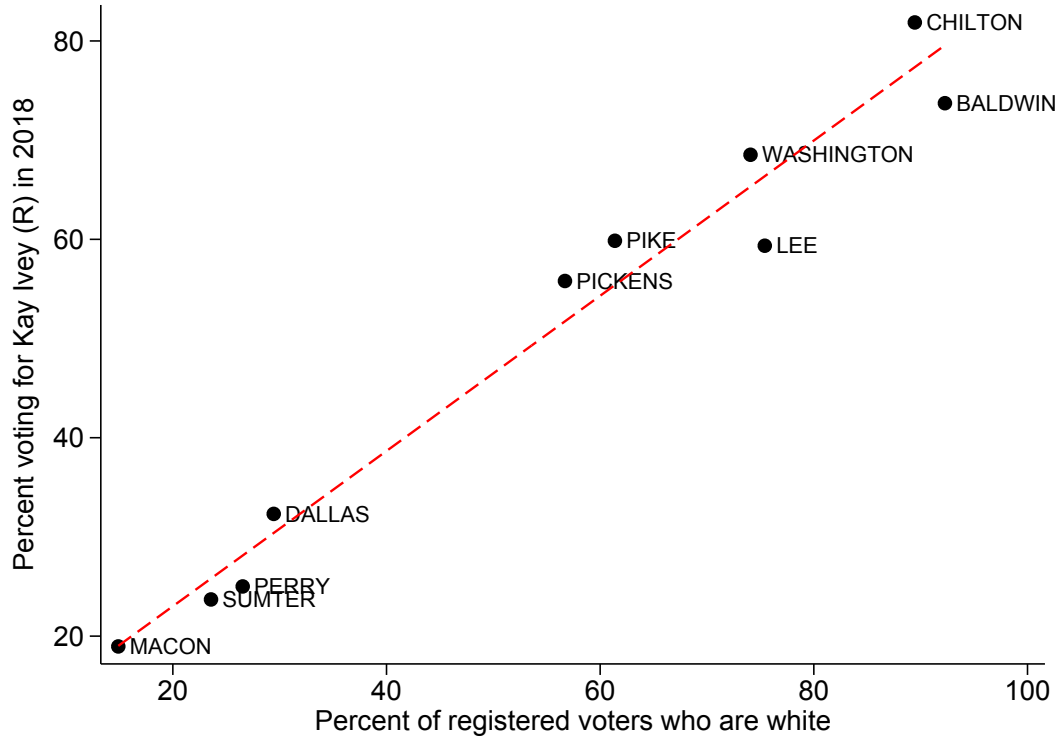
What's in a relationship?

Our methods will largely depend upon the level of measurement of our variables, but our interests are relatively constant.

- Direction of association
- Strength of association
- Statistical significance

A simple example

- H_a: Race $\rightarrow$ Vote Choice
- H_0: No relationship
- How *strong* is this relationship, and which hypothesis is more valid?



Pearson's correlation coefficient

- Pearson's correlation, $r \in [-1, 1]$, is a measure of .
- Formally:

$$r = \frac{\sum(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

- We can check for statistical significance using a t-test:

$$t = r\sqrt{\frac{n-2}{1-r^2}}.$$

An example with two continuous-level variables

- Suppose we observe individuals' height in inches, $H = \{60, 66, 62, 72\}$ and weight in pounds, $W = \{150, 180, 130, 200\}$.
- Then we have $\bar{H} = 65$ and $\bar{W} = 165$.
- Let's use our formula for r to determine the strength of association between H and W.

An example (Continued)

Obs.	H	W	$\bar{H}$	$\bar{W}$	$(H_i - \bar{H})$	$(W_i - \bar{W})$
1	60	150	65	165	-5	-15
2	66	180	65	165	1	15
3	62	130	65	165	-3	-35
4	72	200	65	165	7	35

Obs.	$(H_i - \bar{H})^2$	$(W_i - \bar{W})^2$	$(H_i - \bar{H})(W_i - \bar{W})$
1	25	225	75
2	1	225	15
3	9	1225	105
4	49	1225	245
Sum	84	2900	440

An example (Continued)

$$r = \frac{440}{\sqrt{84 \times 2900}} = \frac{440}{493.56} = 0.89$$

$$t = 0.89\sqrt{\frac{4-2}{1-0.89^2}} = .89\sqrt{\frac{2}{0.21}} = 0.89(3.09) = 2.75$$

Non-linear correlations

- Much of the data we deal with aren't measured continuously.
- We need similar methods of measuring correlations, strengths of association, and statistical significance for these variables too.
- Much of our analysis will rely upon cross-tabulations (crosstabs) and χ^2 tests.

Cross-tabulations

- Useful when examining the relationship between categorical variables (why not scatterplots?).
- We can array the observations across variables' categories to uncover a relationship.

Example of a crosstab

Suppose we surveyed 100 public administrators about their careers and their income, we created a crosstab, and we found:

	Public sector	Private sector	Total
Low income	60	10	70
High income	5	25	30
Total	65	35	100

Table 1: Effects of job type on income

What do we see?

Analysis of independence in crosstabs

- The question is, are our observations independent of chance, or is there some pattern here?
- We can go ahead and state H_a, H_0, α, etc.
- Calculate degrees of freedom (don't include row/column totals): $df = (\#\text{Columns} - 1) \times (\#\text{Rows} - 1)$
- Calculate expected frequencies. For each cell: (Row Total $\times$ Column Total)/n.
- Then:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e},$$

  where f_o is each outcome, and f_e is its expectation.
- Finally, consult a χ^2 table.

Back to our previous example

	Public sector	Private sector	Total
Low income	f_o: 60 f_e: 46	f_o: 10 f_e: 25	70
High income	f_o: 5 f_e: 20	f_o: 25 f_e: 11	30
Total	65	35	100

Table 2: Effects of job type on income

$$\begin{aligned}
\chi^2 &= \frac{(60-46)^2}{46} + \frac{(10-25)^2}{25} + \frac{(5-20)^2}{20} + \frac{(25-11)^2}{11}, \\
&= \frac{196}{46} + \frac{225}{25} + \frac{225}{20} + \frac{196}{11}, \\
&= 42.33.
\end{aligned}$$

Crosstabs with ordinal data

- So far, we've established how to analyze statistical significance on categorical data—be they ordinal or nominal.
- But we'd also like some measure of *strength* of association (like Pearson's r).
- For ordinal data, that measure is sometimes just called γ (read, "gamma").

Example of ordinal crosstabs

Suppose we examine 100 individuals' political ideology as a function of their religious affiliation and found perfect separation. If this were a Pearson's r, we'd have gotten an $r = 1.00$. Deviations from perfect separation would diminish that relationship.

	Not Evangelical	Evangelical	Total
Liberal	30	0	30
Conservative	0	70	70
Total	30	70	100

Table 3: Effects of religion on ideology

Goodman and Kruskal's gamma

- Gamma ($\gamma \in [-1, +1]$) is calculated by looking at "concordant" and "discordant" pairs in the cross-tab.

$$\gamma = \frac{C - D}{C + D}.$$

- A pair of observations is concordant if the subject who is higher on one variable is also higher on the other. Otherwise, they are discordant.

Example of ordinal crosstabs

Let's make the relationship a little more complicated. Is the relationship statistically significant?

	Not Evangelical	Evangelical	Total
Liberal	30	10	40
Conservative	20	40	60
Total	50	50	100

Table 4: Effects of religion on ideology

Example of ordinal crosstabs

Let's look at a slightly more sophisticated crosstab of political ideology and partisanship. Is the relationship statistically significant?

	Liberal	Moderate	Conservative	Total
Democrat	100	50	0	150
Independent	10	80	10	100
Republican	0	50	100	150
Total	110	280	110	400

Table 5: Effects of ideology on partisanship

Relationships between nominal data

- Ordinal measures of association are inappropriate if our cross-tab consists of nominal-level data (unless they're dichotomous).
- We could simply stop with the χ^2 test statistic and say, "There's a relationship."
- But that's a little unfulfilling.
- The ϕ coefficient can be helpful here: $\phi = \sqrt{\frac{\chi^2}{n}}$. So can Cramer's V, each of which varies between 0 and 1 where greater values reflect better fit.

An example of nominal data in a crosstab

Suppose you're interested in the number of children respondents have as a function of their religious affiliation

	Catholic	Protestant	Neither	Total
None	154	317	326	797
One	102	210	147	459
Two	162	377	194	733
Three	113	251	103	467
Four or more	118	216	77	411
Total	649	847	1,371	2,867

$\chi^2 = 92.99, \quad p < 0.000, \quad \phi = 0.18$

Table 6: Effects of religion on child-bearing (Source: GSS 2016)

More than correlation

- We've now thoroughly analyzed how examine relationships between two variables, but bivariate correlations can only take us so far.
- For one thing, they tell us virtually nothing about the effect specific values IVs might have on DVs.
- For another, they fail to take account of other, possibly confounding independent variables.
- This is a job for regression analysis—a subject to which we will next move.