Week 2 Exercises

Exercise 1 Bernoulli grid search

Suppose you design a Bernoulli experiment that generates successes or failures with an unknown probability π. You want to estimate π, so you run the experiment three times and get the outcomes `y <- c(0, 1, 0)`, where `0` is a failure and `1` is a success.

Use ML to estimate `pi`. *But don't find the maximum analytically* or with a hill-climbing algorithm. Instead, use a *grid search* to find the maximum. Use `seq()` to create a range of ten to twenty candidate values for π, compute the log-likelihood for each candidate value, and locate the candidate value that produces the largest log-likelihood. Report your results in figure and a table.

Exercise 2 Equivalence of numerical and closed-form solutions

We found that the sample average is the ML estimate of the parameter λ of the Poisson distribution. For the data set `y <- c(12,  7,  9, 12, 10)` show that optimizing the Poisson log-likelihood function with `optim()` produces the same answer and the closed-form solution.

Exercise 3

DeGroot and Schervish, q. 9, p. 425. Suppose a distribution $f(x;\theta) = \theta x^{\theta-1}$ for $0 < x < 1$ and $\theta > 0$. Find the ML estimator of θ.

Exercise 4

Suppose you have a binary outcome `y <- c(0, 1, 0, 1, 1, 1, 0)`. Suppose you want to estimate the *mean* of this distribution. You'd normally model the binary outcome with a Bernoulli distribution and estimate the parameter Bernoulli parameter π with ML. But instead

of using the Bernoulli distribution, you model these data with a normal distribution. What would your estimate of the mean be? What if you used the Poisson? Exponential? Explain.

No matter what distribution we use, we obtain the *identical* estimate. Not the same in expectation (over a large number of imaginary repetitions). Not the same asymptotically (as the sample size grows large). The exact same number every time. This mean that any property that the Bernoulli estimator has, the others have as well. If the Bernoulli estimator is consistent, then the other three estimators are consistent as well. It's also worth noting that all three are unbiased estimators as well.

This illustrates an important point. The quality of an estimator doesn't always depend on the correctness of all parts of the model. It often depends on what quantity of interest you are targeting.

Exercise 5

Let $\widehat{\theta}$ be the ML estimate of θ. Suppose we are interested in estimating $\psi = \theta^2$. What is the ML estimate of ψ?

Exercise 6 Uniform distribution

Suppose a discrete uniform distribution from 0 to K. The pdf is $f(x; K) = \frac{1}{K}$ for $x \in \{0, 1, ..., K\}$. Suppose I have three samples from the distribution: 276, 159, and 912.

a. Find the ML estimate of K. *Hint: The log-likelihood is discontinuous, so the usually optimization routine might mislead you. But the maximum is immediately apparent once you write out the likelihood.*
b. Find the method of moments estimate of K. *Hint: The mean of this uniform distribution is* $\frac{K}{2}$.
c. Discuss any problems you notice with each estimator.

Exercise 7 Exponential model

The exponential distribution has pdf $f(t; \lambda) = \lambda e^{-\lambda t}$ for $t \geq 0$ and $\lambda > 0$. We sometimes use this distribution to model time spells t, such as the time until an event or the time between events. For example, some political scientists are interested in how long a government lasts after a government formation in a parliamentary system. We might use an exponential distribution to model this time or "duration."

a. Show that the cdf of the exponential distribution is $F(t;\lambda) = \Pr(T \leq t;\lambda) = 1 - e^{-\lambda t}$. Use the cdf to find the survival function $S(t;\lambda) = \Pr(T > t;\lambda) = 1 - F(t;\lambda)$. How can we interpret the survival function (i.e., for input t, what does the survival function return)?
b. Find $\Pr(T > t + s \mid T > s)$ using the survival function and the definition of conditional probability. Compare $P(T > t)$ to $P(T > t + s \mid T > s)$. What do you notice? Interpret the result. Using a specific political outcome as a concrete example (e.g., time between major protests, government durations, time between constitutional amendments), what does this property say about the time until an event occurs, given that it has been some given time since an event occurred?
c. Suppose we collect N random samples $t = \{t_1, t_2, ..., t_N\}$ and model each draw as an exponential random variable random. Find the ML estimator of λ.
d. Find an ML estimator for the mean, which is $\frac{1}{\lambda}$.

Optional: Show that the mean of the exponential distribution is $\frac{1}{\lambda}$. This requires integration by parts.

Exercise 8 Simulating memorylessness in R

Use R to simulate 10,000 draws from an exponential distribution with rate $\lambda = 1$. Each of the 10,000 values represents a duration (i.e., time until an event occurs). For

```
set.seed(123)
durations <- rexp(10000, rate = 1)
```

For durations larger than five, compute the *remaining* duration.

```
remaining_durations <- durations[durations > 5] - 5
```

Use a histogram or plots of the ECDF to compare the distribution of `remaining_durations` to the original `durations`. Explain the similarity, why it's expected, and why it's important.

Exercise 9 The `faithful` Data Set

The `faithful` dataset in base R contains 272 observations of `eruptions` (eruption time in minutes) and `waiting` (waiting time to next eruption in minutes) for the Old Faithful geyser in Yellowstone National Park. See `?faithful` for more details.

```
glimpse(faithful)
```

```
Rows: 272
Columns: 2
$ eruptions <dbl> 3.600, 1.800, 3.333, 2.283, 4.533, 2.883, 4.700, 3.600, 1.95~
$ waiting   <dbl> 79, 54, 74, 62, 85, 55, 88, 85, 51, 85, 54, 84, 78, 47, 83, ~
```

Model `eruptions` as an exponential distribution. Estimate the rate and mean. Use the predictive distribution to evaluate the fit of the exponential model to these data.

Optional. Repeat for `waiting`.

Exercise 10 Herron's hockey data set

Herron's hockey data set contains data on the time between hits[1] in the regulation periods of the 82 regular season games for the Chicago Blackhawks. The data are available via GitHub Gist here. You can load the data directly from the web, but you need to click "Raw" and copy the URL for the raw CSV data. Model `seconds_btw_hits` as an exponential distribution. Estimate the rate and mean. Use the predictive distribution to evaluate the fit of the exponential model to these data. Discuss whether you find the fit surprising and explain why.

```
# load data directly from web
hockey <- read_csv("https://gist.githubusercontent.com/carlislerainey/0bc3018cd2377022fd045e

# quick look
glimpse(hockey)
```

```
Rows: 1,175
Columns: 5
$ game_id                 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, ~
$ period_id               <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 1, 1, 1, 1, 2, ~
$ time_of_hit             <chr> "25S", "2M 8S", "2M 58S", "3M 8S", "4M 13S", ~
$ seconds_played_in_season <dbl> 25, 128, 178, 188, 253, 560, 643, 1407, 1825,~
$ seconds_btw_hits        <dbl> NA, 103, 50, 10, 65, 307, 83, 764, 418, 535, ~
```

[1] From Wikipedia, a hit is defined as: *"Intentionally initiated contact with the player possessing the puck that causes that player to lose possession of the puck. Loss of possession may or may not involve a turnover. If the contact results in a penalty, no hit is awarded."*

Exercise 11 The location-scale t distribution

Some Theory

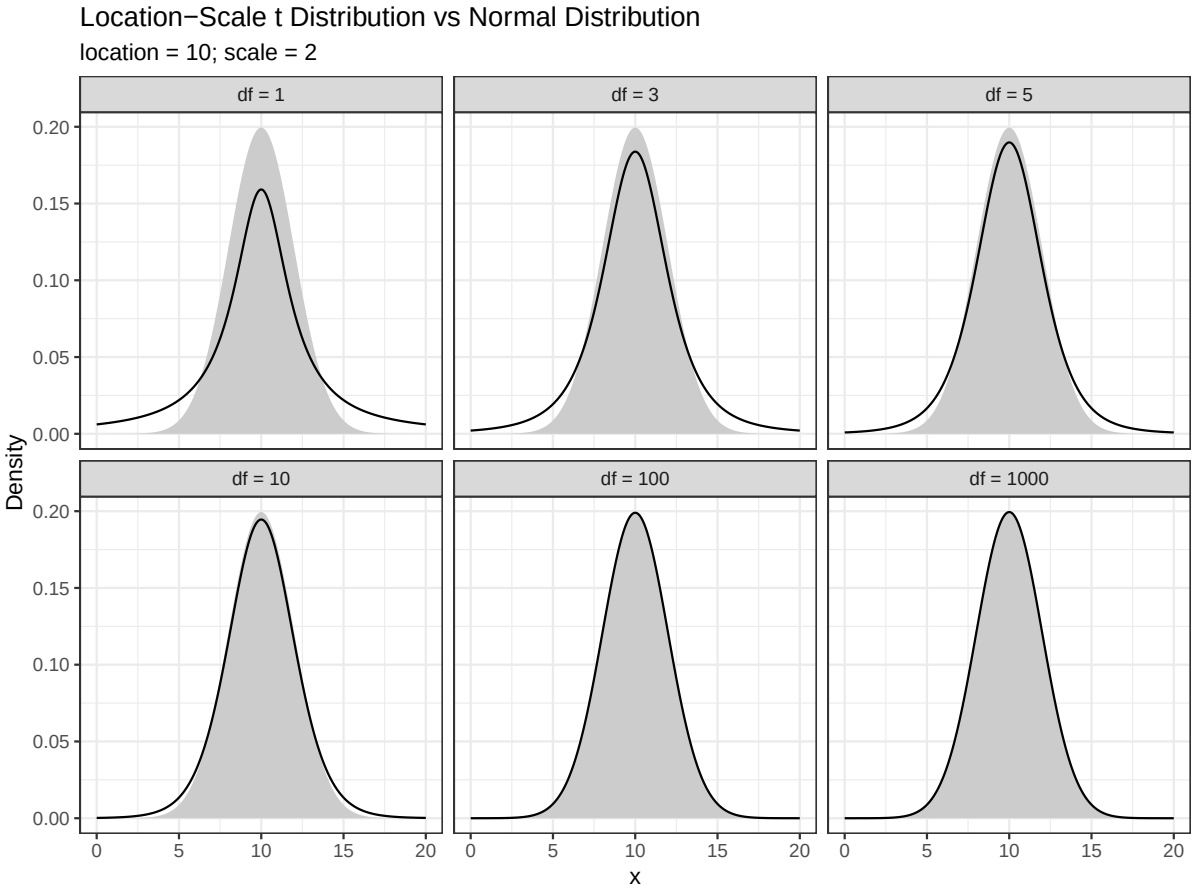
The location-scale t distribution is a flexible model for data that may exhibit heavier-than-normal tails. This is can be helpful when your data have more extreme observations than expected with a normal model.

The pdf of the location-scale t distribution is

$$f(y \mid \mu, \sigma, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\,\sigma}\left[1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}},$$

where $\Gamma(\cdot)$ is the gamma function, μ is the location parameter (which shifts the distribution much like the mean of the normal distribution), σ is the scale parameter (which changes the spread of the distribution much like the SD parameter of the normal distribution), and ν controls the heaviness of the tails—a smaller value of ν produces a heavier tails and the distribution converges to the normal distribution as $\nu \to \infty$. For $\nu > 10$, the t and normal distributions are very similar. We refer to the special case of $\nu = 1$ as the Cauchy distribution.

The figure below compare the location-scale t and normal distributions. As x moves away from the center of the distribution, notice that that the t density remains well above zero for *much* longer when the df is less than five or so.



These heavy tails make the location-scale t distribution useful for robust modeling, where we want to capture central tendencies while being less sensitive to unusual values.

In R, we have the `metRology::dt.scaled()` function that computes the density or likelihood. Some important things to note about this function: the argument `mean` refers to the location parameter μ, though μ isn't always a mean. The argument `sd` refers to the scale parameter σ, though σ isn't the SD. The argument `df` refers to the parameter ν.

Some Questions

a. As a baseline, model the percentage GDP growth in the data below as using a normal model. Use the predictive distribution to assess the fit. What does the normal model miss?
b. As an alternative, try a location-scale t model. Use `optim()` to estimate the parameters μ, σ, and ν. Compare the predictive distribution for the t model to the normal model. (The `metRology::dt.scaled(..., log = TRUE)` function will compute the log-likelihood for the individual observations, and `metRology::rt.scaled()` will draw samples.)
c. What is the ML estimate of the degrees of freedom parameter? Discuss it's importance. What patterns can the t distribution capture that the normal distribution cannot?

Exercise 12 Corrupted data

For the WDI data above, corrupt the data by replacing one observation with a data entry error (something like `pct_gdp_growth` of 10,000 and re-fit the normal and t models. How did the estimates of the *location* of each change? Why? Is this a desirable property?

Exercise 13 *Optional*: R function to simulate the predictive distribution

Note: This is a challenging function to write, so only attempt this problem if you are comfortable with writing other, simpler functions.

We've use my code to simulate the predictive distribution a few times and combine them together with the observed data set really long and repetitive. In his excellent *R for Data Science 2(3)*, Hadley Wickham writes: "A good rule of thumb is to consider writing a function whenever you've copied and pasted a block of code more than twice (i.e. you now have three copies of the same code)."

The inputs might be (1), the name of the outcome variable, (2) an observed data set, (3) a *function* to simulate the fake data sets, and (4) the number of fake data sets to generate.

```
sim_fake <- function(variable, observed, sim_fn, n = 5) {
  ...
}
```

Exercise 14 Reflection

The exercises above ask you to compare fitted distributions to observed data. In some cases, we saw close correspondence between the models and the data. In other cases, the two diverged substantially. You may have noticed that even the poorly fitting distributions tend to have the same mean as the observed data. Is it important that our models mimic the other features of the data? To what extent is the mean the most important feature (or the only important feature) of the distribution? When and why might we care about the other features substantively (e.g., SD, heavy tails, memorylessness)? Why might we care about these other features statistically?