Statistics Writing Sample

Haven Williams

2023-05-10

Overview

What follows below is a summary of the motivating factors behind my chosen workflow of statistical inference and a worked example of such workflow.

Introduction

The Replication Crisis, Broadly

Statistics and the science of uncertainty have taken leaps over the last decade while also uncovering some sordid truths: modern laptops have become yesterday's supercomputers and as the capabilities of scientists have improved dramatically so too has their clarity that many results across all scientific literatures fail to replicate in further studies. This mass realization, commonly known as the "replication crisis," has awakened a new sense of urgency amongst scientists, including social scientists, of the need for better practices. An example of how mainstream this recognition has become, in 2021 the Nobel Prize in Economics was awarded to economists Guido Imbens and Joshua Angrist of MIT for their "methodological contributions to the analysis of causal relationships." It is notable that Angrist is also a prolific writer of concerns related to the replication crisis.

A proposed alternative: Andrew Gelman's framework

One prominent statistician has a potential framework for dealing with complex social science questions in an era of increasing awareness of statistical nuance. Andrew Gelman, acclaimed professor at Columbia University, has a novel way workflow for statistical practice in the social science world that has particular promise for social science.

Gelman flips the conventional framework entirely. Most social science statistical analysis is conducted in the following way: • Assume a null hypothesis of zero effect. • Interpret results of statistical analysis, looking for variables of high levels of "statistical significance." • If those results hold up to scrutiny, assume support of an alternative hypothesis with non-insignificant effect sizes.

Gelman does the following: • Assume a null hypothesis whose effect size and magnitude is predicted by an informed researcher. • Fit a model that has those expectations baked into the analysis. • If the model's results, including both effect sign and magnitude, produce no meaningful differences from the researcher's expectations under scrutiny, then the researcher's predicted effects are at play. If this fails, then there is something truly unknown that requires further scientific study.

Gelman's work holds promise in the social sciences for the following reasons: • Social science effect sizes are often small and thus need high levels of specificity • Social science effects are often dynamic and interrelated and researcher expectations of non-normality of non-constant effect sizes are easier to start with than to discover in data summaries over complex dimensions • In the context of the replication crisis, a workflow that promotes a focus on sign errors and magnitude errors should be paramount.

And Gelman’s work in Bayesian inference alleviates some secondary concerns, especially in a realm like international relations where experimental evidence is often nonexistent.

Bayesian frameworks allow for: • Flexibility of priors in multilevel-modeling: in multilevel Bayesian modelling we have supreme flexibility of prior choices. • We can more easily incorporate formal theory by deriving and directly integrating statistical distributions based on theory into our models. • With census-level application of Bayesian statistics not violating classical frequentist assumptions that we use less than 10% of the population in our analysis, we do not need to rely upon multiple-universe explanations for inference of macroeconomic events.

And while the biggest drawback to Bayesian inference is a steeper learning curve, we are more likely within the framework proposed by Gelman to have correct inferences from the beginning.

What follows is a miniaturized example of Gelman’s proposed hypothesis testing workflow over fake data that simulates support for universal healthcare in the American electorate.

The overall summary is this. I will perform power analysis of a hypothesis test of two proportions to get a desired sample size. I will then simulate data of the desired dimension with two

1: Power Analysis

Suppose that a politician is concerned about the ability for their legislature to pass a bill to create a universal healthcare market in their healthcare sector. The politician believes that if more than 50% of the public agrees with a recently proposed plan that the bill will pass: the politician suspects that support is around 60% based on several exogenous sources of data.

If we simply wanted to find the results of a survey of support in a classic single-issue poll, we could use the canonical formula $ CI = z $, but this formula isn’t appropriate for the question posed. The question is, how certain can we be that the true proportion is different than our proposed proportion, not the magnitude of an issue in isolation. In order to answer this question accurately, we turn to power analysis to see how large a sample we would need to construct a survey for any given level of power.

What we’re looking for is a measure of our statistic that is robust to type 1 and type 2 errors. We must find the total number of standard deviations away from the reference statistic our sample statistic must be in order to assume they are different. That is, we must add together a confidence interval excluding our null statistic and the upper portion of the cumulative distribution of our reference statistic so we are looking at statistics both above and greater than our minimum proportion.

Essentially we set uncertainty around the reference statistic plus uncertainty around the suspected statistic and set them to 0 so our expected result is different from our reference result. In this case we model the uncertainty around the reference statistic as a 95% confidence interval and the uncertainty of the suspected statistic as the 80th percentile of the quantile distribution around the suspected statistic.

We get: $0.5 + 1.96 * s.e. = .6 - 0.84 * s.e.$ for our power analysis. If we substitute in $s.e. = \frac{0.5}{\sqrt{n}}$ for the standard error of two proportions, we calculate that we need a sample of 196 to be 80% certain that our survey returns the correct result that support for the policy is above 50% when we believe the true level of support is 60%. For a difference in any two proportions, the general code is below and the first implementation of the function is the result described, while the second implementation is a much stricter result with a small reference effect and a small suspected effect at the same power level. A final implementation shows a 95% power level for our hypothesized statistics and shows the sample number required would be 324, still a reasonable number for a national survey.

```
#This yields the same results as the power calculator for a conservative null hypothesis
power_calculator <- function(null_effect, alternative_effect, power_level){
  if(alternative_effect > null_effect){
    required_sample_size = ((1.96 + qnorm(power_level)) * .5 / (alternative_effect - null_effect)) ** 2
    print(required_sample_size)
  }
  else {
    required_sample_size = ((-1.96 - qnorm(power_level)) * .5 / (alternative_effect - null_effect)) ** 2
    print(required_sample_size)
  }
}

power_calculator(.5, .6, .8) #196
```

```
## [1] 196.227
```

```
power_calculator(.04, .03, .8) #19622
```

```
## [1] 19622.7
```

```
power_calculator(.5, .6, .95) #324
```

```
## [1] 324.8742
```

2: Fake Data Simulation

1. Generating data

I generate 324 fake voters whose ages are all 18 or older.

```r
# Set the seed for reproducibility
set.seed(123)

# Number of data points
n <- 324

# Generate the data
data <- data.frame(
  # Binary column with 1s 60% of the time
  Support = rbinom(n, 1, 0.6),

  # Continuous income data, assuming a mean of $50,000 and standard deviation of $10,000
  Income = rnorm(n, mean = 50000, sd = 10000),

  # Continuous age data, assuming a mean of 40 and standard deviation of 10
  Age = rnorm(n, mean = 40, sd = 10),

  # Binary column for Republican, Democrat if 0
  Republican = rbinom(n, 1, 0.5),

  # Binary column for Male, 0 if Female
  Male = rbinom(n, 1, 0.5)
)

# Function to check and regenerate observations for ages less than 18. Because the entire observation is redistributed from independent draws of different distributions, this censoring does not bias our fake data.
regenerate_age <- function(df) {
  while(any(df$Age < 18)) {
    df$Age[df$Age < 18] <- rnorm(sum(df$Age < 18), mean = 40, sd = 10)
  }

  return(df)
}

# Use the function to regenerate ages
data <- regenerate_age(data)

# Print the first few rows of the data
head(data)
```

```
##   Support   Income      Age Republican Male
## 1       1 37398.45 49.14773          0    0
## 2       0 82410.40 34.25605          0    1
## 3       1 45831.42 56.26881          0    0
## 4       0 52982.28 36.19043          1    1
## 5       0 56365.70 38.94216          0    1
## 6       1 45162.19 54.04050          1    0
```

```r
summary(data) #Our age function works
```

```
##     Support          Income            Age          Republican    
##  Min.   :0.000   Min.   :23391   Min.   :19.22   Min.   :0.0000  
##  1st Qu.:0.000   1st Qu.:44277   1st Qu.:33.43   1st Qu.:0.0000  
##  Median :1.000   Median :50425   Median :40.29   Median :1.0000  
##  Mean   :0.608   Mean   :50542   Mean   :40.20   Mean   :0.5432  
##  3rd Qu.:1.000   3rd Qu.:57045   3rd Qu.:46.43   3rd Qu.:1.0000  
##  Max.   :1.000   Max.   :82410   Max.   :66.92   Max.   :1.0000  
##       Male       
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0000  
##  Mean   :0.4784  
##  3rd Qu.:1.0000  
##  Max.   :1.0000
```

```
summary(data$Support) #We see a mean of .608 to indicate approximately 60% support.
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   1.000   0.608   1.000   1.000
```

2: Bayesian regression with informed priors

I write a logit model with a normal prior around .6 to represent our beliefs that the true statistic lies at a proportion of 60% of the public favoring the bill. This allows us to probe our indicator variables and see if there is anything else that may cause our level of support.

```
# Specify the model

model <- stan_glm(
  formula = Support ~ Income + Age + Republican + Male,
  data = data,
  family = binomial(link = "logit"),
  prior = normal(location = .6),
  chains = 4,
  iter = 2000,
  cores = 4,
  refresh = 0
)
```

3: Convergence

3a: Summary Statistics

1. Print() summary to see covariate effects.

```
summary(model)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       binomial [logit]
##  formula:      Support ~ Income + Age + Republican + Male
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 324
##  predictors:   5
##
## Estimates:
##               mean   sd   10%   50%   90%
## (Intercept)  0.9    0.8 -0.1   0.9   1.9
## Income       0.0    0.0  0.0   0.0   0.0
## Age          0.0    0.0  0.0   0.0   0.0
## Republican   0.0    0.2 -0.3   0.0   0.3
## Male        -0.4    0.2 -0.7  -0.4  -0.1
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 0.6    0.0  0.6   0.6   0.7
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable
(for details see help('summary.stanreg')).
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  3301
## Income        0.0  1.0  4204
## Age           0.0  1.0  2913
## Republican    0.0  1.0  2563
## Male          0.0  1.0  2548
## mean_PPD      0.0  1.0  3045
## log-posterior 0.0  1.0  1537
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective
sample size, and Rhat is the potential scale reduction factor on split chains (at convergence Rh
at=1).
```

We see that the posterior distribution at the 90th percentile for most of our covariates is 0, so there is not an association between those variables and our outcome variable. We do see support decline in being male. The model is weighted by male since being Male is exogenous to the question, while partisanship is potentially endogenous so there is no weighting on that column. Further simulations should use an interactive effect between the two, but this will dramatically increase the required sample size by a factor of four. This is because the main effect with an interaction is measured by standard error $\sqrt{\sigma^2/(n/4) + \sigma^2/(n/4) + \sigma^2/(n/4) + \sigma^2/(n/4)} = 4\sigma/\sqrt{n}$ and the four is visible in the right hand numerator.

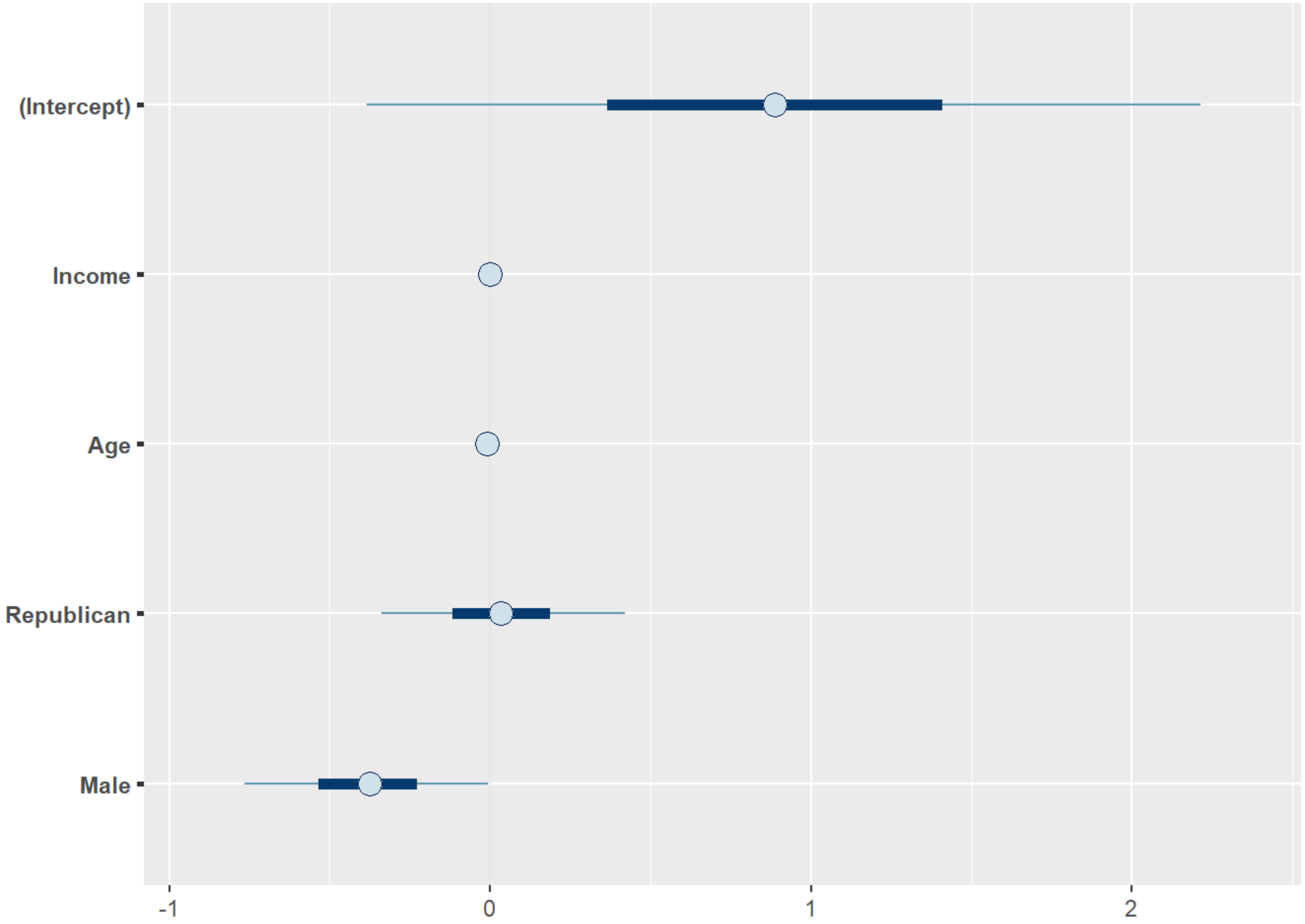
```
summary(data$Male)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0000  0.0000  0.4784  1.0000  1.0000
```

3b: Chain plotting

1. If we see crimps in the data, we can use stacking, a method that uses cross-validation to resample chains and weight them by importance. This is to be used only when
2. We can check convergence visually.
3. These checks may need to be further reevaluated if re-simulating the chains multiple times leads to higher rates of failure. It is always possible that one more additional chain invalidates the results, so checking convergence should not be the only method to diagnose model convergence.

```
mcmc_intervals(model)
```



4: Assessments

4a: Cross validation

Cross validation is a standard component of statistical inference. It involves leaving one or multiple members of a dataset out and repeating the analysis, essentially checking for robustness to outliers.

```
loo(model)
```

```
##
## Computed from 4000 by 324 log-likelihood matrix
##
##          Estimate  SE
## elpd_loo   -220.4 4.4
## p_loo         5.2 0.2
## looic       440.9 8.8
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

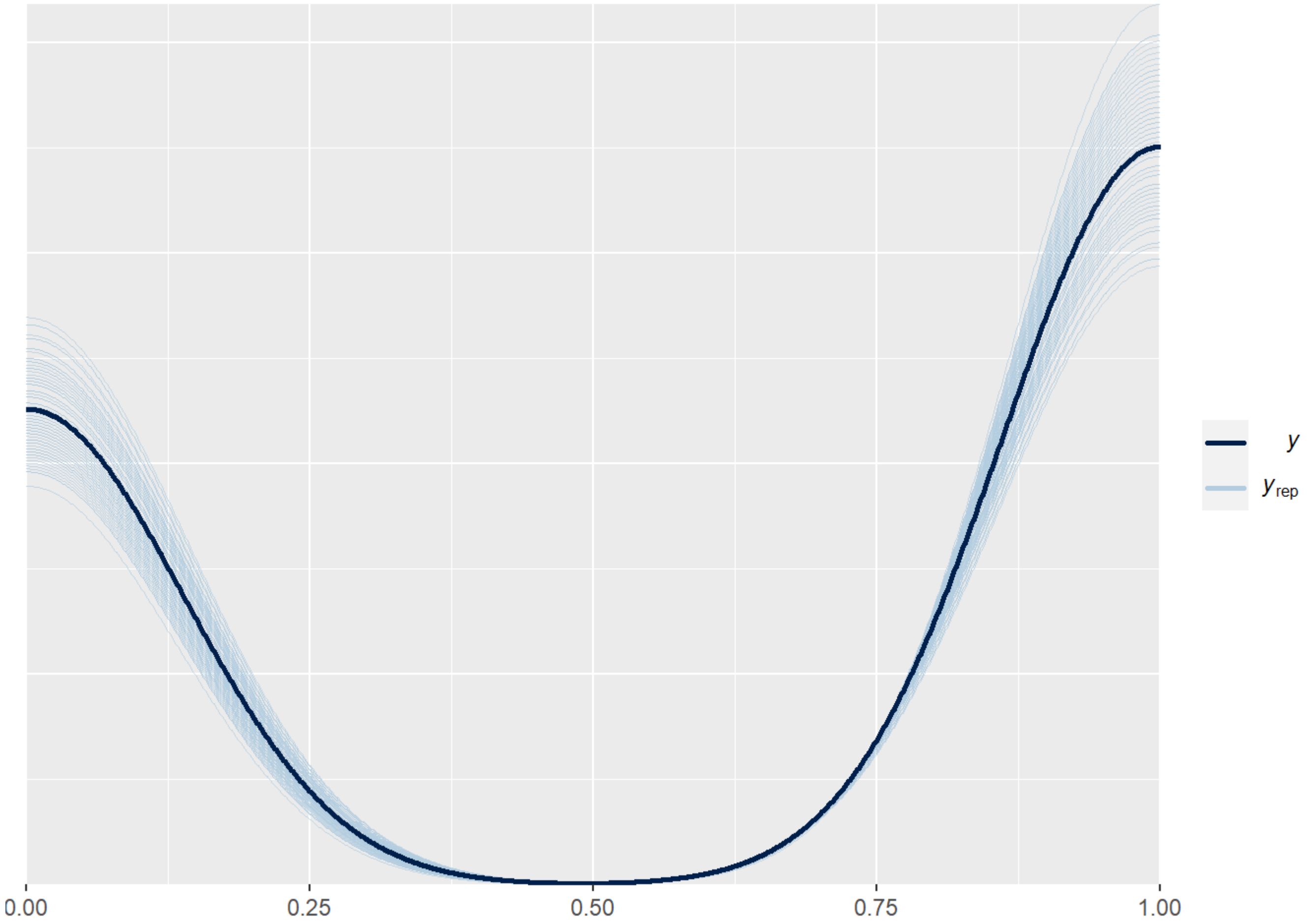
Because all of our Pareto K diagnostics are below .7, the model passes the LOO cross validation test.

4b: Posterior predictive checks

Posterior predictive checks, similar to prior predictive checks, generates fake data based on the posterior distribution and checks whether the posterior distribution from the test matches the posterior distribution from the simulations.

```
#Replicate results from simulations from the model
y_rep <- posterior_predict(model, draws = 100)

#Plot in comparison to the actual data
ppc_dens_overlay(data$Support, y_rep[0:100, ])
```



Since we are generating our data from an admittedly sterile fake dataset, we see our posterior predictive check's density plot maps perfectly, indicating good model fit.

4c: Type S and Type M errors

Type S orders specify the probability that the incorrect sign has been applied to our effect measurement and type M errors specify the factor by which the effet is exagerated. With an effect size of -.4 and a standard deviation of .2 for the effect of being male on support for the legislation, we find the following results from the retrodesign package's self-titled function that calculates the power of a test, the type S error, and the type M error.

```
summary(model)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       binomial [logit]
##  formula:      Support ~ Income + Age + Republican + Male
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 324
##  predictors:   5
##
## Estimates:
##                 mean   sd   10%   50%   90%
## (Intercept)  0.9    0.8 -0.1   0.9   1.9
## Income       0.0    0.0  0.0   0.0   0.0
## Age          0.0    0.0  0.0   0.0   0.0
## Republican   0.0    0.2 -0.3   0.0   0.3
## Male        -0.4    0.2 -0.7  -0.4  -0.1
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 0.6    0.0  0.6   0.6   0.7
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable
(for details see help('summary.stanreg')).
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  3301
## Income        0.0  1.0  4204
## Age           0.0  1.0  2913
## Republican    0.0  1.0  2563
## Male          0.0  1.0  2548
## mean_PPD      0.0  1.0  3045
## log-posterior 0.0  1.0  1537
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective
sample size, and Rhat is the potential scale reduction factor on split chains (at convergence Rh
at=1).
```

```
retrodesign(A = -0.4, s = 0.2)
```

```
## $power
## [1] 0.5160053
##
## $typeS
## [1] 7.263595e-05
##
## $exaggeration
## [1] 1.381926
```

While the test itself remains theoretically very powerful the implementation of the retrodesign function shows a power of .51 when used to measure the effect size and standard deviation of the Male categorical variable. This low level of power shows that, while the survey replicated here might demonstrate a population-level effect, the survey is not large enough to distinguish the effects of specific covariates.

5: Future Extensions

This is only the tip of the iceberg for a full statistical assessment of support in a survey. Further extension (in real world surveying) can and should including multinomial comparisons for levels of support across categorical data including race. The simplicity of this toy example is that is can be easily extended to other areas.

Furthermore, a rigorous analysis could be done with multilevel modelling to capture performance within districts. In such a framework, a multilevel model could show the

And in the case of non-response bias or incomplete survey responses, data imputation may be cautiously used. This is especially useful when looking to combine results of disparate polling instances where highly detailed, district-level data may not exist, yet where known, census-level data can provide starting values that other districts may inform the standard deviations of.