Week 12 Exercises

Research Paper. Continue to make progress on your research paper. The next draft is due Tuesday of Thanksgiving week. This draft should be close to finalized, so that I can make suggestions to help polish the paper.

Workshop. Finalize your materials.

Rehearsals:

- ~~*Cox*. Monday, November 10 at 9am (to about 10am).~~
- *Unordered* and *ordered*. Friday, November 14 at 12pm (to about 2pm).

Workshops: I've solicited RIBC students and sent around the (still active) registration sheet.

- *Unordered*. Tuesday, November 18 from 1pm to 2pm.
- *Ordered*. Wednesday, November 19 from 1pm to 2pm.
- *Cox*. Friday, November 21 from 11am to 12noon.

Exercise 1 Simulate and recover; log-normal with censoring

Do the simulate and recover exercise for a log-normal duration model *with censoring*. You may use any `comparison()` you like.

> **Warning**
>
> For the log-normal, we have $\text{median}(y \mid X) = \exp(X\beta)$ and $E(y \mid X) = \exp\left(X\beta + \frac{\sigma^2}{2}\right)$. You can find these identities listed in the back of most probability textbooks.
> When writing the solution, I realized `predict()` returns the *median* of Y for `survreg(..., dist = "lognormal")`. This means that `predictions()` returns medians and `comparisons()` returns differences in medians.
> But there's also the {flexsurvreg} package. I tested it further and realized `predict()` returns the *mean* of Y for `flexsurvreg(..., dist = "lognormal")`. This means that `predictions()` returns means and `comparisons()` returns differences in means for

`flexsurvreg()` models.
This is an example of software doing things you don't realize or expect, but you can catch with a simulate and recover experiment.

```
## --- simulate a fake data set: log-normal duration model with censoring ---

# set sample size
n <- 1000  # large so that CI is not super wide

# create explanatory variables
## numeric variable
set.seed(123)
x1 <- rnorm(n, 0, 0.5)  # sd = 0.5 is my preferred scale for interpretation
## qualitative variable
x2 <- sample(c("Label 1", "Label 2"),
             size = n,
             replace = TRUE,
             prob = c(0.75, 0.25))  # about 75% will be "Label 1"

# create coefficients
b0 <- 0
b1 <- 0.6
b2 <- -0.4
sigma <- 0.4

# simulate event times from LogNormal with meanlog = eta and sdlog = sigma
eta <- b0 + b1 * x1 + b2 * (x2 == "Label 2")
t_unobs <- rlnorm(n, meanlog = eta, sdlog = sigma)


c <- 1 # censored here; about 25% will be censored

# observed time and censoring indicator
time <- ifelse(t_unobs > c, c, t_unobs)
status <- as.integer(t_unobs < c)  # 1 = event observed, 0 = censored

# combine into data frame
data <- data.frame(time, status, x1, x2) |>
  glimpse()
```

```
Rows: 1,000
Columns: 4
$ time   <dbl> 0.6086345, 0.8253431, 1.0000000, 1.0000000, 1.0000000, 1.000000~
$ status <int> 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, ~
$ x1     <dbl> -0.28023782, -0.11508874, 0.77935416, 0.03525420, 0.06464387, 0~
$ x2     <chr> "Label 1", "Label 1", "Label 1", "Label 1", "Label 1", "Label 1~
```

```r
# for log-normal
# - E(T | x) = mean(T | x) = exp(eta + sigma^2 / 2)
# - median(T | x) = exp(eta)

eta_hi <- b0 + b1 * max(x1) + b2 * 1  # x2 == "Label 2"
eta_lo <- b0 + b1 * min(x1) + b2 * 1

# medians
median_hi <- exp(eta_hi)
median_lo <- exp(eta_lo)

# means
mean_hi <- exp(eta_hi + (sigma^2) / 2)
mean_lo <- exp(eta_lo + (sigma^2) / 2)

# predict() returns medians (I learned!), so predictions() and comparisons() return medians
truth_fd <- median_hi - median_lo
truth_fd
```

```
[1] 1.483823
```

```r
# --- recover the quantity of interest ---

# load packages
library(survival)
library(flexsurv)
library(marginaleffects)

# fit log-normal AFT with censoring
fit <- survreg(Surv(time, status) ~ x1 + x2, data = data, dist = "lognormal")

summary(fit) # note that log(.4) = -0.9162907
```

```
Call:
survreg(formula = Surv(time, status) ~ x1 + x2, data = data,
    dist = "lognormal")
               Value Std. Error      z      p
(Intercept) -0.00304    0.01778  -0.17   0.86
x1           0.55779    0.03058  18.24 <2e-16
x2Label 2   -0.39787    0.03103 -12.82 <2e-16
Log(scale)  -0.93861    0.03168 -29.63 <2e-16

Scale= 0.391

Log Normal distribution
Loglik(model)= -284.5   Loglik(intercept only)= -501.5
    Chisq= 433.89 on 2 degrees of freedom, p= 6.1e-95
Number of Newton-Raphson Iterations: 5
n= 1000
```

```
# recover qis: these are medians!
predictions(fit,
            variables = list(x1 = "minmax"),
            newdata  = datagrid(x2 = "Label 2"))
```

```
      x2 Estimate Std. Error    z Pr(>|z|)     S 2.5 % 97.5 %
 Label 2    0.306    0.00899 34.0   <0.001 841.0 0.288  0.324
 Label 2    1.654    0.07073 23.4   <0.001 399.1 1.515  1.792

Type: response
```

```
comparisons(fit,
            variables = list(x1 = "minmax"),
            newdata  = datagrid(x2 = "Label 2"))
```

```
      x2 Estimate Std. Error    z Pr(>|z|)     S 2.5 % 97.5 %
 Label 2     1.35      0.072 18.7   <0.001 257.0  1.21   1.49

Term: x1
Type: response
Comparison: Max - Min
```

```r
# predict is returning median(Y) not E(Y) here
grid <- tibble(x1 = c(min(x1), max(x1)), x2 = "Label 2") |>
  glimpse()
```

```
Rows: 2
Columns: 2
$ x1 <dbl> -1.404887, 1.620520
$ x2 <chr> "Label 2", "Label 2"
```

```r
predict(fit, type = "response", newdata = grid)
```

```
       1        2
0.305887 1.653676
```

```r
median_hi; median_lo
```

```
[1] 1.77236
```

```
[1] 0.2885369
```

```r
mean_hi; mean_lo
```

```
[1] 1.919975
```

```
[1] 0.3125683
```

```r
# try with flexsurv
fit_fs <- flexsurvreg(Surv(time, status) ~ x1 + x2, data = data, dist = "lnorm")
predict(fit_fs, type = "response", newdata = grid)
```

```
# A tibble: 2 x 1
  .pred_time
       <dbl>
1      0.330
2      1.79
```

Exercise 2 Claims and CIs

Read Rainey (2014) and McCaskey and Rainey (2015). These papers describe in detail my advice to "only make a claim if the claim holds for the *entire* confidence interval."

Give two examples (see below) of how current practice deviates from this advice. Explain a how my advice applies to these two situations. Explain why these deviations from my advice matter (or not).

Examples:

1. McCaskey and Rainey (2015): Claiming "substantively significant" if the point estimate is meaningful (and statistically significant).
2. Rainey (2014): Claiming "no effect" when an estimate is not statistically significant.

Exercise 3 Power

Background

Suppose you are conducting a simple, balanced treatment-control experiment. You know that the SD of the outcome will be about $\widetilde{SD}$. Rainey (2025) shows that the SE of the estimated effect will be about $\frac{2 \cdot \widetilde{SD}}{\sqrt{2 \cdot n}}$, where n is the number of participant *per condition*. Further, the minimum detectable effect with 95% power is $3.3 \cdot SE$.

Details of the Experiment

- Your outcome is affective polarization, measured as a 101-point feeling thermometer directed at the out-party ("Republicans" or "Democrats"). This outcome typically has an SD of about 25 in surveys like the ANES or those conducted on crowd-sourced platforms like MTurk.
- Your treatment a two-minute video showing Republicans and Democrats enjoying a beer together.

If you have a simple, balanced treatment-control experiment that you'd rather write about, feel free to replace the details above with the analogous information.

Questions

To answer these question, you can refer to Rainey (2025) if helpful–there are lots of useful rules there. But the *Background* above is sufficient!

1. For the treatment and outcome above, what is the smallest treatment effect that is substantively meaningful. Alternatively, what is the largest treatment effect that is substantively trivial. Come up with a precise number on the 101-point feeling thermometer scale.[1]
2. For samples of size 100, 200, 500, and 2,000 *per condition*, what is the minimum detectable effect with 95% power?
3. Suppose you are applying for grant funding for this experiment. Respondents cost $2.00 each. Write a paragraph (to be included in the grant) justifying a specific number of respondents. Describe and justify the treatment effects you are targeting (i.e., the smallest substantively meaningful effect), the power target (e.g., 80%, 90%, or 95%), and the sample size to obtain that power.

Solution.

Part 1.

I chose $m = 3$ on the 101-point feeling thermometer scale as my smallest effect of substantive interest.

Part 2.

```
sd_tilde <- 25
n <- c(100, 200, 500, 2000, 1500)  # per condition
se <- 2*sd_tilde/sqrt(2*n)
mde <- 3.3*se

cbind(n, se, mde)
```

```
        n        se       mde
[1,]  100 3.5355339 11.667262
[2,]  200 2.5000000  8.250000
[3,]  500 1.5811388  5.217758
[4,] 2000 0.7905694  2.608879
[5,] 1500 0.9128709  3.012474
```

Part 3.

A sample size of 3,000 respondents (1,500 per condition) gives us a well-powered experiment. In recent pilot data, the out-party feeling thermometer has a standard deviation of about 25

[1] For example, a 5-point treatment effect might mean that the average in the treatment group is 55 and the average in the control group is 50.

points. Using the rules from Rainey (2025), this implies a minimum detectable effect with 95% power of $3.3 \cdot \frac{2 \times 25}{\sqrt{2 \times 1,500}} \approx 3$. An effect of about 3 points roughly corresponds to a "small" substantive effect. For context, Lovakov and Agadullina (2021) show that about 25% of effects in social psychology are smaller than about 15% of a standard deviation. We are well-powered to detect and effect of about $\frac{3}{25} = 12\%$ of a standard deviation.

Exercise 4 Literatures, Part 1

Read Gelman and Carlin (2014). Use the `retrodesign()` function in the {retrodesign} to compute the power, Type S error, and Type M errors for the designs above (i.e., for the various sample sizes above). See `?retrodesign`. Intepret the values.

An example is below. Note that `retrodesign()` does not accept a vector for `s`, so you need to be creative to compute across multiple sample sizes where `s` is changing.[2]

```
library(retrodesign)

n <- 500  # per condition
sd_tilde <- 25
se <- 2*sd_tilde/sqrt(2*n)
retrodesign(2.0, s = se, alpha = 0.10)
```

```
$power
[1] 0.3538025

$type_s
[1] 0.005111622

$type_m
[1] 1.831827
```

Solution.

[2] You can use a for-loop or `Vectorize()`.

```r
# load packages
library(retrodesign)
library(tidyverse)
library(tinytable)

# parameters
n <- c(100, 200, 500, 2000)  # per condition
sd_tilde <- 25
se <- 2*sd_tilde/sqrt(2*n)

# put into a tibble and run retrodesign() the tidy way
rd_tbl <- tibble(n = n,
                 sd_tilde = sd_tilde,
                 se = se) |>
  mutate(rd = map(se, ~ retrodesign(2.0, s = .x, alpha = 0.10))) |>
  unnest_wider(rd) |>
  rename(`Assumed SD` = sd_tilde,
         SE = se,
         Power = power,
         `Type S` = type_s,
         `Type M` = type_m) |>
  glimpse()
```

```
Rows: 4
Columns: 6
$ n            <dbl> 100, 200, 500, 2000
$ `Assumed SD` <dbl> 25, 25, 25, 25
$ SE           <dbl> 3.5355339, 2.5000000, 1.5811388, 0.7905694
$ Power        <dbl> 0.1537903, 0.2063418, 0.3538025, 0.8119281
$ `Type S`     <dbl> 8.800223e-02, 3.511429e-02, 5.111622e-03, 1.837667e-05
$ `Type M`     <dbl> 3.789970, 2.726906, 1.824342, 1.137455
```

```r
# print table
tt(rd_tbl, digits = 3)
```

Exercise 5 Literatures, Part 1

Read Arel-Bundock et al. (2025).

n	Assumed SD	SE	Power	Type S	Type M
100	25	3.536	0.154	0.0880022	3.79
200	25	2.5	0.206	0.0351143	2.73
500	25	1.581	0.354	0.0051116	1.82
2000	25	0.791	0.812	0.0000184	1.14

Question 1

Arel-Bundock et al. write that the median (observed) power in political science is about 10%. The authors of the paper suggest that this problematic. Why? Do you agree?

Hint: Think about the difference between *observed* or *actual* power that Arel-Bundock et al. (2025) focus on and power to detect the smallest effect of interest. These are rarely the same.[3] Which should we care about: power to detect the *actual effect* or power to detect the *smallest effect of interest*?

Question 2

You are an editor of a journal and have received two thoughtful reviews for a well-done survey experiment. It's time to make a decision on this paper.

Previous work suggests that a treatment should have a positive effect, but this paper suggests that the treatment has a *negative* effect. The authors' estimate is negative and statistically significant.

- Reviewer A says that the experiment is otherwise well-done, but claims that the sample size is too small. They cite Gelman and Carlin (2014) and Arel-Bundock et al. (2025) and suggest that the journal shouldn't publish underpowered work.
- Reviewer B also observed that the experiment is underpowered, but claims that this isn't a problem because the authors' test protects them against Type I errors as usual.[4]

1. Explain each perspective more fully.
2. How do you adjudicate between these perspectives? Describe the tradeoffs between a policy of (1) requiring that all papers be well-powered (Reviewer A's stance) and (2) not considering the power once data have been collected (Reviewer B's stance). Which of these policies would you advocate for?

[3]For example, let's say that we clearly identify the smallest effect of interest m and conduct an experiment that's well-powered to detect an effect of size m. But if the actual effect is much smaller than m, then our experiment has an actual power closer to 5%.

[4]Remember that the Type I error rate is *at most* 5% when the null is true, so the authors are protected in the sense that $\Pr(\text{make claim} \mid \text{claim is false}) \leq 0.05$. Recall that power relates to Type II errors, which the authors definitely haven't made here, because they are rejecting the null.

3. Years later, you are no longer an editor, but have accumulated some formal and informal power in the discipline. You decide to return to this issue. What norms, practices, or rules might you change to address this issue?

Exercise 6 Application

Choose a paper that you've read recently that estimates a causal effect (or a descriptive difference). *For the sake of this exercise, an experiment with a simple design works best, but any substantively interpretable estimate will work.*

Note the estimate (a particular number, like 0.73), its standard error (also a particular number, like 0.22),[5], and their interpretation[6] (e.g., "substantively large").

1. First, construct the bins of negligible and meaningful effects. In most cases, authors don't offer these directly, so you'll need to fill in any gaps.
2. Second, multiply the standard error by 2.5. This is the effect that the design can detect with 80% power.[7] Would you consider this well-powered? Use `retrodesign()` to compute the Type M and Type S errors using the authors' estimated SE and the smallest substantively meaningful effect you identified in Part 1. Interpret.
3. Third, compare the authors interpretation to their 90% confidence interval.

    a. Do they make claims that do not hold for the entire 90% CI? (i.e., Do they make claims they shouldn't?)
    b. Do they *not* make claims that *do* hold for the entire 90% CI? (i.e., Do they not make claims they could?)

References

Arel-Bundock, Vincent, Ryan C. Briggs, Hristos Doucouliagos, Marco M. Aviña, and T. D. Stanley. 2025. "Quantitative Political Science Research Is Greatly Underpowered." *The Journal of Politics*, September, 000–000. https://doi.org/10.1086/734279.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations." *Perspectives on Psychological Science* 9 (6): 641–51. https://doi.org/10.1177/1745691614551642.

McCaskey, Kelly, and Carlisle Rainey. 2015. "Substantive Importance and the Veil of Statistical Significance." *Statistics, Politics and Policy* 6 (1-2). https://doi.org/10.1515/spp-2015-0001.

[5] Sometimes you have to reverse engineer the SE from the CI or p-value.

[6] The interpretation might be sparse or absent. Or it might be implicit, like "statistically significant," which I usually read as "greater than zero" or "less than zero."

[7] Multiplying by 3.3 give the effect that the design can detect with 95% power.

Rainey, Carlisle. 2014. "Arguing for a Negligible Effect." *American Journal of Political Science* 58 (4): 1083–91. https://doi.org/10.1111/ajps.12102.

———. 2025. "Power Rules: Practical Advice for Computing Power (and Automating with Pilot Data)." http://dx.doi.org/10.31219/osf.io/5am9q_v2.