STA 360/602L: Module 2.2

Operationalizing data analysis; selecting priors

Dr. Olanrewaju Michael Akande

Outline

- Operationalizing data analysis
- Example: rare events
- Selecting priors and potential problems

Operationalizing data analysis

How should we approach data analysis in general?

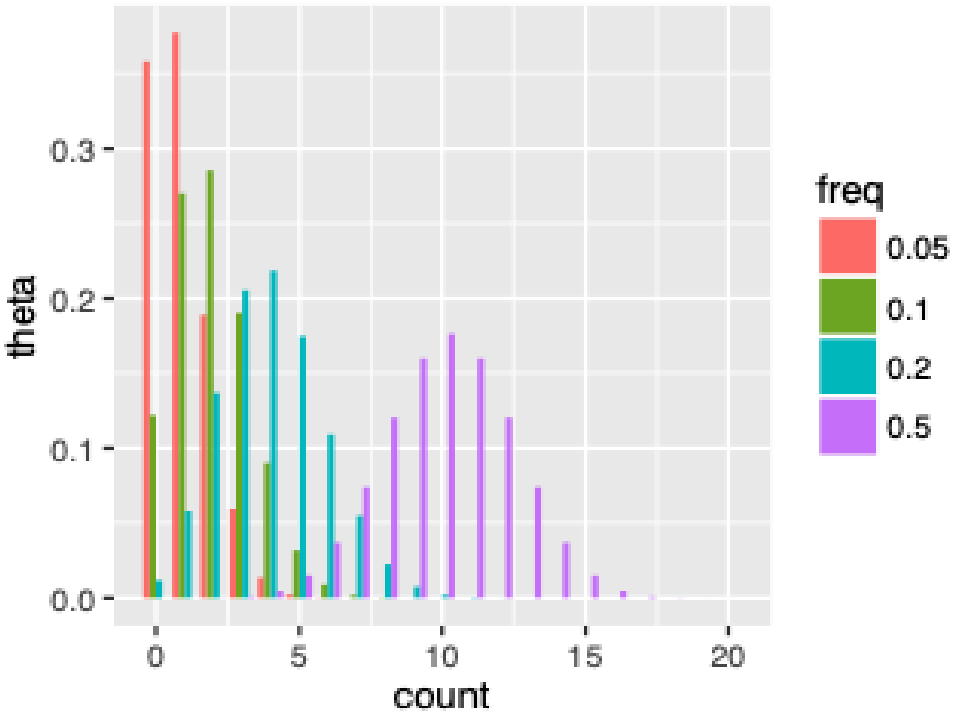
- Step 1. State the question.
- Step 2. Collect the data.
- Step 3. Explore the data.
- Step 4. Formulate and state a modeling framework.
- Step 5. Check your models.
- Step 6. Answer the question.

Example: rare events

- Step 1. State the question:
  - What is the prevalence of an infectious disease in a small city?
  - Why? High prevalence means more public health precautions are recommended.
- Step 2. Collect the data:
  - Suppose you collect a small random sample of 20 individuals.
- Step 3. Explore the data:
  - Let Y denote the unknown number of infected individuals in the sample.

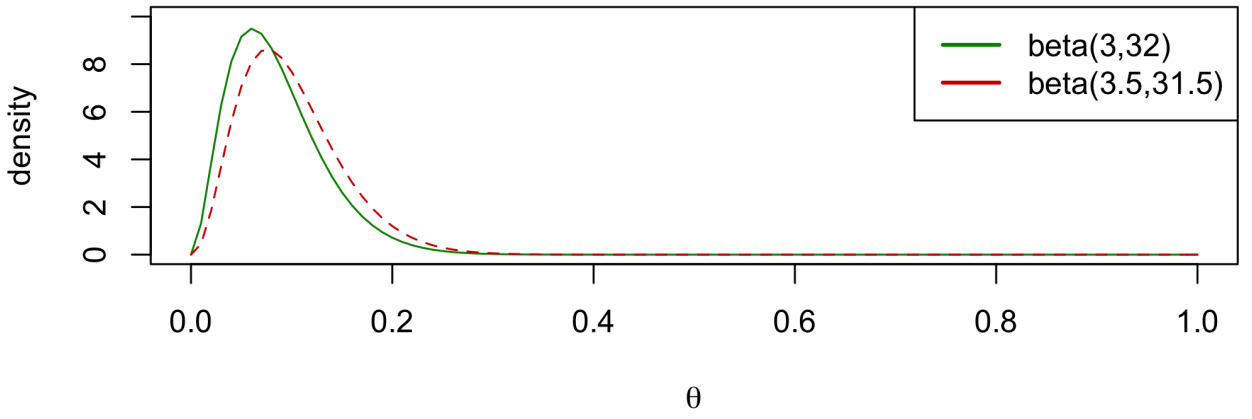
Example: rare events

- Step 4. Formulate and state a modeling framework:
  - Parameter of interest: θ is the fraction of infected individuals in the city.
  - Sampling model: a reasonable model for Y can be $\text{Bin}(20, \theta)$



Example: rare events

- Step 4. Formulate and state a modeling framework:
  - Prior specification: information from previous studies – infection rate in "comparable cities" ranges from 0.05 to 0.20 with an average of 0.10. So maybe a standard deviation of roughly 0.05?
  - What is a good prior? The **expected value** of θ close to 0.10 and the **standard deviation** close to 0.05.
  - Possible option: $\text{Beta}(3.5, 31.5)$ or maybe even $\text{Beta}(3, 32)$?



Quick beta-binomial recap

- Binomial likelihood:

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

- $+$ Beta Prior:

$$\pi(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} = \text{Beta}(a,b)$$

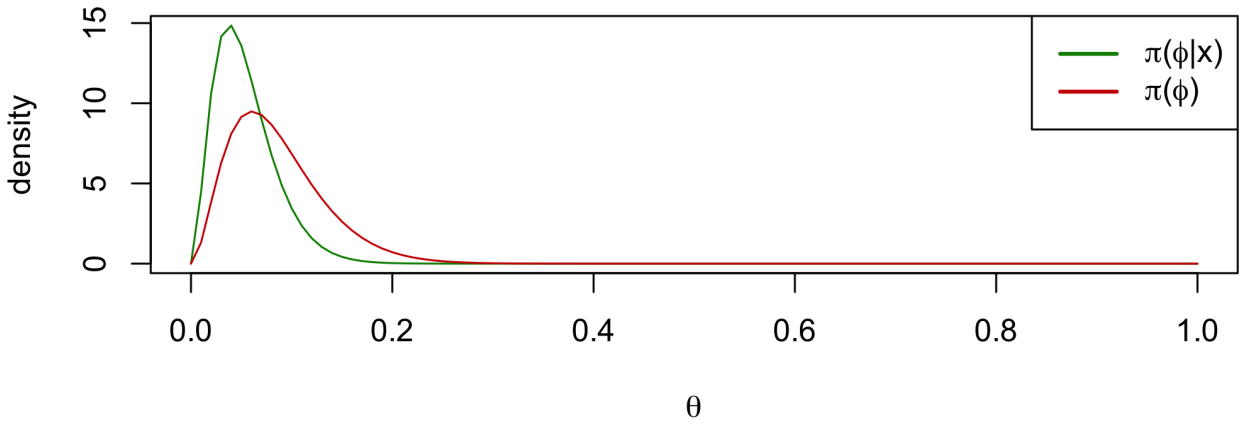
- $\Rightarrow$ Beta posterior:

$$\pi(\theta|y) = \frac{1}{B(a+y,b+n-y)}\theta^{a+y-1}(1-\theta)^{b+n-y-1} = \text{Beta}(a+y,b+n-y).$$

- Recall: If $\theta \sim \text{Beta}(a,b)$, then
  - $\mathbb{E}[\theta] = \frac{a}{a+b}$
  - $\mathbb{V}[\theta] = \frac{ab}{(a+b)^2(a+b+1)}$

Example: rare events

- Step 4. Formulate and state a modeling framework:
  - Under $\text{Beta}(3, 32)$, $\Pr(\theta < 0.1) \approx 0.67$.
  - Posterior distribution for the model:
    $\pi(\theta|Y = y) = \text{Beta}(a + y, b + n - y)$
  - Suppose $Y = 0$. Then, $\pi(\theta|Y = y) = \text{Beta}(3, 32 + 20)$



Example: rare events

- Step 5. Check your models:
  - Compare performance of posterior mean and posterior probability that $\theta < 0.1$.
  - Under $\text{Beta}(3, 52)$,
    - $\Pr(\theta < 0.1|Y = y) \approx 0.92$. More confidence in low values of θ.
    - For $\mathbb{E}(\theta|Y = y)$, we have

$$\mathbb{E}(\theta|y) = \frac{a+y}{a+b+n} = \frac{3}{52} = 0.058.$$

    - Recall that the prior mean is $a/(a+b) = 0.09$. Thus, we can see how that contributes to the prior mean.

$$\begin{aligned}\mathbb{E}(\theta|y) &= \frac{a+b}{a+b+n} \times \text{prior mean} + \frac{n}{a+b+n} \times \text{sample mean} \\ &= \frac{a+b}{a+b+n} \times \frac{a}{a+b} + \frac{n}{a+b+n} \times \frac{y}{n} \\ &= \frac{35}{52} \times \frac{3}{35} + \frac{20}{52} \times \frac{0}{n} = \frac{3}{52} = 0.058.\end{aligned}$$

Example: rare events

- Step 6. Answer the question:
  - People with low prior expectations are generally at least 90% certain that the infection rate is below 0.10.
  - $\pi(\theta|Y)$ is to the left of $\pi(\theta)$ because the observation $Y = 0$ provides evidence of a low value of θ.
  - $\pi(\theta|Y)$ is more peaked than $\pi(\theta)$ because it combines information and so contains more information than $\pi(\theta)$ alone.
  - The posterior expectation is 0.058.
  - The posterior mode is 0.04.
    - Note, for $\text{Beta}(a, b)$, the mode is $(a - 1)/(a + b - 2)$.
  - The posterior probability that $\theta < 0.1$ is 0.92.

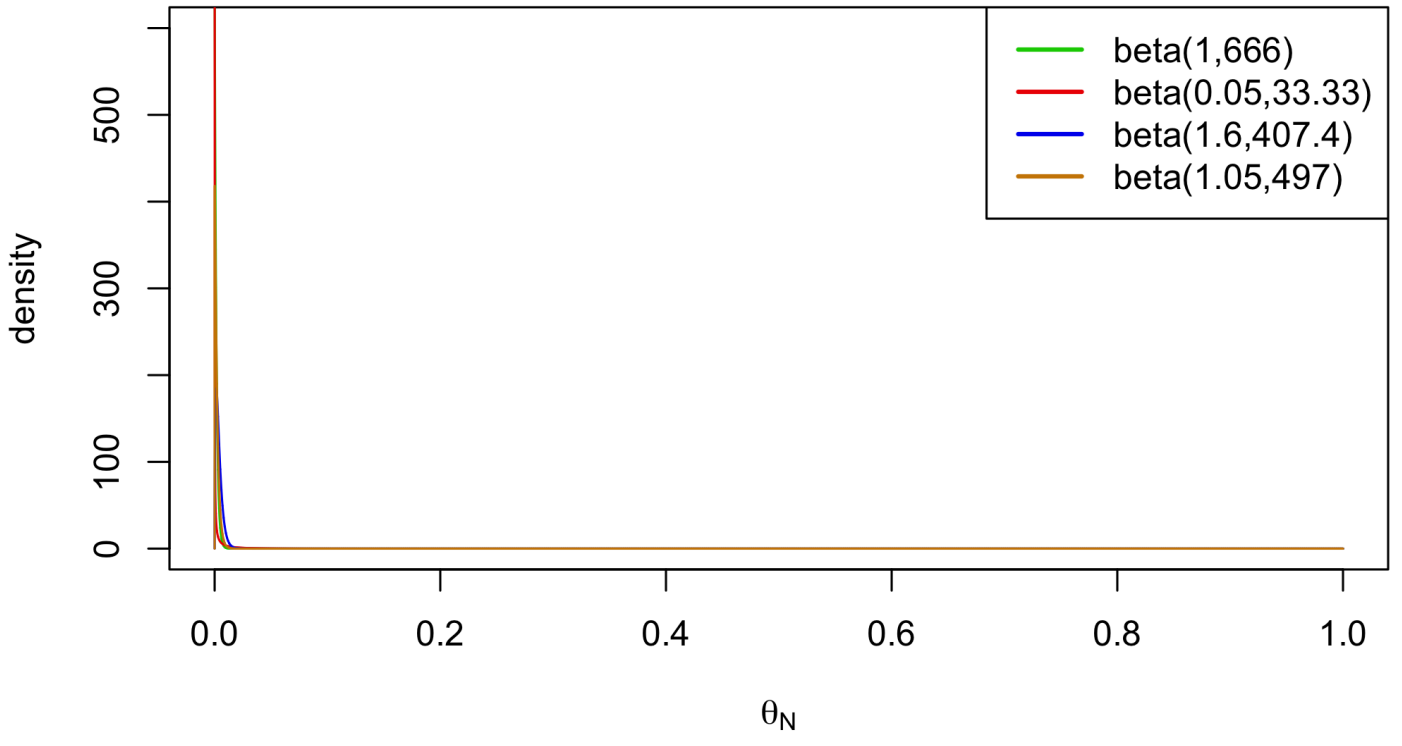
Cautionary tale: parameters at the boundary

- Tuyl et al. (2008) discuss potential dangers of using priors that have $a < 1$ with data that are all 0's (or $b < 1$ with all 1's). They consider data on adverse reactions to a new radiological contrast agent.
- Let θ_N: probability of a bad reaction using the new agent.
- Current standard agent causes bad reactions about 15 times in 10000, so one might think 0.0015 is a good guess for θ_N.
- How do we choose a prior distribution?

Potential prior distributions

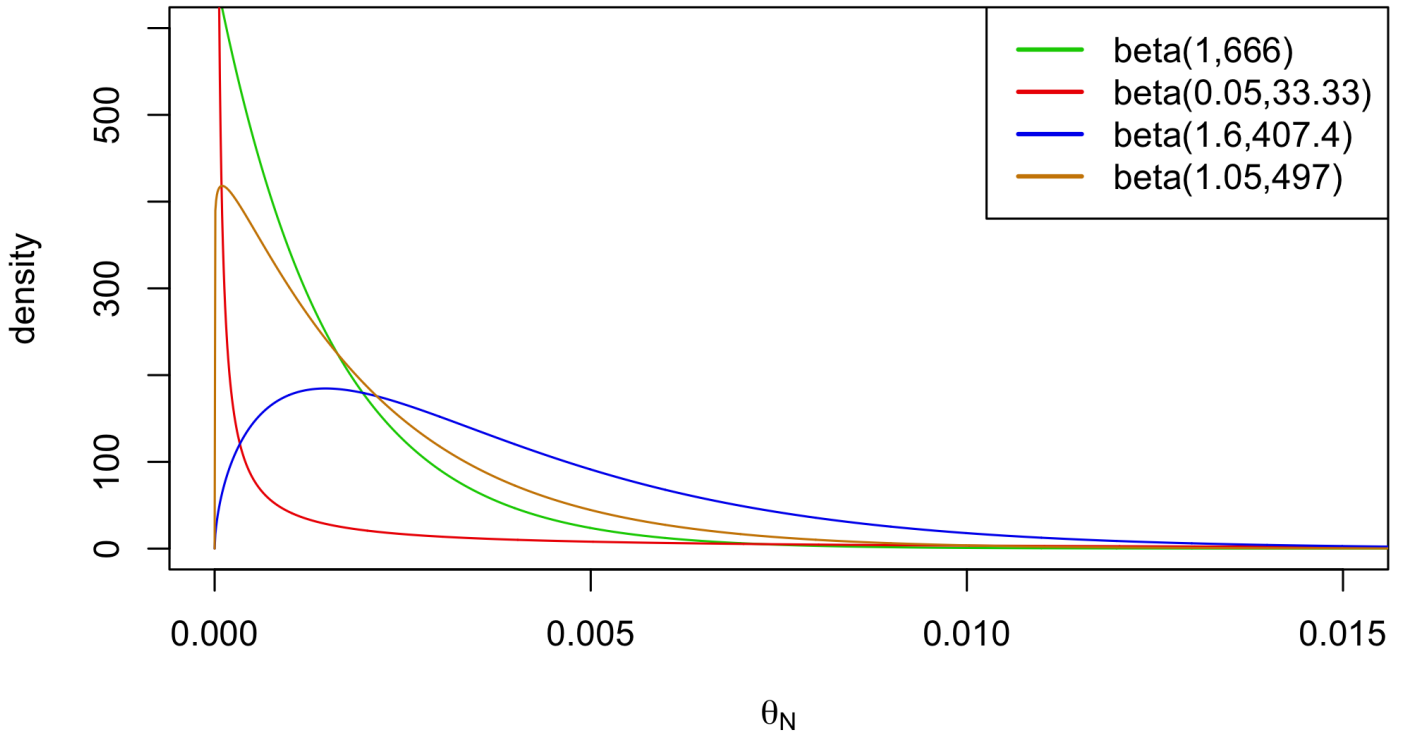
- One might consider a variety of choices centered on $15/10000 = 0.0015$:
  - Prior 1: Beta(1,666) (mean 0.0015; 1 prior bad reaction in 667 administrations)
  - Prior 2: Beta(0.05,33.33) (mean 0.0015; 0.05 prior bad reactions in 33.38 administrations)
  - Prior 3: Beta(1.6, 407.4) (mode 0.0015; 409 prior administrations)
  - Prior 4: Beta(1.05, 497) (median 0.0015; 498.05 prior administrations)
- Does it matter which one we choose?

Potential prior distributions



Potential prior distributions

Let's zoom in:



Potential prior distributions

- Let's take a closer look at properties of these four prior distributions, concentrating on the probability that $\theta_N < 0.0015$.
- That is, new agent not more dangerous than old agent.

	Be(1,666)	**Be(0.05,33.33)**	**Be(1.6,407.4)**	**Be(1.05,497)**
Prior prob	0.632	0.882	0.222	0.500
Post prob (0 events)	0.683	0.939	0.289	0.568
Post prob (1 event)	0.319	0.162	0.074	0.213

- Suppose we have a single arm study of 100 subjects.
- Consider the two most likely potential outcomes:
  - 0 adverse outcomes observed
  - 1 adverse outcome observed

Problems with the priors

- After just 100 trials with no bad reactions, the low weight (33.38 prior observations) prior indicates one should be 94% sure the new agent is equally safe as (or safer than) the old one.
- The low weight prior largely assumes the conclusion we actually hope for (that the new agent is safer), thus it takes very little confirmatory data to reach that conclusion.
- Is this the behavior we want?
- Take home message: be very careful with priors that have $a < 1$ with data that are all 0's (or $b < 1$ with all 1's).
- Given that we know the adverse event rate should be small, we might try a restricted prior e.g. Unif(0,0.1).

What's next?

Move on to the readings for the next module!