A Web-Based Application to Support Applied Bayesian Data Analysis

James Uanhoro
Quantitative Research, Evaluation & Measurement
Department of Educational Studies, Ohio State University

Abstract

I present a freely accessible web-based Bayesian data analysis application built on the Stan computation engine. The goal of the application is to make Bayesian data analysis accessible while demonstrating two practices of applied Bayesian data analysis. The first practice is that of including subjective prior information in data analysis. The second practice is communicating study effects probabilistically using posterior samples. To demonstrate these practices, the application elicits prior information from the user and returns plots that communicate effects probabilistically. The application implements two-group comparisons for: unbounded continuous outcomes (t-regression and quantile regression); bounded continuous outcomes (beta regression); and binary outcomes (binomial regression). The application is under continuous development; by the date of presentation, the application will include additional techniques.

Link to application: `https://www.jamesuanhoro.com/project/bms/`

Notation in text:
$\mathbf{y}$: continuous variable; $\mathbf{x}$: binary variable ($\mathbf{x} \in \{0, 1\}$);
$\Phi(\cdot)$: Standard normal quantile function
Generative models always follow mean-scale notation.

Words in text: 1993

Introduction

Bayesian estimation is an increasingly common approach to data analysis in the social sciences. An appealing justification for Bayesian methods is that stated by Kruschke (2013): a Bayesian analysis provides a rich description of parameters of interest relative to the information provided by analogous frequentist analysis. A challenge to greater adoption of

Bayesian data analysis is that the majority of Bayesian statistical packages are script-based. This increases the burden on substantive researchers. Another challenge to widespread adoption is one many advocates of Bayesian methods have heard: "how do I choose the prior?"

In this paper, I present a web-based application that attempts to make Bayesian methods accessible to substantive researchers. The operation of the application is grounded in two beliefs:

1. The diligent researcher always has prior information, although a data analyst may be required to translate this information into a prior distribution.

2. One should communicate model insights probabilistically using posterior samples (distributions) of parameters.

These beliefs stem from a subjective Bayesian approach (Goldstein, 2006), and an interest in encouraging practices that scale to complex problems. Specific practices that follow from these beliefs are:

1. I match the expected range of parameters to the 95% quantile interval of distributions for specifying prior distributions (Greenland, 2006).

2. I emphasize posterior samples (distributions) for communicating study results, and de-emphasize Bayes factors. Bayes factors are a common recommendation in quantitative psychology (e.g. Dienes & Mclatchie, 2018; Morey, Rouder, Verhagen, & Wagenmakers, 2014; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2015). However, Bayes factors are not readily scalable (Gelman & Rubin, 1995), unlike posterior samples which are always available.

The application sits atop Stan (Carpenter et al., 2017), accessed via Python. Stan uses the No-U-Turn sampler allowing for highly efficient posterior sampling (Hoffman & Gelman, 2014). In the next section, I outline the statistical models implemented in the application, and conclude with a demonstration comparing faculty salaries by gender.

Outline of statistical models

Heteroskedastic $t-$regression: Two−group comparison for unbounded continuous data

For continuous outcome $\mathbf{y}$ with binary group membership $\mathbf{x}$, I adopt the Bayesian $t-$regression model by Kruschke (2013), which allows for outliers in the data and heteroskedasticity by group membership:

$$\mathbf{y} \sim t\left(\nu,\ \alpha + \beta\mathbf{x},\ e^{\delta+\gamma\mathbf{x}}\right) \tag{1}$$

where ν is the degrees of freedom of the t distribution − smaller values of ν imply outliers in $\mathbf{y}$. The scale equation is on the log-scale to ensure the scale is always positive. For priors, I assume:

$$\begin{aligned} &\nu \sim \text{Gamma}(\text{shape} = 1,\ \text{rate} = 0.1),\ \alpha \sim \text{Cauchy}(0, s_\alpha = 5),\ \delta \sim t(3, 0, s_\delta = 1) \\ &\beta \sim \mathcal{N}\big(0, s_\beta = u_\beta \big/ \Phi^{-1}(0.975)\big),\ \gamma \sim \mathcal{N}\big(0, s_\gamma = \ln(u_\gamma) \big/ \Phi^{-1}(0.975)\big) \end{aligned} \tag{2}$$

The gamma prior on ν assumes a 95% quantile interval of (0.25, 36.9), permitting distributions of the data that range from *pathological* ($\nu \leq 1$) to near-normality ($\nu > 30$) (Ding, 2014). α and δ are intercepts for the mean and log-scale respectively. I assign them large-variance priors as the researcher may not have much information on the intercepts (group mean) themselves.

However, diligent and experienced researchers will have expectations for the differences between groups. The user is asked for the largest difference between both groups they would find believable (u_β). By setting $s_\beta = u_\beta \big/ \Phi^{-1}(0.975) = u_\beta \big/ 1.96$, we define a scale for β that a-priori assumes a 95% chance that β lies in the $(-u_\beta, u_\beta)$ range, based on the *empirical rule*.

The user is also asked for a maximum ratio of both group standard deviations, u_γ. I set $s_\gamma = \ln(u_\gamma) \big/ \Phi^{-1}(0.975)$. The log transformation on u_γ transforms the ratio to an

additive operation and dividing the log-transformed value by $\Phi^{-1}(0.975)$ a-priori assumes a 95% chance that γ falls in the $\big(-\ln(u_\gamma), \ln(u_\gamma)\big)$ range.

The user is required to select a number of iterations between 1000 and 2000. The application runs the set number of iterations across 4 chains. The first half of iterations are used to warm-up the sampler, the final half are retained for posterior inference. The user also has to enter a desired quantile interval for summarizing inference (95% by default).

Heteroskedastic quantile regression: Two−group comparison for unbounded continuous data

I adopt the Bayesian quantile regression model by Yu and Moyeed (2001) which is based on the asymmetric Laplace distribution (ALD). The ALD permits location-scale modeling at a specific percentile of the outcome. The ability to test group differences at specific percentiles can be substantively important. For example, an intervention may target students with low proficiency in a particular subject, but the intervention is delivered to students at all levels of proficiency. In such an instance, testing the treatment effect at a lower percentile better assesses the intervention than a test of mean differences.

The Bayesian model is:

$$\mathbf{y} \sim \text{ALD}\left(\alpha + \beta\mathbf{x},\ e^{\delta+\gamma\mathbf{x}},\ p\right) \tag{3}$$

where $\alpha + \beta\mathbf{x}$ and $\delta + \gamma\mathbf{x}$ are the models for the mean and log-scale respectively, while p is the percentile of interest provided by the researcher. The model makes the same assumptions about priors as in equation 2, requesting user inputs for differences in group means and the ratio of group standard deviations.

Beta regression: Two−group comparison for bounded continuous data

Beta regression (Smithson & Verkuilen, 2006) is a flexible regression model for bounded continuous data. Bounded data are commonplace in educational settings, most commonly in the form of test scores, GPA. Additional examples include percentages and

averaged Likert scores. Although, one may analyze bounded data using models that ignore the bounds (e.g. normal and t distributions), this practice of ignoring the bounds can result misleading inference especially when the data are clustered at the bounds. An example is analyzing an outcome with a marked ceiling or floor effect.

Prior to analysis, the data are first re-scaled to the unit interval (0, 1) using the formula: $\mathbf{y}' = (\mathbf{y} - \min(\mathbf{y}))/(\max(\mathbf{y}) - \min(\mathbf{y}))$, where the minimum and maximum values are based on the range of the response scale not the sample statistics. The re-scaling ensures that the data are probabilities. Secondly, the beta distribution assumes the data fall between 0 and 1 (exclusive). Hence, if there are 0 and/or 1 values after re-scaling the data to the unit interval, I implement a common re-scaling (Smithson & Verkuilen, 2006): $\mathbf{y}'' = [\mathbf{y}'(N-1) + 1/2]/N$, where N is the total sample size.

After the data are transformed, the Bayesian model is:

$$\mathbf{y}'' \sim \text{beta}(p_0 + p_d\mathbf{x},\ \boldsymbol{\kappa}_{[\mathbf{x}+1]}) \tag{4}$$

where the beta distribution is parameterized by the mean and sample size (Kruschke, 2011). p_0 is the average probability when $\mathbf{x} = 0$ (the intercept) and p_d is the difference in probabilities between both groups. $\boldsymbol{\kappa}$ is the sample size parameter which is different for both groups. For priors, I assume:

$$\begin{aligned} &\boldsymbol{\kappa} \sim \text{Gamma}(2, .1),\ p_0 \sim \text{beta}(u_0, u_0) \\ &p_d \sim \mathcal{N}\left(0,\ u'_d \Big/ \Phi^{-1}(0.975)\right), \quad u'_d = u_d \Big/ \Big(\max(\mathbf{y}) - \min(\mathbf{y})\Big) \end{aligned} \tag{5}$$

The sample size parameter is set up to permit a large variety of positive values. If the user suggests that the data are extreme and close to the bounds, u_0 is set to 1 permitting a flat prior on p_0. Otherwise, u_0 is set to 2, such that there is increasingly lower prior probability on extreme values for p_0. The user provides an estimate of the maximum difference between the groups, u_d. This value is re-scaled to the unit interval (u'_d) and the

empirical rule is again used to construct the prior, such that there is a 95% chance that p_d lies between $(-u'_d, u'_d)$. Finally, the estimated group means, as well as the difference between groups, are converted back to the original scale and returned to the user.

Binomial logistic regression: Two−group comparison for binary data

For binary outcomes, I utilize a simple binomial logistic regression model:

$$\begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \sim \text{Binomial}\left(\begin{bmatrix} t_1 \\ t_2 \end{bmatrix}, \begin{bmatrix} \left(1 + e^{-(\alpha+\beta)}\right)^{-1} \\ (1 + e^{-\alpha})^{-1} \end{bmatrix} \right) \tag{6}$$

where s_1 and s_2 are the number of *successes* for groups 1 and 2 out of t_1 and t_2 trials respectively. α is the expected log-odds of success for group 2, while β is the difference between group 1 and group 2 on the log-odds scale.

For priors, I assume: $\alpha \sim \mathcal{N}(0, s_\alpha)$ and $\beta \sim \mathcal{N}(0, s_\beta)$; s_α and s_β are based on user input. The user is asked if the event is known to be an extreme event. For extreme events, we set $s_\alpha = 5$, such that the average probability of success for group 2 is a-priori assumed to have a near 0 lower and near 1 upper limit. Otherwise, we set $s_\alpha = 2$ such that the average probability of success for group 2 is assumed to have a 95% a-prior interval of $\big((0.019, .981),\ \text{inv-logit}\big(\Phi^{-1}(.975) \times 2\big)\big)$.

Next, the program requires that the user set an upper limit for the odds ratio (u_β). The program recommends 2 for outcomes that are difficult to change or for controversial interventions and 10 for relations that are obvious. I set $s_\beta = \ln(u_\beta) \Big/ \Phi^{-1}(0.975)$ since β is additive on the log-odds scale. The remaining user inputs relate to the number of posterior samples and the requested quantile intervals as with other methods.

Demonstration: Comparing salaries of female and male faculty within a college

I assess gender differences in 2017 salary data for 141 tenured faculty within a college at a large public mid-western institution, see Figure 1. The salaries were publicly available

data while gender was determined using pronouns in faculty bios. There were 82 women (average salary = US$98K, SD = US$27K, median = US$93K) and 59 men (average salary = US$115K, SD = US$43K, median = US$100K). As seen in Figure 1 and as supported by the descriptive statistics, the data were right-skewed and the male salaries appeared to have a higher location and variation than the female salaries.

We begin by analyzing these data with the t-regression model. The data were divided by 10,000 such that the salaries are in units and tens. I present the series of steps to perform the analysis in Figure 2 − the male salaries were entered for group 1 hence differences were calculated as male − female salaries. I assumed that that average difference in salaries would not exceed US$50,000 (5 in step 3), and that the ratio of group standard deviations would not exceed 4. The model summaries are presented in Figure 3. The program also creates four files available for download:

1. A plot for location differences presented in Figure 4.
2. A plot for scale ratios similar to the location differences plot.
3. A summary file containing summary statistics (including the potential scale reduction factor) for the estimated parameters, and the posterior samples for the estimated parameters that may be used for model criticism (Gelman, Meng, & Stern, 1996).
4. A rank plot visualizing the ranks of posterior samples for each parameter across chains (Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2019), see Figure 5.

As seen in Figure 4, there was more than 95% chance that male faculty were paid more than female faculty on average. However, the US$8,000 difference (Figure 3) pales in comparison to the sample mean differences (115K−98K=17K) and is closer to the sample median difference (100K−93K). This is because the t-regression identified outliers in the data (as noted by a degrees of freedom value of 2.2, see Figure 3). And the outliers are down-weighted during estimation. Also interesting is the difference in scale between both groups. The scale of male salaries was 76% $((1.757 - 1) \times 100\%)$ larger than the scale

of female salaries suggesting much greater variation in the salary data for men, as seen in Figure 1. The effective sample size values are large enough for reliable inference and the rank plots in Figure 5 suggested the sampler converged and the different chains mixed adequately for the mean difference and scale ratios.

Given the importance of this topic, I opt for the quantile approach to further delve into the data. We assess the gender differences in salaries when salaries are low (25th percentile), at the median and when salaries are high (75th percentile). I simply modified the user inputs to specify the 25th percentile as seen in Figure 6 and repeated this process for the 50th and 75th percentiles. And at the 75th percentile, I specified a maximum believable salary difference of US$100K given how high salaries could rise.

The application reports all the plots and files as before but we only present the model summaries (Figure 7) and location difference probability plots (Figure 8) for the 75th percentiles results. The average salary difference (alongside 95% intervals) at the 25th, 50th and 75th percentiles were US$4,180 (95% interval: −2,810, 11,010), US$6,890 (−1,280, 15,140) and US$28,940 (12,380, 47,640) respectively. All this suggests that highly paid women (75th percentile = US$105K) were paid a lot less than highly paid men (75th percentile = US$134K). Also, worth noting is the scale ratio at high-end salaries (1.81 for men:women). In addition to being paid more, men's salaries are less determined than women's salaries. I note that these results by themselves do not indicate bias (Billard, 2017).

Conclusion

I have outlined the statistical model behind the application for the four Bayesian methods implemented in the application, and demonstrated the functionality in the case of a continuous outcome. In the demonstration, a Bayesian approach enhanced our ability to understand and communicate parameters of interest as seen in the summary plots that permit direct probabilistic statements about parameters. Moreover, the requirements for performing these analyses is a sound understanding of the subject matter and a good un-

derstanding of the outcome measure. These are expectations that one should have of any diligent researcher. Hence the application provides a way for diligent researchers to gain rich information about patterns in data in simple applications. I look forward to enhancing the application to perform more complex techniques in the coming months. By the time of presentation, I expect the application to include two sample comparisons for count and ordinal data.

References

Billard, L. (2017, jan). Study of Salary Differentials by Gender and Discipline. *Statistics and Public Policy*, *4*(1), 1–14. doi: 10.1080/2330443X.2017.1317223

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). doi: 10.18637/jss.v076.i01

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin and Review*, *25*(1). doi: 10.3758/s13423-017-1266-z

Ding, P. (2014). Bayesian robust inference of sample selection using selection-t models. *Journal of Multivariate Analysis*, *124*, 451–464. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0047259X1300256X` doi: https://doi.org/10.1016/j.jmva.2013.11.014

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760. Retrieved from `http://www.jstor.org/stable/24306036`

Gelman, A., & Rubin, D. B. (1995). Avoiding Model Selection in Bayesian Social Research. *Sociological Methodology*, *25*, 165–173. Retrieved from `http://www.jstor.org/stable/271064` doi: 10.2307/271064

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*(3), 403–420. Retrieved from `https://projecteuclid.org:443/euclid.ba/1340371036` doi: 10.1214/06-BA116

Greenland, S. (2006, jan). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology*, *35*(3), 765–775. Retrieved from `https://doi.org/10.1093/ije/dyi312` doi: 10.1093/ije/dyi312

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623. Retrieved from `https://dl.acm.org/citation.cfm?id=2627435.2638586`

Kruschke, J. K. (2011). *Doing Bayesian data analysis:A tutorial with R and BUGS*. Academic Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. Retrieved from `http://doi.apa.org/getdoi.cfm?doi=10.1037/a0029146` doi: 10.1037/a0029146

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science. *Psychological Science*, *25*(6), 1289–1290. Retrieved from `http://journals.sagepub.com/doi/10.1177/0956797614525969` doi: 10.1177/0956797614525969

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *20*(December). doi: 10.1037/met0000061

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71. doi: 10.1037/1082-989X.11.1.54

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2019). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *arXiv preprint arVix: 1903.08008*. Retrieved from `https://arxiv.org/abs/1903.08008`

Yu, K., & Moyeed, R. A. (2001, oct). Bayesian quantile regression. *Statistics and Probability Letters*, *54*(4), 437–447. doi: 10.1016/S0167-7152(01)00124-9

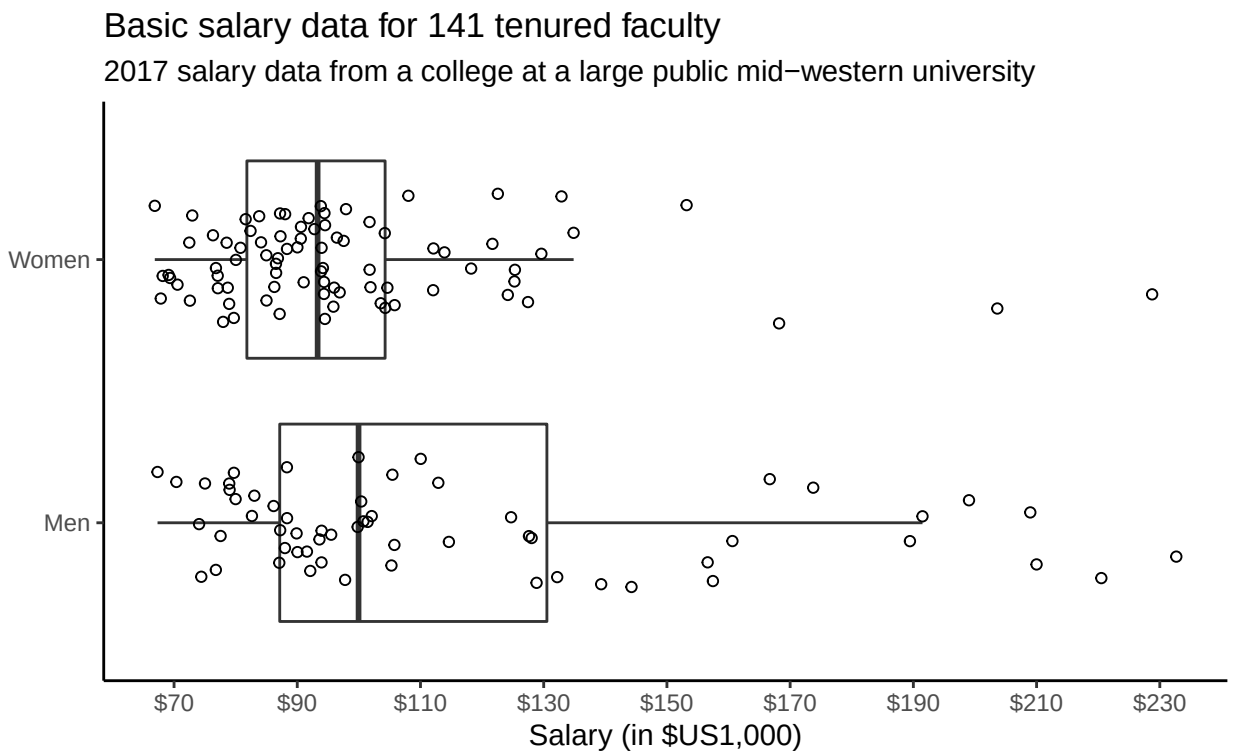

Figure 1. Basic salary for 141 tenured faculty within a college at a large public mid-western institution by gender. The boxplot shows the median and interquartile range by group. Each data point represents a salary.

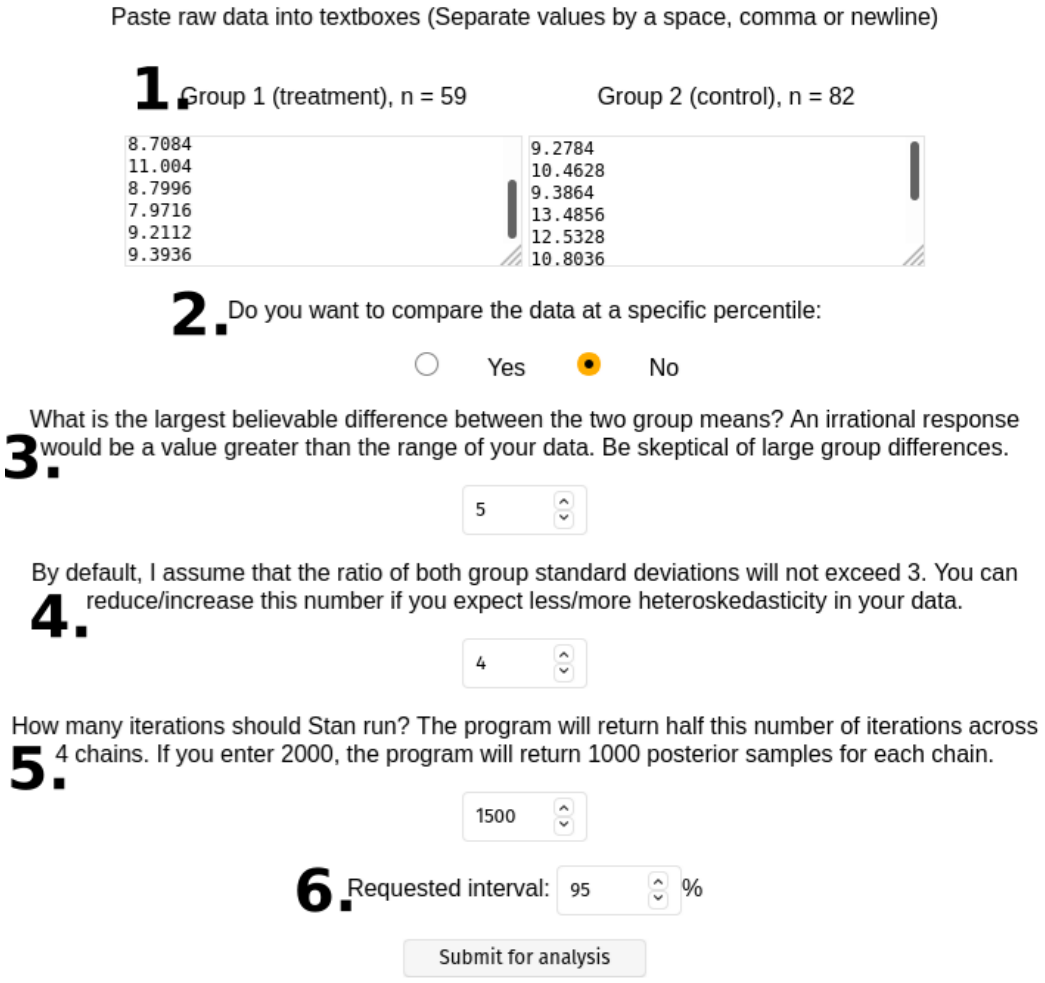

Figure 2. Series of steps to perform the t-regression on the faculty salary data. Selecting *no* in Step 2 (default) performs t-regression.

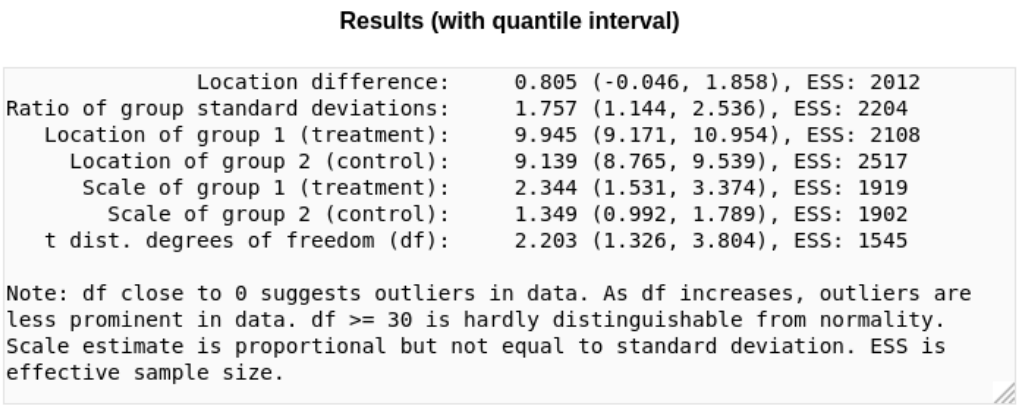

Figure 3. Model summary for t-regression results.

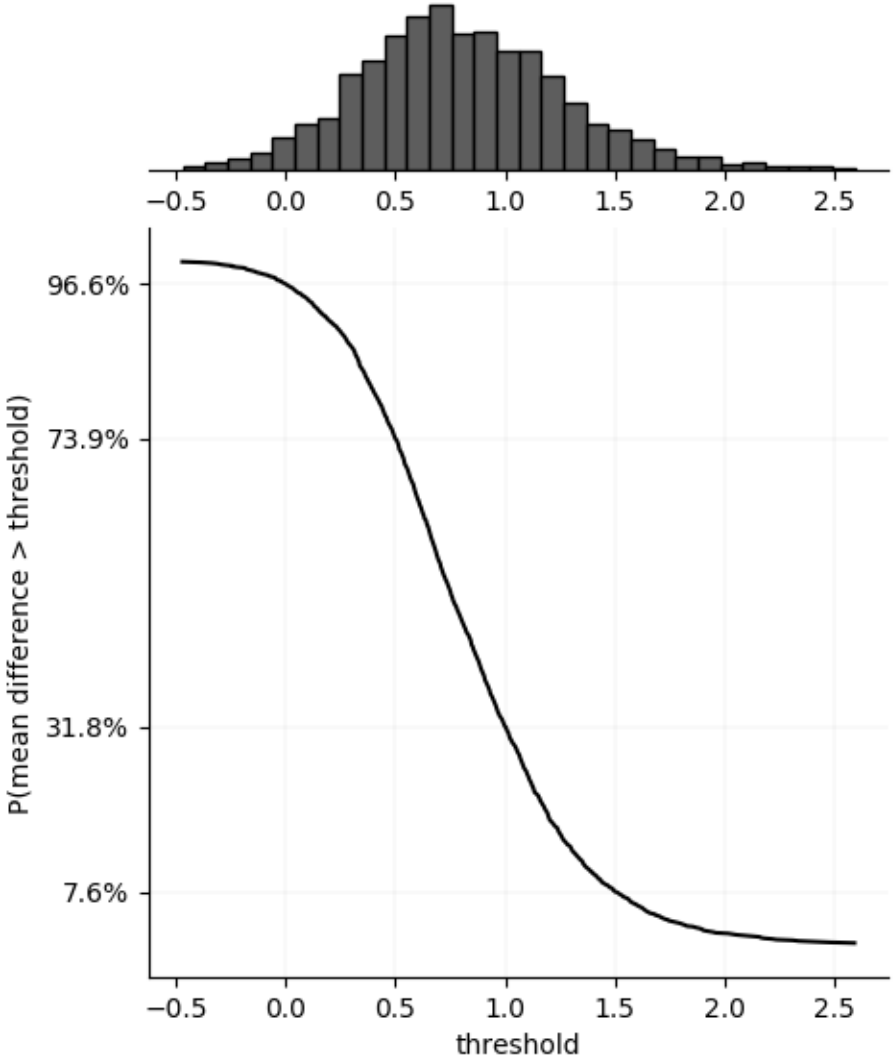

Figure 4. Using the posterior distribution to evaluate the average gender difference in salaries. From the histogram, it is clear that most posterior samples were greater than zero. From the line chart, one is able to evaluate the evidence supporting the salary difference at any threshold. For example, there was 96.6% chance that male salaries were on average higher than female salaries. However, the probability that the average salary difference exceeded US$10,000 (1.0 on x-axis) was lower (31.8%).

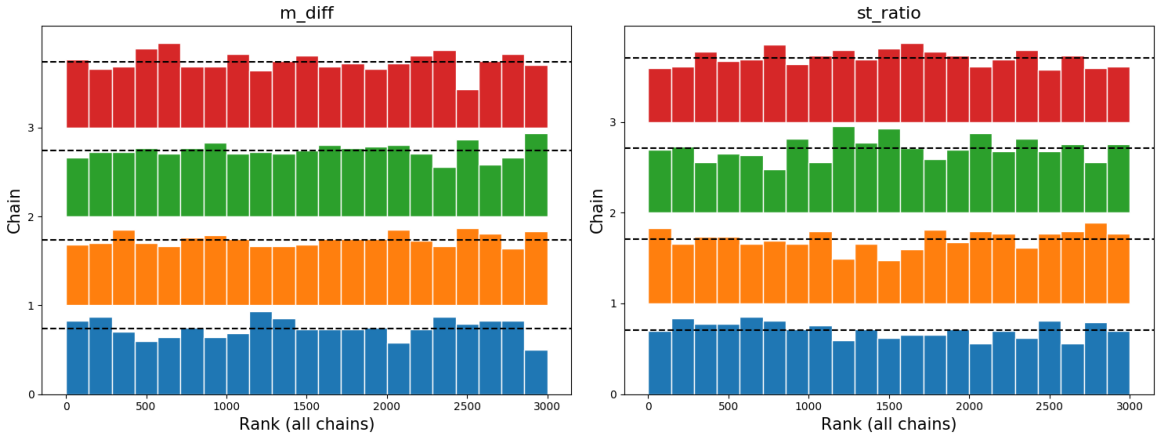

Figure 5. Rank plots of the mean difference and scale ratios. The posterior samples are rank transformed and plotted for each chain. If the different chains have the same target and have fully mixed, each histogram should be uniformly distributed. This plot is an improvement over trace plots. With trace plots, it is sometimes difficult to assess if chains have mixed as the chains overlap visually and a single chain might obscure the trajectory of other chains (Vehtari et al., 2019).

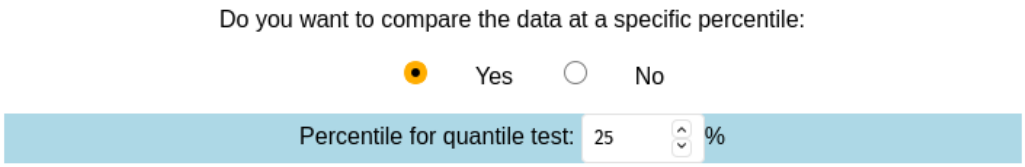

Figure 6. Modified step 2 in the earlier steps for t-regression and specified the desired percentiles to perform the quantile regression approach.

```
              Location difference:     2.894 (1.238, 4.764), ESS: 2071
Ratio of group standard deviations:     1.808 (1.306, 2.438), ESS: 2092
   Location of group 1 (treatment):     13.389 (11.873, 15.106), ESS: 2362
     Location of group 2 (control):     10.495 (9.920, 11.204), ESS: 2116
      Scale of group 1 (treatment):     1.612 (1.257, 2.064), ESS: 3373
        Scale of group 2 (control):     0.901 (0.733, 1.106), ESS: 1904
   t dist. degrees of freedom (df):
```

Figure 7. Model summary for results at the 75th percentile.

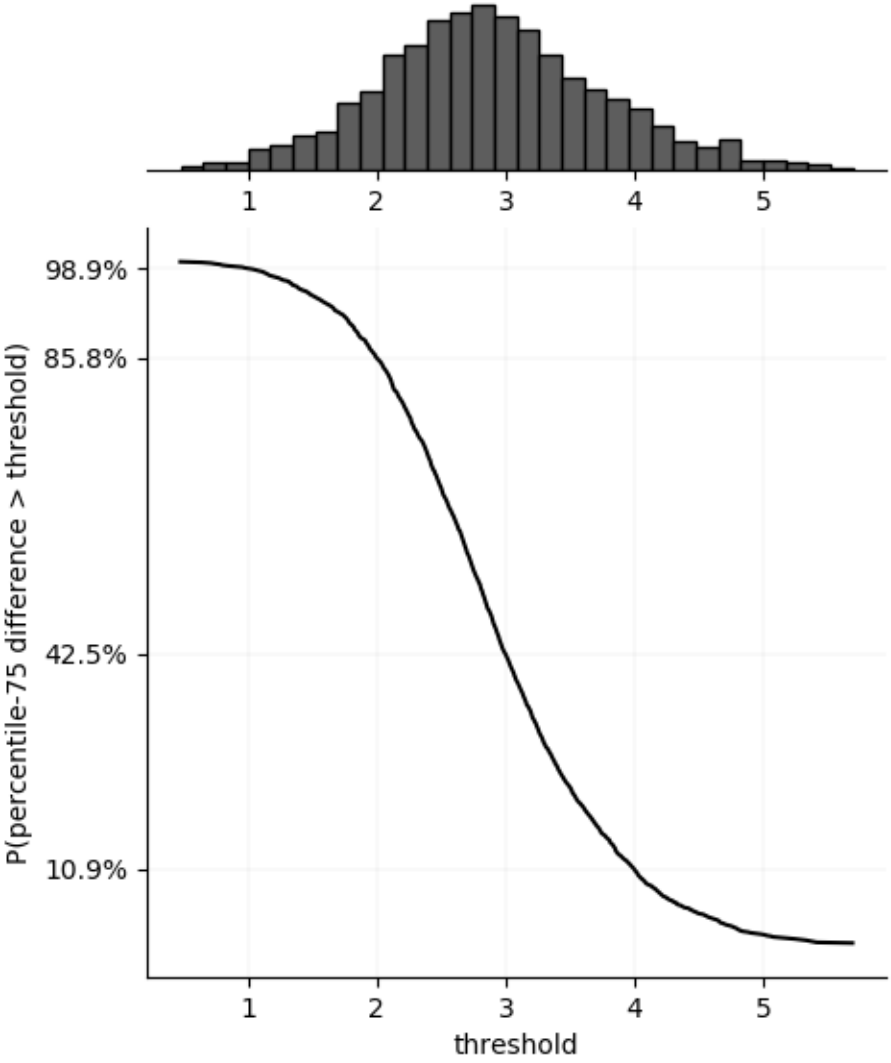

Figure 8. Using the posterior distribution to evaluate the gender difference in salaries at the 75th percentile. From the histogram, all posterior samples were greater than zero, hence we can be highly certain that 75th percentile of male salaries was higher than 75th percentile female salaries. Moreover, the probability that this difference exceeded US$10,000 was 98.9%.