Lecture 1: Review and Background

Professor Alexander Franks

2023-01-12

Class Resources

Required Textbook

- Bayes Rules: https://www.bayesrulesbook.com/

Course Pages

- Class website on Canvas: (https://https://www.canvas.ucsb.edu/)
- Nectir for course related questions and discussion: https://ucsb.nectir.io/group/pstat115-w23
  - On Canvas site
- Gradescope: https://www.gradescope.com/courses/344698 (Enrollment code: BBRN6B)
  - On Canvas site

Grades

- 35% - expect approximately 5 homeworks
- 20% - Midterm (February 9)
- 10% - Quizzes
- 5% - Participation (Section attendance)
- 30% - Final exam (March 21)

Homework

- There will be approximately 6 homeworks (35% of your grade total)
- You will typically have 1-2 weeks to complete the homeworks
- You are allowed to work with a partner
  - Add partners name to your assignment
- Every student *must* submit their own assignment on gradescope
- Homework turned in within 24 hrs after the deadline without prior approval will receive a 10 pt deduction (out of 100)
- Homework will not be accepted more than 24 hrs late.

Homework submission format

- All code must be written to be reproducible in Rmarkdown
- All derivations can be done in any format of your choosing (latex, written by hand) but must be legible and *must be integrated into your Rmarkdown pdf*.
- All files must be zipped together and submitted to Gradescope
- Ask a TA *early* if you have problems regarding submissions.

Labs and Quizzes

- There will be a handful of "pop" quizzes throughout the quarter.
- There are no makeups, but the lowest quiz grade will be dropped from your final score.
- Quizzes (10%) will be multiple choice and will test your comprehension of the basic concept.
- Participation (5%). Includes lecture attendance, section attendance, and nectir posts.

2 Free misses

Class Policies

- All questions should be posted on nectir, *not by email* (unless they are personal or grade-related)

RStudio Cloud Service

- Log on to pstat115.lsit.ucsb.edu
  - Cloud based rstudio service
  - Log in with your UCSB NetID
- Use [https://tinyurl.com/32ra4at4] https://tinyurl.com/32ra4at4) to sync new material (BOOKMARK THIS)
- Make sure you can write and compile an R markdown (`Rmd`) document online
- Text formatting is minimal but syntax is simple

Logistics

- First homework out by end of week
  - Due January 22 (Sunday)
- Try pstat115.lsit.ucsb.edu
  - Cloud based rstudio service
  - Log in with your UCSB NetID
- Canvas website

Resources

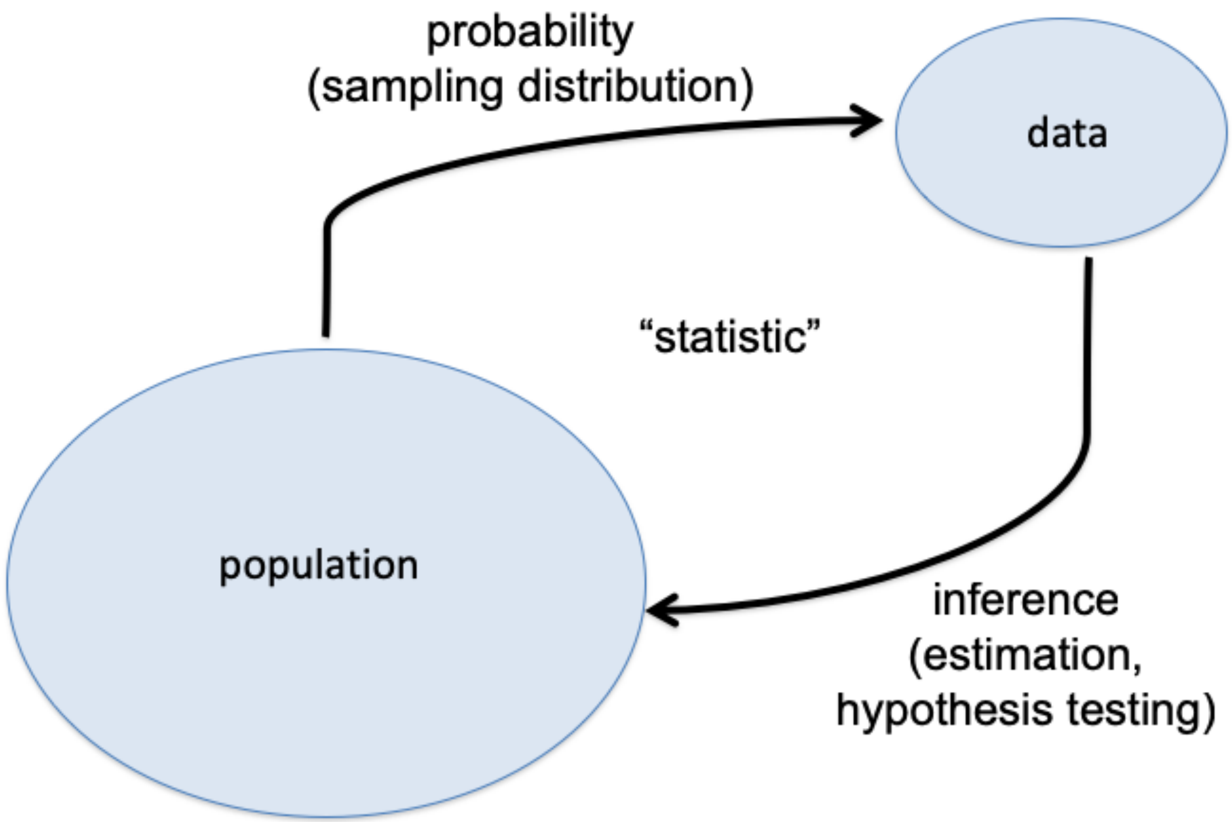
Look at the resources folder in cloud for

- A fantastic probability review sheet
- Probability density information
- Bayes Rules: Chapters 1 and 2

Logistics

- Homework 1 will be out by the end of the week in `winter23/homework/homework1.Rmd`
- Do not change the name of the file or the directory
- Autograding
    - Leave code cells that look like `. = ottr::check("tests/q1a.R")`

Population and Sample



Population and Sample

- The *population* is the group or set of items relevant to your question
  - Usually very large (often conceptualize a population as infinite)
- Sample: a finite subset of the population
  - How is the sampling collected (representative?)
  - Denote the sample size with n

Population and Sample

- Our goal is (usually) to learn about the population from the sample
  - Population parameters encode relevant quantities
  - The **estimand** is the thing we what to infer and is usually a function of the population parameters

Random variables

- A random variable, Y has variability, can take on several different values (possibly infinitely may), and is associated with a distribution.
  - The distribution determines the probability that the r.v. will take a specific value.
- Notation:
  - Y (uppercase) denotes a random variable
  - y (lowercase) is a *realization* of that random variable and is not random

$$Y \sim Bin(\underbrace{n}_{=10}, \underbrace{\theta}_{=0.5})$$

$$y = 6$$

Constants

- Constants: quantities with 0 variance.
    - Constants can be *known* (e.g. observed data) $y = 6$ heads
    - Constants can be *unknown* (not observed)

$\theta = 0.5$ (in frequentist paradigm)

Setup

- The *sample space* $\mathcal{Y}$ is the set of all possible datasets we could observe. We observe *one* dataset, y, from which we hope to learn about the world.

  $y \in \mathcal{Y}$, $y = 6$ heads

- The *parameter space* Θ is the set of all possible parameter values θ

  $\Theta = [0, 1]$ (probability of heads)

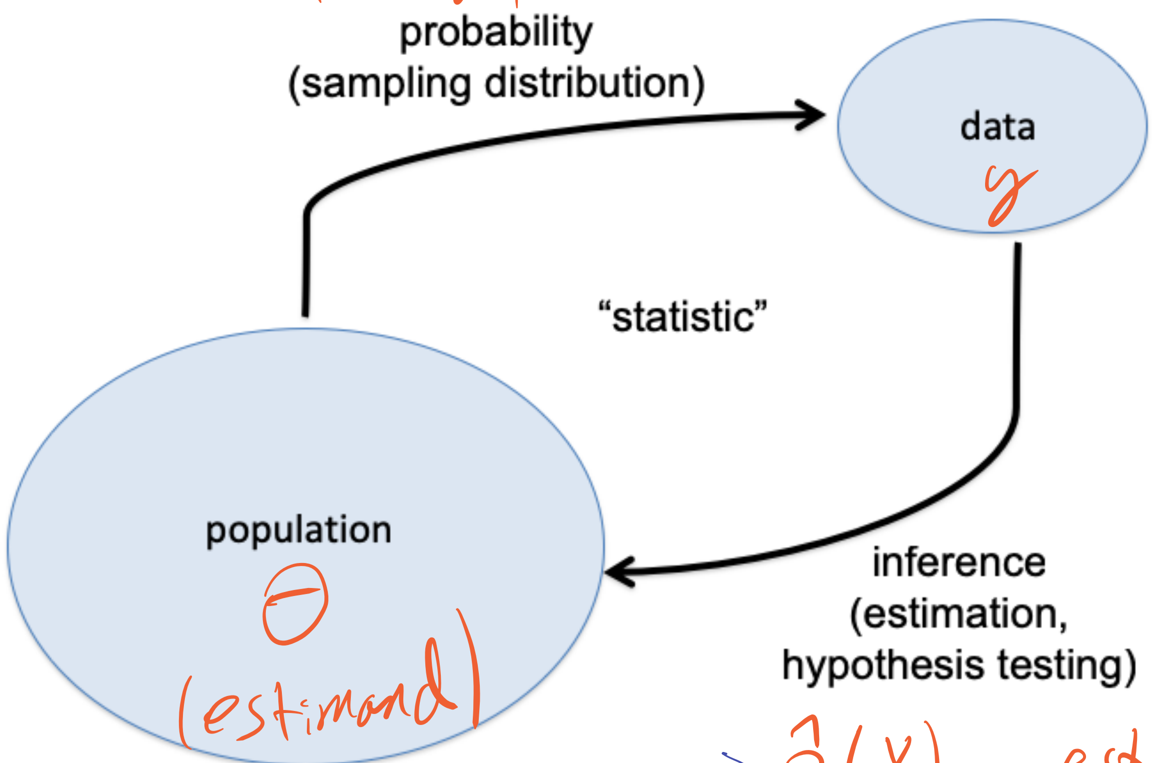
- θ encodes the population characteristics that we want to learn about

- Our *sampling model* $p(y \mid \theta)$ describes our belief about what data we are likely to observe for a given value of θ.

  $P(y=6 \mid \theta) \sim \text{Bin}(10, \theta)$

Notation

$P(Y=y \mid \theta)$



observed

variability $\rightarrow$ $\hat{\theta}(Y)$ estimator

constant $\rightarrow$ $\hat{\theta}(y)$ estimate

The Likelihood Function

- The likelihood is the "probability of the observed data" expressed as a function of the unknown parameter:
- A function of the unknown constant θ.
- Depends on the observed data $y = (y_1, y_2, \ldots, y_n)$

$$L(\theta) \propto P(y \mid \theta)$$

known data.

unknown

Independent Random Variables

- $Y_1, \ldots, Y_n$ are random variables
- We say that $Y_1, \ldots, Y_n$ are *conditionally* independent given θ if
- Conditional independence means that Y_i gives no additional information about Y_j beyond that in knowing θ

$$P(Y_1, Y_2, \ldots Y_n \mid \theta) = \prod_{i=1}^{n} P(Y_i \mid \theta)$$

$$P(Y_1, Y_2, \ldots Y_n) = \prod^{n} P(Y_i)$$

~~conditional indep $\Rightarrow$ Independence.~~

Example: A binomial model

- Assume I go to the basketball court and takes 5 free throw shots
- Model the number of made shots I make using a $\text{Bin}(5, \theta)$

  Probability shot (True Skill)

  - What are the key assumptions that make these a reasonable emodel?
- θ represents my true skill (the fraction of shots I make)
- How can we estimate my true skill?

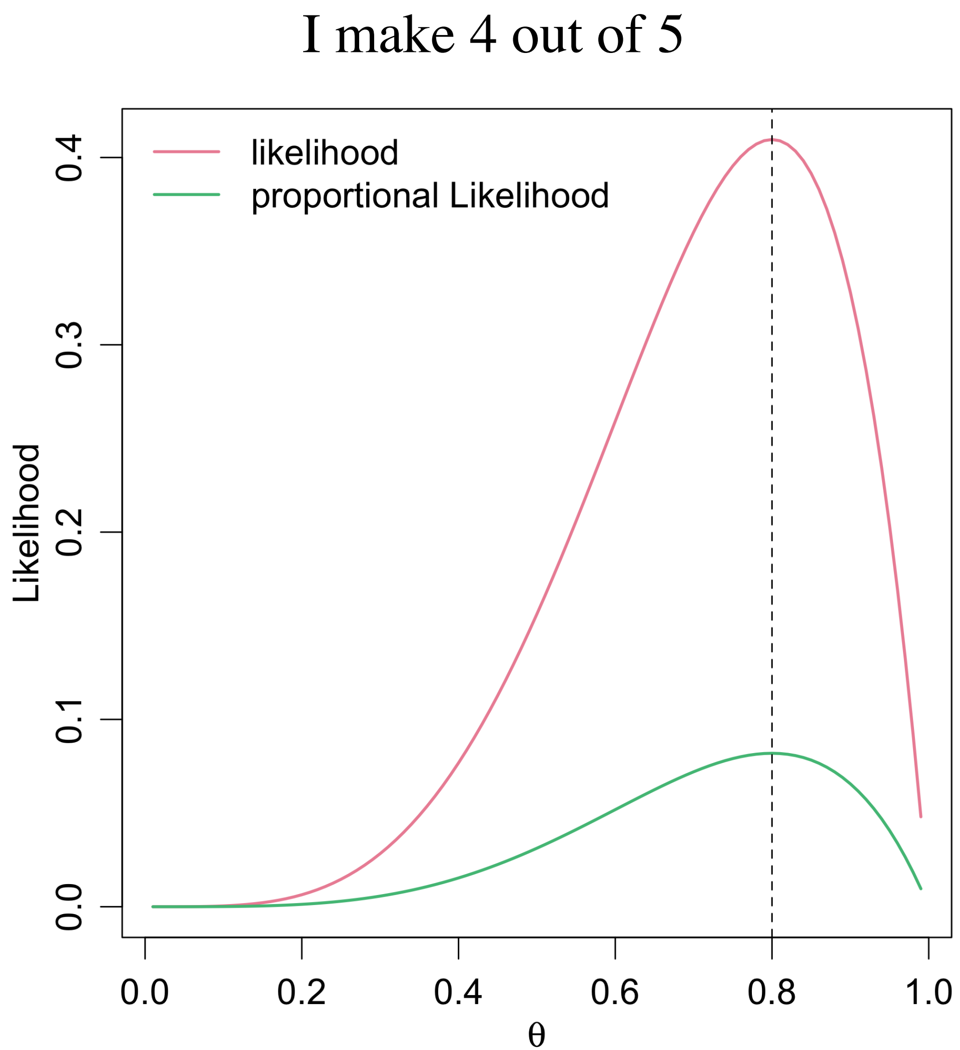
No θ, constant multiplier.

Likelihood: $L(\theta) = P(y \mid \theta) = \cancel{\binom{5}{y}} \theta^y (1-\theta)^{5-y}$

$\propto \theta^y (1-\theta)^{5-y}$ (made 4)

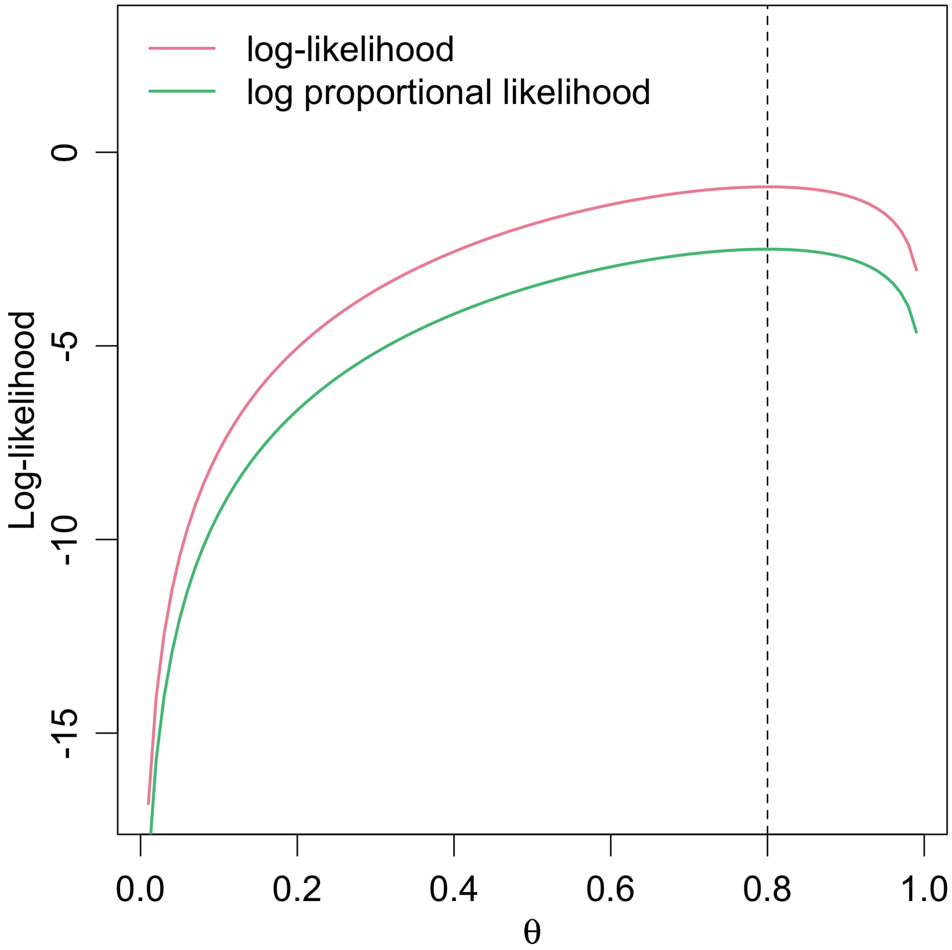
$\propto \theta^4 (1-\theta)^1$

The binomial likelihood



The log-likelihood

$\log L(\theta) \triangleq \ell(\theta)$



Maximum Likelihood Estimation

- The *maximum likelihood estimate* (MLE) is the value of θ that makes the data the most "likely", that is, the value that maximizes $L(\theta)$

- To compute the maximum likelihood estimate:

  1. Write down the likelihood and take its log:

  $$\log(L(\theta)) = \ell(\theta)$$

  2. Take the derivative of $\ell(\theta)$ with respect to θ:

  $$\ell'(\theta) = \frac{d\ell(\theta)}{d\theta}$$

  3. Solve for $\hat{\theta}$ such that $\ell'(\theta) = 0$

Maximum Likelihood Estimation

$\log(ab) = \log a + \log b$

$\log(a^b) = b \log a$

1. $L(\theta) \propto \theta^{y}(1-\theta)^{n-y}$

2. $\log L(\theta) = y \log \theta + (n-y)\log(1-\theta)$

3. $\ell'(\theta) = \dfrac{y}{\theta} - \dfrac{n-y}{1-\theta} = 0$

$$\frac{y}{\theta} = \frac{n-y}{1-\theta} \Rightarrow (n-y)\theta = y(1-\theta)$$

$$\hat{\theta}_{MLE} = \frac{y}{n}$$

Example: Binomial

- Assume we are polling the presidential race in the upcoming election
- We poll 25 random students in the class $Y_1, \ldots Y_n$ from $n = 25$
- Y_i is either 0 (Trump) or 1 (Biden)
- $Y_i \sim \text{Bern}(\theta)$, where $\text{Bern}(\theta)$ is equivalent to $\text{Bin}(1, \theta)$
  - Bernoulli random variables is a binomial with one trial
  - Assume our class is a simple random sample of the population
- How do we estimate θ for multiple observations?

$$
\begin{aligned}
L(\theta) &= P(y_1, \ldots y_{25} \mid \theta) \\
&= \prod_{i=1}^{25} P(y_i \mid \theta) \\
&= \prod^{25} \binom{1}{y_i} \theta^{y_i} (1-\theta)^{1-y_i}
\end{aligned}
$$

Example: the likelihood for independent Bernoulli's

$$\theta^{a}\,\theta^{b} = \theta^{(a+b)}$$

$$
\begin{aligned}
p(y_1, y_2, \dots, y_n|1, \theta) &= p(y_1, y_2, \dots, y_n|\theta)\\
&= p(y_1|\theta)p(y_2|\theta)\dots p(y_n|\theta)\\
&= \prod_{i=1}^{n} p(y_i|\theta)\\
&= \prod_{i=1}^{n} \binom{1}{y_i} \theta^{y_i}(1-\theta)^{(1-y_i)}\\
&= \underbrace{\cancel{\left[\prod_{i=1}^{n} \binom{1}{y_i}\right]}}_{h} \underbrace{\theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i}}_{g(\sum y_i, \theta)}\\
&= L(\theta)
\end{aligned}
$$

Multiplicative constant.

Sufficient Statistics

- Let $L(\theta) = p(y_1, \ldots y_n \mid \theta)$ be the likelihood and $s(y_1, \ldots y_n)$ be a statistic

$\sum y_i$

$\bar{y}$

$\left(\sum_{i=1}^{5} y_i, \sum_{i=5}^{n} y_i\right)$

- $s(y)$ is a sufficient statistic if we can write:

$$L(\theta) = h(y_1, \ldots y_n)g(s(y), \theta)$$

  - g is only a function of s(y) and θ only
  - h is *not* a function of θ

- This is known as the *factorization theorem*

$\left(\prod^{25} \binom{1}{y_i}\right)$ $\quad \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}$

- $L(\theta) \propto g(s(y), \theta)$

$s(y_1, .. y_n) = \sum y_i$ is sufficient!

Sufficient Statistics

- Intuition: a sufficient statistic contains all of the information about θ
  - Many possible sufficient statistics Σy_i equiv $\bar{Y}$
  - Often seek a statistic of the lowest possible dimension (minimal sufficient statistic)
  - What are some sufficient statistics in the previous binomial example?

Σy_i $\quad$ $\underline{(y_1, y_2, \ldots y_n)}$

$\bar{Y}$

$\cancel{(y_1, y_2, \ldots y_{n-1})}$

$$L(\theta) \propto \underbrace{h(y_1, \ldots, y_n)}_{\substack{\| \\ 1}} \underbrace{g(s, \theta)}_{\prod^{25} \binom{1}{y_i} \theta^{y_i} (1-\theta)^{1-y_i}}$$

$(y_1, \ldots y_n)$

Σy_i

Estimators and Estimates

- In classical (frequentist) statistics, θ is an unknown constant
- An **estimator** of a parameter θ is a function of the random variables, Y

  Capitalized

  - E.g. for Binomial$(1, \theta)$: $\hat{\theta}(Y) = \frac{\sum_i Y_i}{n}$

    Rule for estimating Θ.

  - An estimator is a random variable
  - Interested in properties of estimators (e.g. mean and variance)

Estimators and Estimates

- $\hat{\theta}(y)$ as a function of realized data is called an **estimate**
  - Plug in observed data $y = (y_1, \ldots y_n)$ to estimate θ
  - An estimate is a non-random constant (it is has 0 variability)
  - E.g. in the binomial$(1, \theta)$, $\hat{\theta} = \bar{y} = \frac{\sum_i y_i}{n}$ $= .8$ is the maximum likelihood estimate for the binomial proportion.

Bias and Variance

- Estimators are random variables. What are some r.v. properties that are desirable?

- Unbiased: on average $\hat{\theta}(Y)$ is θ

$$E[\hat{\theta}(Y)] = \theta$$

- Small variance, $Var(\hat{\theta}(Y))$ is small.

(not good on it's own, but can be good, if bias is low)

- Consistent: Get θ if $n \to \infty$

Accuracy: $E\left[\left(\hat{\theta}(Y) - \theta\right)^2\right]$

Bias and Variance

- Estimators are random variables. What are some r.v. properties that are desirable?

- Bias: $E[\hat{\theta}] - \theta = 0$

  - $E[\hat{\theta}] - \theta = 0$ means the estimator is unbiased

  - E.g. expectation of the binomial MLE: $E[\hat{\theta}] = E[\frac{\sum Y_i}{n}] = \theta$

$$E\left[\frac{1}{n}\sum Y_i\right] = \frac{1}{n}\sum E[Y_i]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\theta = \theta$$

- $\text{Var}(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$

E.g. variance of the binomial MLE is $= \frac{1}{n^2}\text{Var}(\sum Y_i)$

$$\text{Var}[\hat{\theta}] = \text{Var}(\frac{\sum Y_i}{n}) = \frac{\theta(1-\theta)}{n}$$

$$\text{Var}(a + bY) = b^2\,\text{Var}(Y), \quad \text{Var}(Y_1 + Y_2) \overset{\text{indep}}{=} \text{Var}(Y_1) + \text{Var}(Y_2)$$

$Var(Y_1 + Y_2) = Var(Y_1) + Var(Y_2) + 2Cov(Y_1, Y_2)$

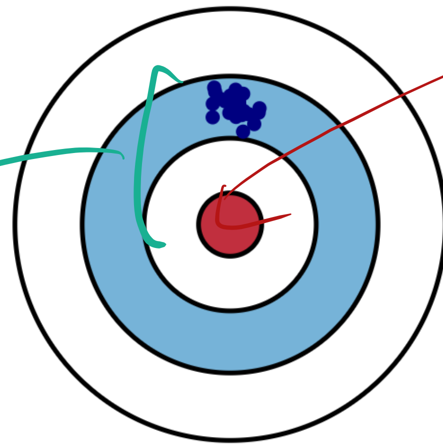
Bias and Variance

- Want estimators that have low bias and variance because this implies low overall error
- Mean squared error equals bias^2 + variance

Bias

The average difference between the prediction and the response

$E[\hat{\theta}] - \theta$



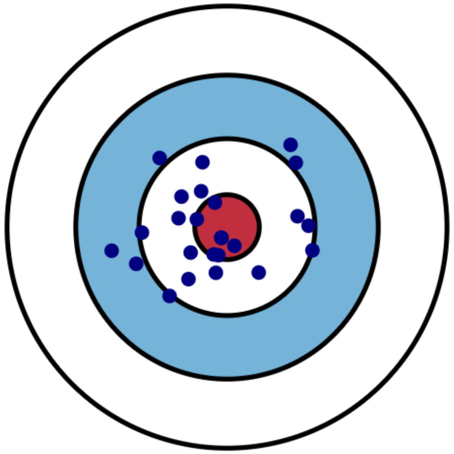
θ

Statistical definition of bias:

$$E[\hat{\theta} - \theta]$$

Variance

How variable is the prediction about its mean?



Statistical definition of variance:

$$E[\hat{\theta} - E[\hat{\theta}]]^2$$

Bias and Variance