Absence of evidence and evidence of absence in the effects of same sex marriage laws on rates of opposite sex marriage: mutual reinforcement of research, scholarly growth and the classroom

Alexis Dinno
Assistant Professor, School of Community Health
Portland State University

Puzzling assertions about same sex marriage

Opponents to legalization of same sex marriage have positioned it as an "assault" seeking to "weaken," "destroy" and "undermine" *opposite sex* marriage.

In a recent ruling of the United States Court of Appeals for the Ninth Circuit in Perry vs. Brown (the repeal of CA Prop 8) the proponents argue "if the definition of marriage between a man and a woman is changed, it would fundamentally redefine the term from its original and historical procreative purpose. This shift in purpose would weaken society's perception of the importance of entering into marriage to have children, which would *increase the likelihood that* [opposite sex] *couples would choose to cohabitate rather than get married*."

This kind of language is deployed by representatives, judges and pundits.

(A) Ridiculous! (B) That's a testable hypothesis!

Are changes in state rates of opposite sex marriage different in states legalizing same sex marriage, than in states with no legal same sex marriage?

And while we are at it what about in states with strong same sex civil unions?

And what about weak same sex civil unions?

Dinno, A. and Whitney, C. (2013). Same sex marriage and the perceived assault on opposite sex marriage. *PLoS ONE*, Revise and resubmit.

Data were obtained…

- Reported number of marriages by states and year from 1989-2009 from the National Center for Health Statistics
- Reported number of same-sex marriages for those states permitting them (communication with state health and vital records departments); these data permitted us to measure *opposite sex marriages* in each year.
- Estimated population age 18 and older by year and state from the US Bureau of the Census
- Proportion of the year during which same sex marriage and strong and weak civil unions were in effect from state legislative records.

Modeling marriage rates

Marriage rates are *non-stationary* processes (because $r_{ti} \approx r_{t-1i} + \zeta_i + \varepsilon_{ti}$), which makes application of traditional regression models invalid (not i.i.d.).

No good! $\rightarrow$ $r_{ti} \sim \beta_0 + \beta_t t + \beta_m m_{ti} + \beta_s s_{ti} + \beta_w w_{ti} + \varepsilon_{ti}$

Modeling marriage rates

Marriage rates are *non-stationary* processes (because $r_{ti} \approx r_{t-1i} + \zeta_i + \varepsilon_{ti}$), which makes application of traditional regression models invalid (not i.i.d.).

No good! → $r_{ti} \sim \beta_0 + \beta_t t + \beta_m m_{ti} + \beta_s s_{ti} + \beta_w w_{ti} + \varepsilon_{ti}$

(Excerpted from Dinosaur Comic 1897 by Ryan North http://www.qwantz.com/index.php?comic=1897)

Modeling (nonstationary) marriage rates

Fortunately, there's a not too involved analytic technique developed by the econometricians to deal with this type of situation termed a *single-equation generalized error correction model* which models *change* in marriage rates, rather than simply marriage rates:

$$\begin{aligned}\Delta r_{ti} = \beta_{0i} &+ \beta_{\mathrm{c}}\left[r_{t-1i} - (m_{t-1i} + s_{t-1i} + w_{t-1i} + ms_{t-1i})\right] \\ &+ \beta_{\Delta s}\Delta s_{ti} + \beta_s s_{t-1i} + \beta_{\Delta m}\Delta m_{ti} + \beta_m m_{t-1i} \\ &+ \beta_{\Delta w}\Delta w_{ti} + \beta_w w_{t-1i} + \beta_{\Delta ms}\Delta ms_{ti} + \beta_{ms} ms_{t-1i} \\ &+ \varepsilon_{ti} + \mu_{0i}\end{aligned}$$

Where:

$x_{t-1} = x$ last year,

$\Delta x_t = x_t - x_{t-1}$(the difference between x this year and last year)

Generalized error correction models

This type of model describes change in marriage rates in terms of three kinds of effects of each policy:

1. There may be an 'instantaneous' effect: the number of opposite sex marriages jumps as policy is implemented

2. There may be a 'lagged short term' effect: the linear trend in the number of opposite sex marriages changes while policy is in effect

3. There may be a 'long run' effect: the dynamic equilibrium implied by $r_{ti} \approx r_{t-1i} + \zeta_i + \varepsilon_{ti}$ may be shifted while the policy is in effect

Results presented in our first submission

We found no relationship between rates of opposite sex marriage and same sex marriage or strong or weak same sex union laws.

Table 1. Effects of same sex marriage and union laws on opposite sex marriage rates (N=1071)

	estimate[a]	s.e.[b]	95% CI[c]	q-value[d]
Instantaneous short run effects of				
same sex marriage w/o strong unions	0.0001	0.0013	-0.0025, 0.0027	> 0.9999
same sex marriage & strong unions	-0.0007	0.0014	-0.0035, 0.0021	> 0.9999
strong same sex unions w/o marriage	-0.0003	0.0007	-0.0016, 0.0010	> 0.9999
weak same sex unions	-0.0004	0.0006	-0.0016, 0.0008	> 0.9999
Lagged short run effects of				
same sex marriage w/o strong unions	-0.0003	0.0015	-0.0031, 0.0026	> 0.9999
same sex marriage & strong unions	-0.0004	0.0031	-0.0064, 0.0056	> 0.9999
strong same sex unions w/o marriage	0.0000	0.0007	-0.0014, 0.0014	> 0.9999
weak same sex unions	0.0002	0.0007	-0.0011, 0.0015	> 0.9999
Long run run effects of				
same sex marriage w/o strong unions	-0.0037	0.0153	-0.0336, 0.0262	> 0.9999
same sex marriage & strong unions	-0.0279	0.0756	-0.1760, 0.1203	> 0.9999
strong same sex unions w/o marriage	-0.0067	0.0076	-0.0215, 0.0081	> 0.9999
weak same sex unions	-0.0036	0.0083	-0.0200, 0.0127	> 0.9999

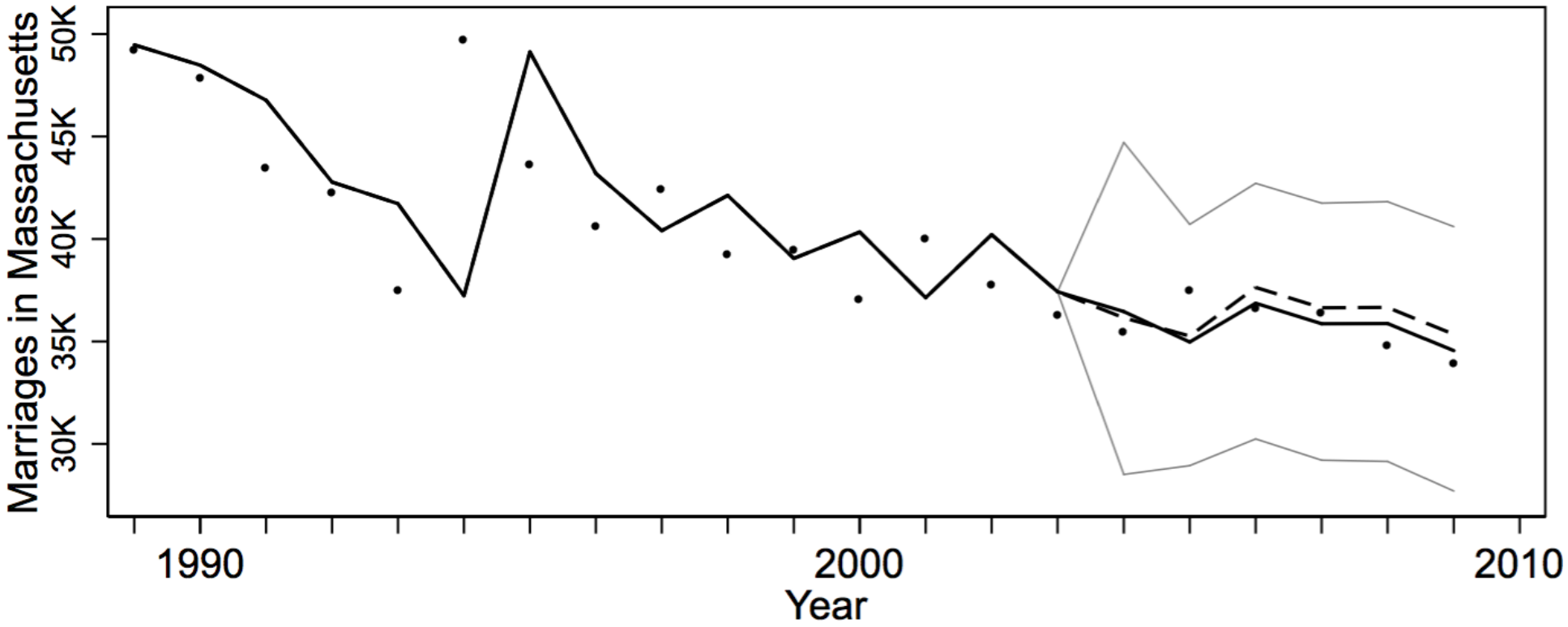
[a] The arithmetic mean of the estimates from all ten imputed data sets.
[b] Combined standard errors account for both within- and between-imputation estimate variance.
[c] 95% confidence intervals are given by the estimate $\pm 1.96 *$ s.e..
[d] q-values are p-values adjusted upward to account for twelve multiple comparisons; compare to $\alpha/2$.

Reported effects and β estimates

The presented tables report *instantaneous short run* effects, *short run lagged* effects, and *long run* effects, but the GECM I presented estimated β terms. The reported *short run lagged* and *long run* effects are combinations of the β estimates.

Check it out, yo:

$$\begin{aligned} \text{instantaneous short run effect of } x &= \beta_{\Delta x} \\ \text{lagged short run effect of } x &= \beta_x - (\beta_c + \beta_{\Delta x}) \\ \text{long run effect of } x &= (\beta_c - \beta_x) / \beta_c \end{aligned}$$


Marriages in Massachusetts
30K
35K
40K
45K
50K
1990
2000
2010
Year

One reviewer took issue with our conclusion

There is no relationship between implementation of same sex marriage or strong or weak same sex union laws and rates of opposite sex marriage.

Reviewer:
Putting the premise of the paper aside, the empirics do not settle the argument about whether rates of OSM decline in response to the implementation of SSM. The tests have very low statistical power (as indicated by the wide confidence bands around the predictions) and thus the finding of a null result tells us little.

It's also worth noting that the findings are of the "wrong" sign: coefficients on most of the SSM predictors in Table 1 are negative and the dotted lines fall above the solid lines in Figure 1. So if we don't get hung up on statistical significance, this paper actually confirms the argument of those opposed to SSM that it leads to a relative drop in OSM rates.

Three responses to this critique (1st of three)

1. We reject the idea that all effects are true effects and simply require a large enough sample size. Our interpretation is than any apparent state-level effects are due to chance alone. (Also: N = 1071, with dozens of years with same sex marriage or union laws in effect)

Three responses to this critique (2nd of three)

2. The issue of the "wrong sign" can be formally assessed: assuming that there is no relationship between same sex marriage and union laws (i.e. observed effects are due to chance alone), then we would expect 6 of the twelve effects reported in Table 1 to be positive, and 6 to be negative, in contrast to the 3 positive and 9 negative we did report.

We could formalize such an expectation as a null hypothesis with a binomial distribution, and p_0 would correspond to the probability that any reported finding is positive equals 0.5 and the number of tests n=12. Under these circumstances, the $P(X \leq 3) = 0.073 \geq 0.025$ which fails even liberal willingness to make a Type I error (α would need to equal 0.146 to conclude that enough of the reported effects were of the "wrong" sign).

Three responses to this critique (3rd of three)

Absence of evidence is not evidence of absence.

—Altman, 1995

We would like either to provide *evidence of absence of an effect*, or to revise our conclusion.

But, how might we provide *evidence of absence of an effect*? Or, to put it another way, how do we provide evidence that two quantities are *equivalent*?

Altman, D. G. and Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal*, 311(7003):485.

Equivalence testing: evidence of *sameness*?

Thankfully, I did not need to invent a new statistical machinery to provide evidence of equivalence.

Equivalence testing has its origins in pharmacology and clinical epidemiology, where a drug manufacturer does not want to be held to a higher standard than another manufacturer with respect to bringing a new drug to market.

Instead, they want to provide evidence that the new drug performs *equivalently* to the existing, FDA-approved, drug.

Equivalence is defined by ranges of tolerance

The methodological basis originates with the *one-sided hypothesis test*, and the logic that, if one selects a *tolerance* (ε)—some quantity which the researchers or regulators value as "making no difference" between two measures.

For example, a decision to apply for a grant might change if the award was $1000 versus $1,000,000 (those are *different* quantities). But a researcher might find a grant awarding $999,000 *equivalent* to one awarding $1,000,000 if their tolerance is ± $5,000.

The actual value of ε varies depending on the nature of the data and/or hypothesis test, on the nature of the question, and even on regulatory guidelines (e.g. the FDA requires ε to be a factor of 1.25 by an existing standard in bioequivalence trials: standard/1.25 and standard*1.25).

Tolerances can also be framed in terms of the test statistic under H_0.

Reframing null hypotheses for equivalence

This idea of tolerance in mind, one might want to know if the difference between two quantities (e.g. rates of opposite sex marriage with and without same sex marriage) is between $-\varepsilon$ and ε. This translates into what is termed the *two one-sided test* approach to equivalence testing:

$$H_{01}^{-}: \mu_1 - \mu_2 \geq \varepsilon \qquad \text{or} \qquad H_{02}^{-}: \mu_1 - \mu_2 \leq -\varepsilon$$

If we reject *both* these hypotheses, and we then conclude *equivalence* within ε.

The '–' denotes a "negativist" null hypothesis of equivalence which is the converse of the more common "positivist" null hypotheses of *no difference*.

Tests of difference and equivalence together

When tests of difference and equivalence are combined, four possibilities result: (1) conclude equivalence or (2) conclude difference based on congruent decisions, (3) find trivial difference—that is, is difference *is* present, but is so small as to be ignorable, and (4) indeterminate findings, where there is not enough power to reject either positivist or negativist null hypotheses.

		H_0^+ (positivist)	
		Reject H_0^+	**Not reject H_0^+**
H_0^- (negativist)	**Reject H_0^-**	*Trivial difference* (overpowered)	Conclude *equivalence*
	Not reject H_0^-	Conclude *difference*	*Indeterminate* (underpowered)

Meanwhile… I'm teaching biostatistics

Why am I only teaching my students how to test for *difference* when I am at the same time teaching my epidemiology students about publication and researcher biases *against* negative findings?

Why am I not also stressing the importance of testing for *equivalence*?

Intro Biostats-level equivalence testing: TOST

As suggested, the two one-sided tests approach to determining equivalence does not require radically new math for any of the basic hypothesis tests. For example, a paired test of *mean difference* uses the hopefully familiar *t* test:

H_0^+: $\mu_1 - \mu_2 = 0$, and the samples means are assumed to be *t*-distributed, with some degrees of freedom, ν.

$$t_\nu = \frac{\bar{d}}{s_{\bar{d}}};\ \text{where } d = x_1 - x_2$$

The test statistics for H_{01}^-: $\mu_1 - \mu_2 \geq \varepsilon$ and H_{02}^-: $\mu_1 - \mu_2 \leq -\varepsilon$ are simply:

$$t_{1\nu} = \frac{\varepsilon - \bar{d}}{s_{\bar{d}}} \text{and } t_{2\nu} = \frac{\bar{d} + \varepsilon}{s_{\bar{d}}}$$

Where *both* tests are right-hand: $P(T \geq t_1)$ and $P(T \geq t_2)$

How to integrate new stuff that I was learning?

Do I wait until I am perfectly competent with the methods before teaching?

What if there are no teachers of the material upon which I can model new material in my course?

Teaching something is an *excellent* way to learn it.

Do I use the published articles that inspired me and my own awareness of such methods?

Do I write my own material?

Useful lessons in integrating new material

Maintain a strong commitment to consistent language (e.g. θ vs. δ vs. ε for the tolerance).

Assign work with new material as required, but (initially) sneakily grading it as extra credit.

Be up front with the students about the novelty of the material, and solicit both their patience and their feedback; attend to their struggles.

Develop software to provide identical computational functionality/support for equivalence tests (i.e. the tost package, type "findit tost" in Stata).

Avoid muddier issues (i.e. there are more powerful ways to construct the tests than the simple TOST approach outlined here).

Prepare student material in 'thick-text' handouts… essentially writing a portion of new textbook chapter

My personal growth a scholar

Work with students and my attempts to implement methods in software lead me to the UMP tests.

Practice in prepping, teaching, and coding up the software gave me the confidence to respond with a revision to my manuscript which included tests of equivalence to provide evidence of absence of effects of same sex marriage on rates of opposite sex marriage.

Equivalence test results in resubmission

Opposite sex marriage were found equivalent with and without same sex marriage.

Using a UMP *t* test of mean equivalence, we rejected the negativist null hypotheses of difference given a liberal tolerance (ε = 0.5 standardized units), a strict tolerance (ε = 0.25 standardized units), and even a very strict tolerance (ε = 0.125 standardized units) across the board for same sex marriage and strong and weak same sex civil unions.

Equivalence test results

Table 2. Equivalence tests for dynamic effects on opposite sex marriage rates (N=1071)

	t[a]	$\mathrm{P}(\lvert t\rvert < \tilde{C}_{0.5})$[b,c]	$\mathrm{P}(\lvert t\rvert < \tilde{C}_{0.25})$[b,c]	$\mathrm{P}(\lvert t\rvert < \tilde{C}_{0.125})$[b] (q)[d]
Instantaneous short run effects of				
same sex marriage w/o strong unions	0.0741	0.0000	0.0000	0.0078 (0.047)
same sex marriage & strong unions	-0.5095	0.0000	0.0000	0.0191 (0.023)
strong same sex unions w/o marriage	-0.4456	0.0000	0.0000	0.0176 (0.023)
weak same sex unions	-0.5782	0.0000	0.0000	0.0208 (0.023)
Lagged short run effects of				
same sex marriage w/o strong unions	-0.1730	0.0000	0.0000	0.0108 (0.032)
same sex marriage & strong unions	-0.1435	0.0000	0.0000	0.0099 (0.040)
strong same sex unions w/o marriage	0.0181	0.0000	0.0000	0.0051 (0.061)
weak same sex unions	0.3044	0.0000	0.0000	0.0141 (0.028)
Long run run effects of				
same sex marriage w/o strong unions	-0.2426	0.0000	0.0000	0.0126 (0.030)
same sex marriage & strong unions	-0.3700	0.0000	0.0000	0.0270 (0.027)
strong same sex unions w/o marriage	-0.8857	0.0000	0.0000	0.0286 (0.029)
weak same sex unions	-0.4364	0.0000	0.0000	0.0260 (0.026)

[a] The quotient of the Table 1 estimates and their standard errors.
[b] The critical value $\tilde{C}_\varepsilon = F_{\alpha=0.05,1,\mathrm{df}=n-k,\varepsilon}$ where F is a quantile function of the noncentral F-distribution, the degrees of freedom are $n-k=1060$ from equation 2, and ε is the noncentrality parameter of F, and the $\mathrm{P}(|t| < \tilde{\theta}_\varepsilon)$ is the cumulative density of $F_{1,\mathrm{df}=n-k,\varepsilon}$ at t [56]. Because under the null hypothesis of *difference*, one of the two single-tails of the tests *must* be rejected, these p-values should be compared to α rather than to $\alpha/2$ for the common interpretation of false rejection under null hypotheses of difference [56, 60].
[c] The q-values for $\varepsilon = 0.5$ and $\varepsilon = 0.25$ are not explicitly reported because the figures remain just as the p-values within the precision of this table.
[d] $q = 12p/i$, where i is the position of ordered p-values from smallest to largest. When stepping down from largest to smallest i, all hypotheses are rejected including and subsequent to the first with $q \leq 0.05$ to control the FDR for twelve multiple comparisons.

Fin

Thank you to Professor Kelly Gonzales and Meghan Crane.

Questions?

Some citations of interest

On *negativist null hypotheses* and *uniformly most powerful tests of equivalence*: Reagle, D. P. and Vinod, H. D. (2003). Inference for negativist theory using numerically computed rejection regions. *Computational Statistics & Data Analysis*, 42(3):491–512.

A seminal paper on *two one-sided tests of equivalence*: Schuirmann, D. A. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Pharmacometrics*, 15(6):657–680.

On analogous confidence intervals supporting inference about equivalence and the interpretation of both *tests of equivalence* and *tests of difference*: Tryon, W. W. and Lewis, C. (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*, 13(3):272–277.

The current state of the art in *uniformly most powerful tests of equivalence*: Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC Press, second edition.