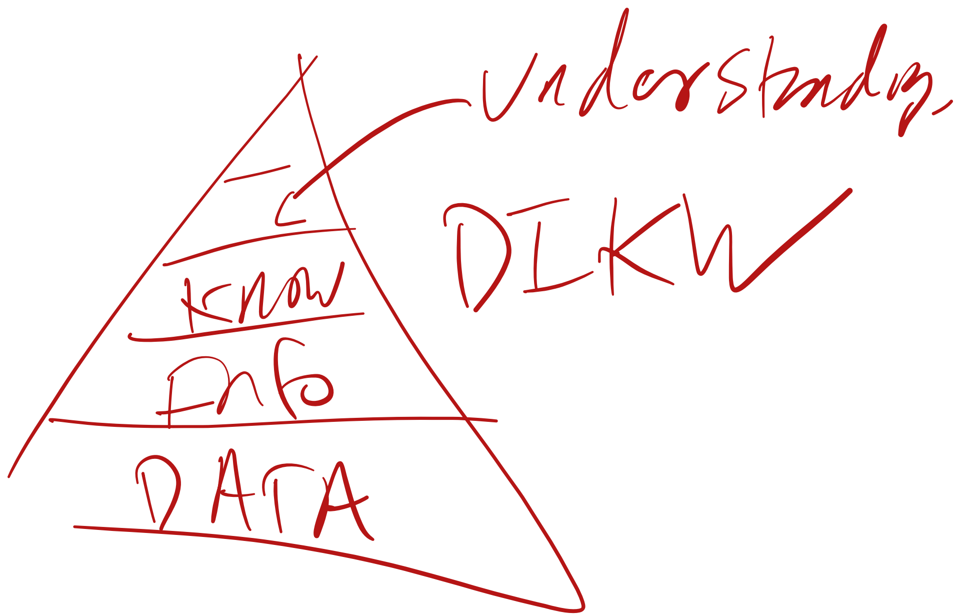
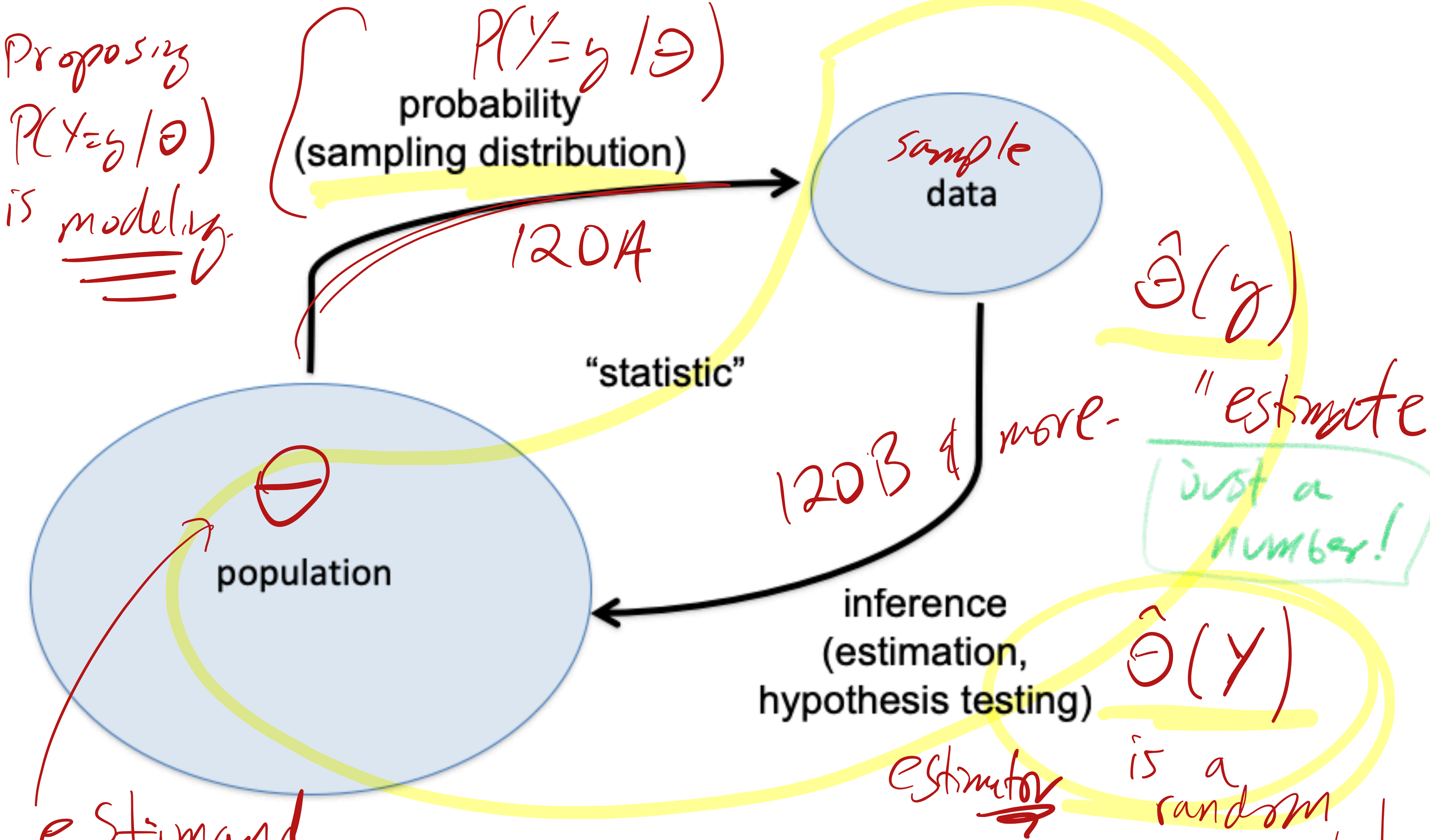
Modeling and Uncertainty

Understanding from knowledge

- Propose a “model” for the data
- Understand mechanisms and/or causality
- Allow us to generalize from the sample to the population





What are some principled choices for $\hat{\theta}$?

Warmup Example

Note

What is the typical family size (children only)?

Summarizing the Data

- Summary: c
- Data: $y_1, \ldots, y_n$
- Error: $y_1 - c, \ldots, y_n - c$
- Loss: l: $R \rightarrow R^+$

Summarizing the Data

Average Loss: $\frac{1}{n}\sum l(y_i, c)$ is also known as **Empirical Risk**

We can try to find the value c that minimizes the empirical risk for any loss function.

$$c = \min_c \frac{1}{n}\sum_{i=1}^{n} l(y_i, c)$$

$$l(y_i, c) = (y_i - c)^2 \Rightarrow c = \bar{y}$$

$$l(y_i, c) = |y_i - c| \Rightarrow c = med(y_1, \ldots, y_n)$$

Minimize the Average Loss

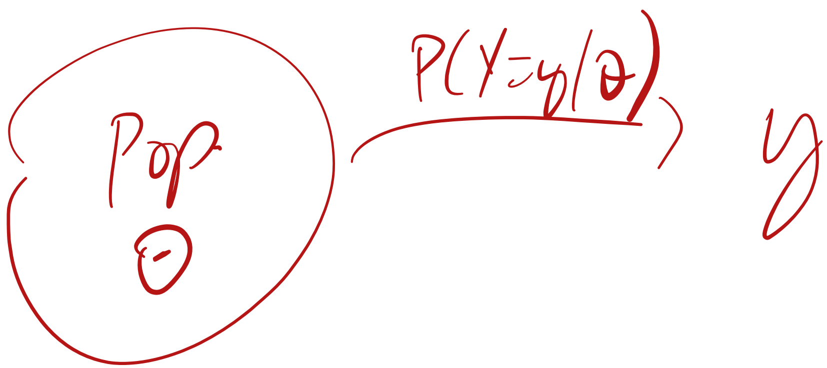
$$\frac{1}{n}\sum l(y_i, c) = \frac{1}{n}\sum (y_i - c)^2$$

The Sample Average Minimizes the Empirical Risk

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 <= \frac{1}{n}\sum_{i=1}^{n}(y_i - c)^2$$

Loss Functions and Risk

- A loss function is a real-valued function, L, of a random variable, Y, and a parameter value θ: $L(\theta, Y) \in \mathbb{R}$.
- Reminder: Y is a the random variable, y is observed data (const), and θ is an unknown parameter (also constant)
- Loss measures performance, indicating “how far” the parameter is from the data



$$\hat{\Theta} = \min \frac{1}{n}\sum L(\Theta, y)$$

Some Example Loss functions.

Name	Definition	Example
Squared Error	$L_2(Y-\theta(X))^2$	Least Squares, Regression
Absolute Error	$L_1(\lvert Y-\theta(X)\rvert)$	Robust regression
Zero-One Loss	$I(Y=\theta(X))$ 0 if $Y \neq \theta(x)$ 1 if $Y = \theta(x)$	Classification

Loss and Risk

- The loss function is random since how much is "lost" depends on what data is sampled
- We want estimators that are likely to give us small loss
- Idea: minimize average loss
- Risk is the expected value of a loss function

$$R_P(\theta) \equiv E_P[L(\theta, Y)]$$

minimize expected loss

where $Y \sim P_\theta$, i.e. P denotes the distribution of Y, which is parameterized by θ

Don't know it, in practice.

Empirical Risk

- Risk can be defined with respect to different distributions
  - the true unknown data generating distribution P
  - in practice, use the known data empirical distribution P_n.
- Empirical risk: $\frac{1}{n}\sum_i^n L(y_i, \theta)$
- A very broad class of statistical inference can be framed in terms of risk optimization.
- Risk minimization: $\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n}\sum_i L(y_i, \theta)$

Empirical Risk Minimization

- For the squared error loss function, the mean minimizes the empirical risk
- For absolute loss, the median minimizes the empirical risk
- How does the distribution of the data factor in?

- How do I decide what loss fn makes sense?

What is modeling?

Data Generating Process (DGP)

- DGP: a statistical model for how the observed data might have been generated , assumption about $P(Y=y|\theta)$
- Often write the DGP using pseudo-code. Example:

```
for (i in 1:N)
  - Generate y_i from a Normal(0, 1)
return y = (y_1, ... y_N)
```

- The DGP should tell a “story” about how the data came to be
- Can translate the DGP into a statistical model

Fwd model

Probability and Inference

The Binomial Distribution

- Discrete distribution
- PMF: $P(Y=y) = \binom{K}{y} p^y (1-p)^{n-y}$.
- Support: $[0, 1, \ldots, K]$
- K independent trials, each with success probability p.

Binomial Examples

- Number of heads in k flips of a coin
- Number of made basketball shots
- Number of patients cured by an experimental drug

The Poisson Distribution

- Discrete distribution
- PMF: $Pr(Y=y|\lambda) = \lambda^y e^{-\lambda} / y!$
- Support: $[0, 1, \ldots, \infty)$
- Counts events in a fixed interval of time (or space)
  - Assumes events occur with constant rate
  - Events independent of the time since the last event

- Insurance claims
- Cars passing a intersection
- Even on Zoom

$$Bin(n, p) = \binom{n}{y} p^{y} (1-p)^{n-y}$$

$E[Y] = n \times p$

Define $p = \frac{\lambda}{n}$

$$= \binom{n}{y} \left(\frac{\lambda}{n}\right)^{y} \left(1-\frac{\lambda}{n}\right)^{n-y}$$

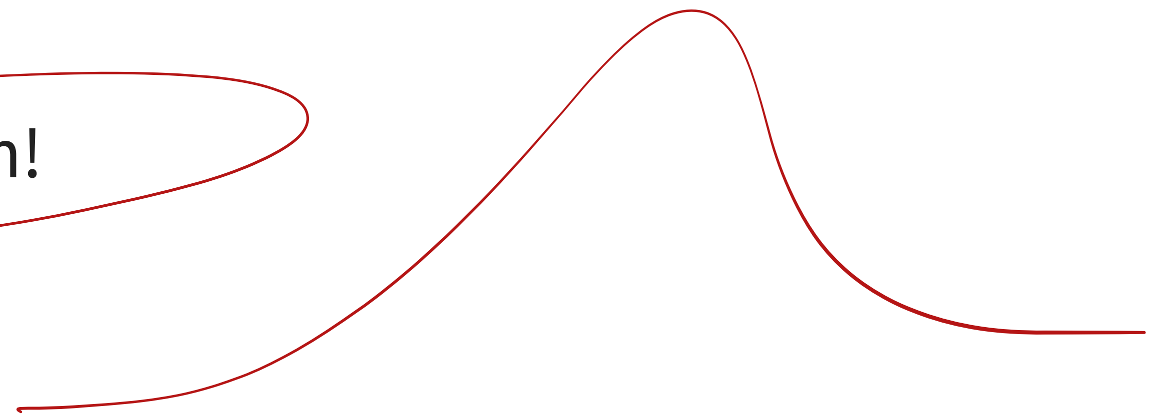
$$\lim_{n \to \infty} \qquad \frac{e^{-\lambda} \lambda^{y}}{y!}$$

Poisson random variable examples

- Cars passing an intersection in a fixed time
- Number of times a neuron in the brain fires
- Number of emails received in a day
- The number of patients arriving in an emergency room between 10 and 11 pm

The Normal distribution

- Continuous
- pdf $P(Y=y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$
- support: $(-\infty, \infty)$
- "bell-shaped data"
  - Central limit theorem!



The Normal Distribution

- Measurement error
- Test scores
- Approximating sums of independent variables

Exponential Distribution

- Continuous distribution
- pdf:
- Support:
- Often used to model time-to-event data
  - Memoryless property
  - Lengths of the times between events in Poisson process

Exponential Distribution Examples

- Time until a radioactive particle decays
- The time it takes for my next email to arrive
- Distance between mutations on a DNA strand

The Likelihood Function

- The likelihood is the "probability of the observed data" expressed as a function of the unknown parameter: $L(\theta; y) = p(y \mid \theta)$
- A function of the unknown constant θ.
- Depends on the observed data $y = (y_1, y_2, \ldots, y_n)$
- Minimizing the negative log likelihood is empirical risk minimization for a loss function determined by the model!

$$Y_1, \ldots Y_n \sim N(\mu, 1)$$

$$L(\mu) = \Pr(y_1, \ldots y_n \mid \mu) =$$

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}$$

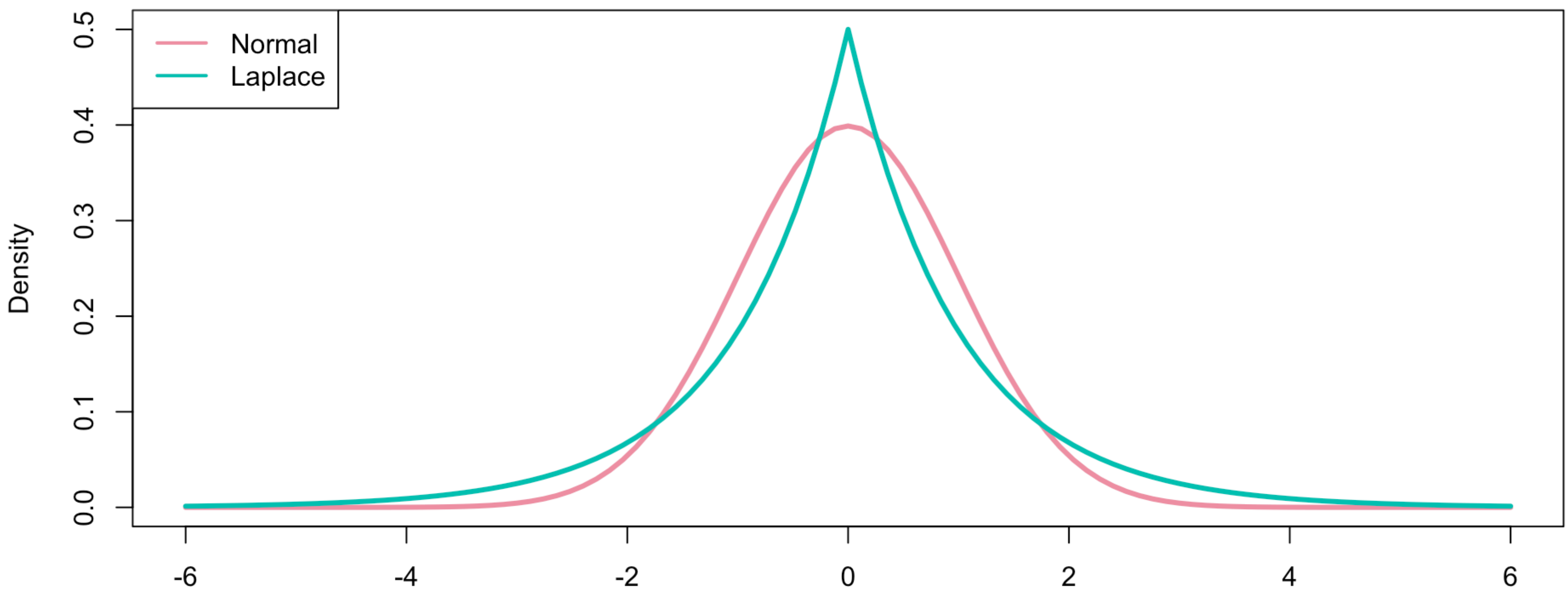
$$\log L(\mu) = -\sum_{i=1}^{n} (y_i-\mu)^2 / 2\sigma^2 + \ldots$$

Maximum Likelihood and Risk Optimization

What is the log-likelihood for iid normal random variable?

Laplace Random Variables

Laplace density: $p(y \mid \theta) = \frac{1}{2} e^{-|y-\theta|}$



Laplace Random Variables

What is the log-likelihood for iid Laplace random variables?

Poisson Example

- Wildlife biologists want to model how plentiful fish are in a particular river
- They count the number of fish, Y, passing a particular bottleneck in the river
- The count the fish in a total of n days
- When is a Poisson model reasonable?

Summary (loss first approach)

- Given any loss function, we can find estimate that minimizes the empirical risk
  - For estimating a "location parameter" MAE more robust to outliers than MSE
- Can choose a loss function based directly on its properties (i.e. robustness)
- Sometimes a loss function corresponds to a physical cost (e.g. in dollars)

Summary (model first approach)

- If we have a probability model, use it to identify the associated risk minimization problem
  - The negative log likelihood defines the loss function
  - Maximum likelihood for the mean of a **normal distribution** is equivalent to **MSE minizmation**
  - Maximum likelihood for the mean of **Laplace distribution** is equivalent to **MAE minimization**

Composing Statistical models

Mixture models

- A mixture model is a probabilistic model for representing the presence of sub-populations
- The sub-population to which each individual belongs is not necessarily known
- When z_i is not observed, we sometimes refer to it as a clustering model
    - “unsupervised learning”

Example Data Generating Process (DGP)

- The state wildlife biologists want to model how many fish are being caught at a state park.
- When visitors leave they are asked how many fish were caught.
- Some visitors do not go fishing, (they are guaranteed to catch 0 fish)
- Some go fishing but still don't catch any.
- Don't know who fishes and who doesn't.

$\Pr(Y = y \text{ fish} \mid \text{go fishing}) \sim \text{Pois}$

$\Pr(\# \text{ go fish}) \sim \text{Bin}(n, \pi)$

$s(y)$ statistic.

Just a number. Depends on sample!

$s(Y)$

Random variable.

Mixture Models

$$Z_i = \begin{cases} 0 & \text{if the } i^{th} \text{ if visitor doesn't go fishing} \\ 1 & \text{if the } i^{th} \text{ if visitor goes fishing} \end{cases}$$

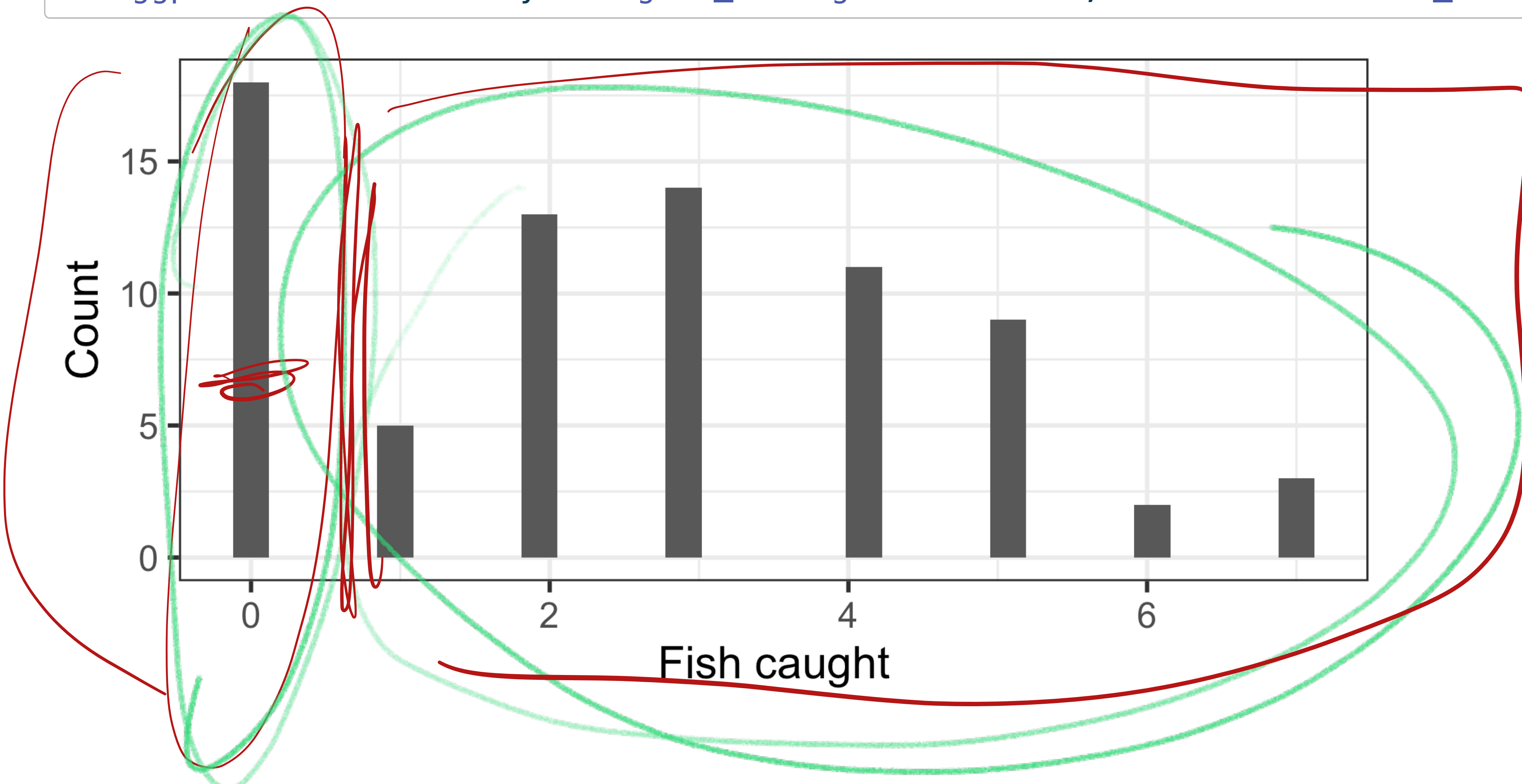
$$Z_i \sim \text{Bin}(1, p)$$

$$Y_i \sim \begin{cases} 0 & \text{if } Z_i = 0 \\ \text{Pois}(\lambda) & \text{if } Z_i = 1 \end{cases}$$

- p is the fraction of visitors that go fishing
- λ is the rate at which a visitor catches fish

Mixture Models

```
z <- ifelse(rbinom(75, 1, 0.8), "Fish", "Don't Fish")
y <- rpois(75, lambda=ifelse(z=="Fish", 3, 0))
ggplot(data.frame(x=y)) + geom_histogram(aes(x=x), bins=30) + theme_bw(base
```



$$\Pr(Y=y \mid \lambda, \pi) = \frac{\lambda^y e^{-\lambda}}{y!}$$

$$\begin{cases} (1-\pi) + \pi e^{-\lambda} & \text{if } y=0 \\ \pi \dfrac{\lambda^y e^{-\lambda}}{y!} & \text{if } y>0 \end{cases}$$

$$E[Y \mid Y>0] = \sum_{y=1}^{\infty} \Pr(Y=y \mid \lambda, y>0)\, y$$

$$= \sum_{y=1}^{\infty} \frac{\frac{\lambda e^{-\lambda}}{y!}}{1-e^{-\lambda}}\, y =$$

$$= \frac{1}{1-e^{-\lambda}} \underbrace{\sum_{y=1}^{\infty} \frac{\lambda e^{-\lambda}}{y!} y}_{\lambda}$$

$$=$$

$$\frac{\lambda}{1-e^{-\lambda}} = E[Y \mid Y>0]$$

$$\frac{\lambda}{1-e^{-\lambda}} = \bar{y}^{(>0)}$$

(MOM)

Solve for λ

Approx: $\frac{\lambda}{1-e^{-\bar{y}^{(>0)}}} = \bar{y}^{(y>0)}$

$$\lambda = \bar{y}^{(>0)}\left(1 - e^{-\bar{y}^{(>0)}}\right)$$

π?

$$\Pr(Y=0) = (1-\pi) + \pi e^{-\hat{\lambda}}$$

$$\text{Frac } 0\text{'s} \approx (1-\pi) + \pi e^{-\hat{\lambda}}$$

Solve for $\pi \longrightarrow \hat{\pi}$

$$\underbrace{\Pr(\hat{\lambda} = \ell \mid \lambda, \pi)}$$ (sampling distn of estimator)

Characterizing Uncertainty

I have $(\hat{\pi}, \hat{\lambda})$ from my

obs. sample. How much does

this vary about true (π, λ)?

How can I construct CI?

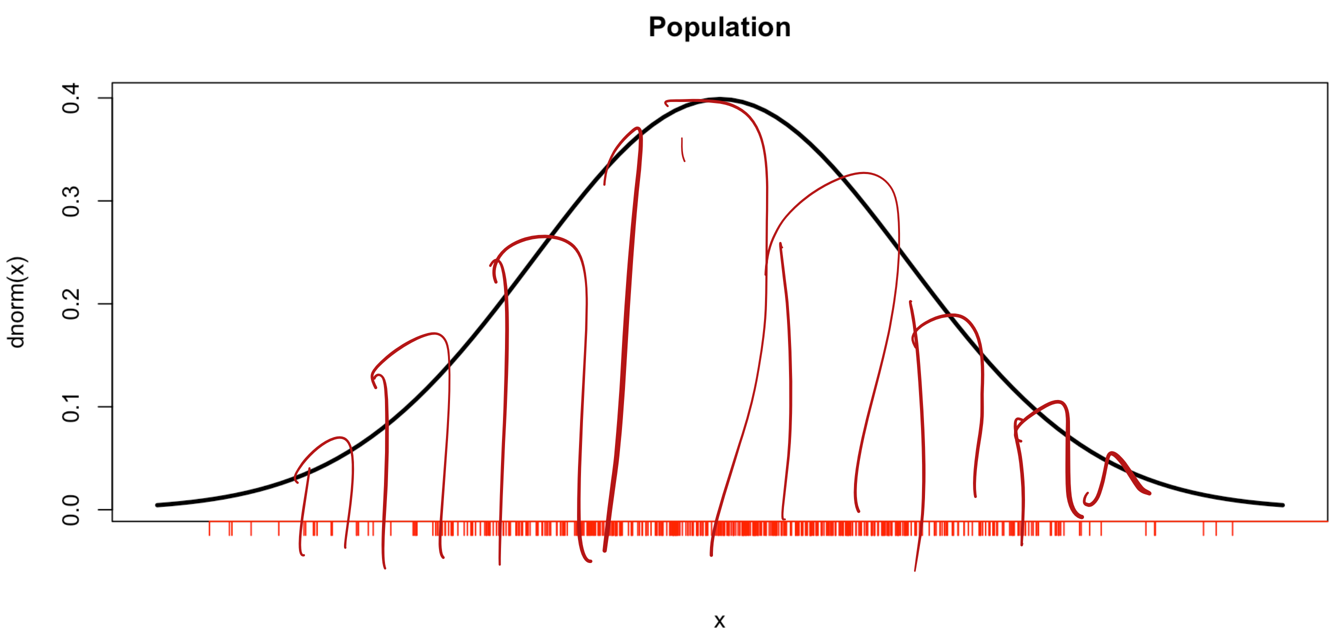
$y_1, \ldots y_n \sim \underline{N(\mu, 1)}$

$\hat{\mu} = \bar{y} \sim N(\mu, \frac{\sigma^2}{n})$ (120B)

- Variance of an estimator is due to *sampling* from a population
  - If you were to repeatedly draw new samples of the same size how much would your estimates vary?
  - e.g. if $Y_i \sim N(\mu, \sigma^2)$ then $\mathrm{Var}(\bar{Y}) = \sigma^2/n$
- *Bootstrap resampling* is a widely applicable tool for estimating the sampling properties of an estimator
  - It is a nonparametric technique: don't need to assume a true distribution of Y
  - Useful for deriving estimates' distributions when mathematically difficult or for small samples.
  - Constructing confidence intervals

The bootstrap

How do we “simulate” repeated sampling?



Answer: pretend the sample is our “population”

The Bootstrap

Generatge a new set of IID observations

$$\underbrace{Y_1, \cdots, Y_n}_{\text{original obs.}}$$

$$Y_1^*, \cdots, Y_n^*$$

where

$$P\left(Y_\ell^* = Y_i\right) = \frac{1}{n}, \quad \forall i = 1, \cdots, n$$

Sampling w/ replacement.

Repeat this process B times to obtain:

1000

$$\left(Y_1^{*(1)}, \cdots, Y_n^{*(1)}\right) \rightarrow \hat{\lambda}^{(1)}, \hat{\pi}^{(1)}$$

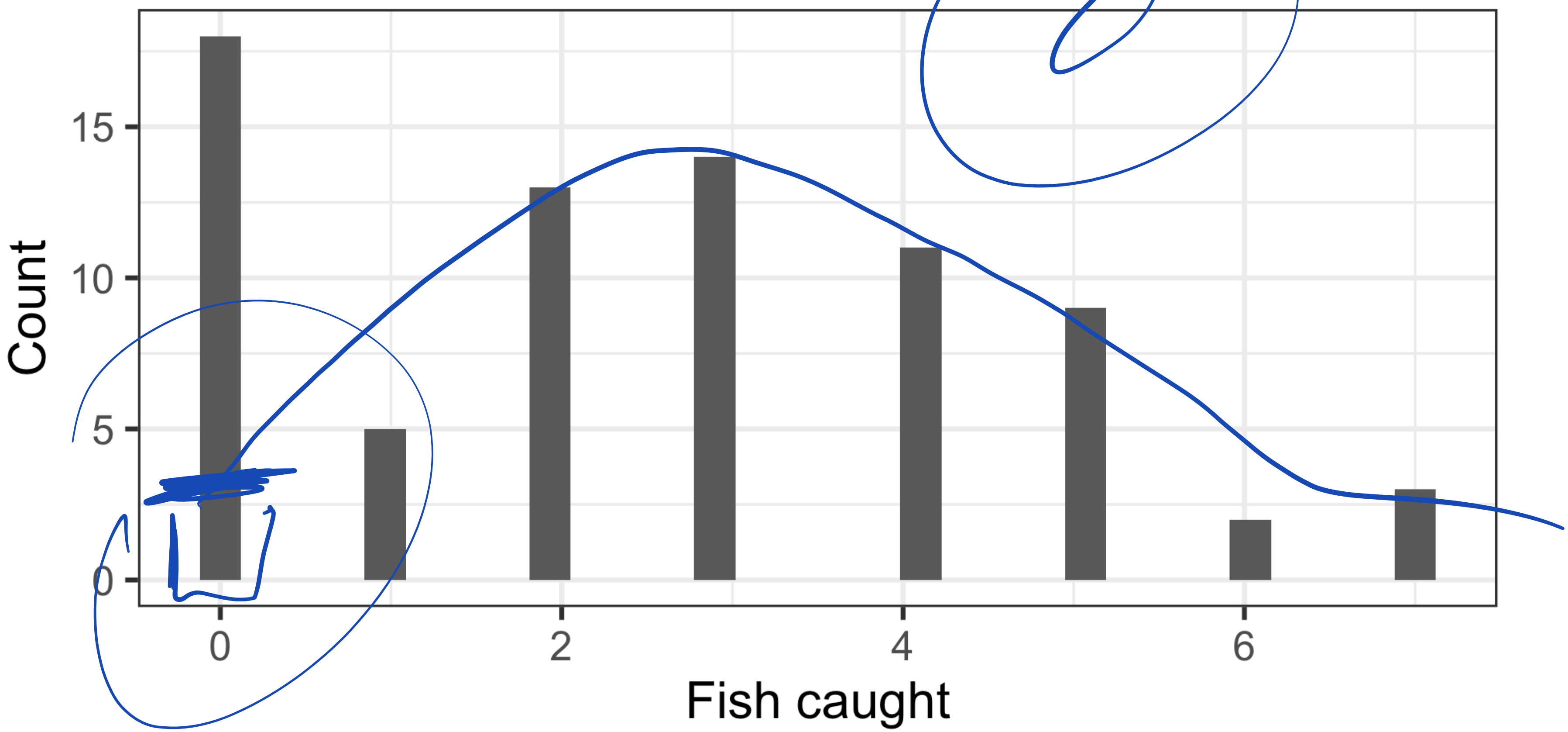
$$\left(Y_1^{*(2)}, \cdots, Y_n^{*(2)}\right) \quad \hat{\lambda}^{(2)}, \hat{\pi}^{(2)}$$

$$\vdots$$

$$\left(Y_1^{*(B)}, \cdots, Y_n^{*(B)}\right) \quad \hat{\lambda}^{(B)}, \hat{\pi}^{(B)}$$

What is the expected fish caught per visitor?

5



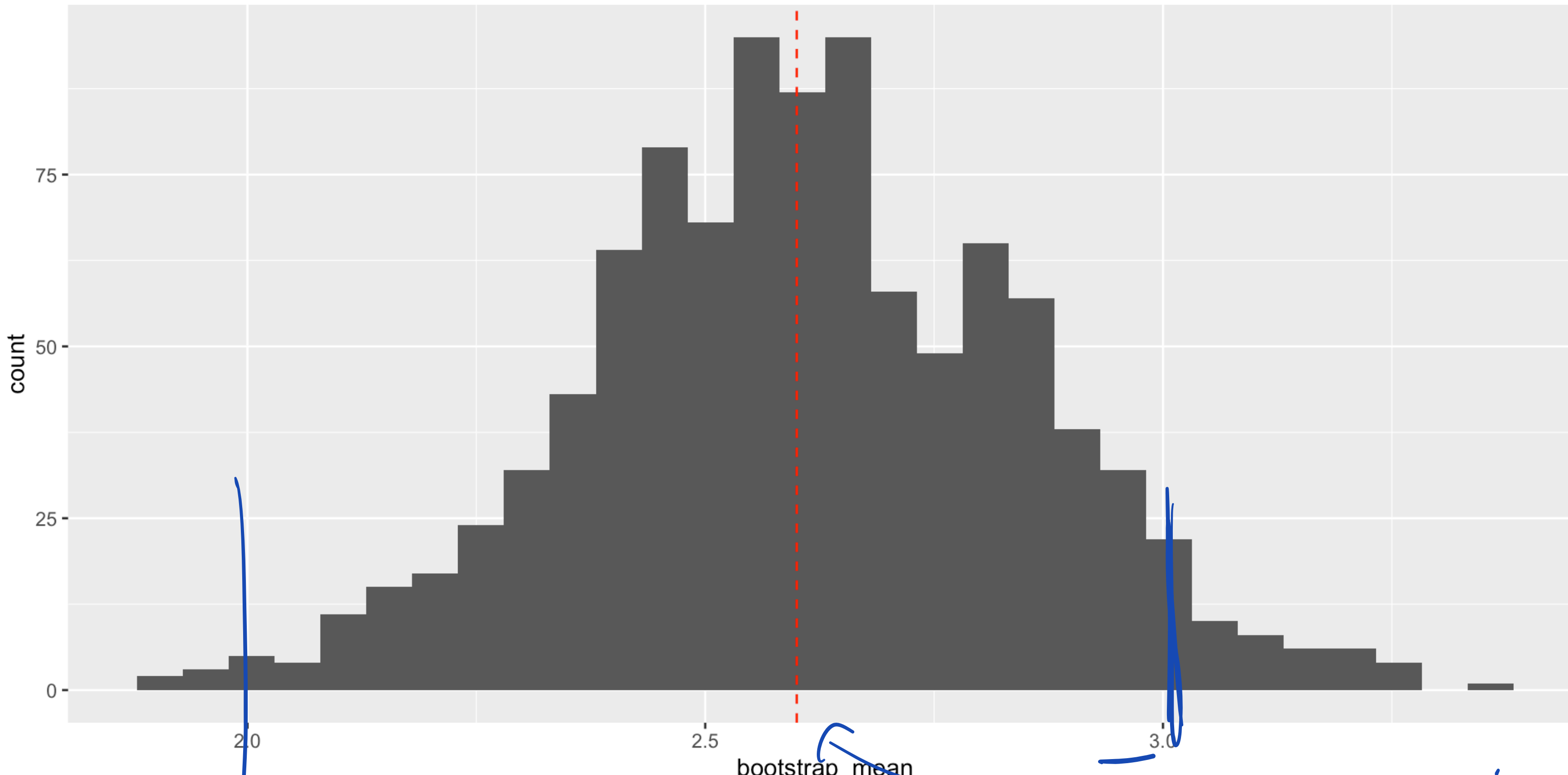
What is the uncertainty in our estimate?

Bootstrapping:

```
bootstrap_data <- tibble(fish=y) |> bootstraps(times=1000) |>
  mutate(bootstrap_mean = map_dbl(splits, \(x) mean(as_tibble(x)$fish)))

bootstrap_data |> ggplot() + geom_histogram(aes(x=bootstrap_mean)) +
  geom_vline(aes(xintercept=mean(y)), col="red", linetype="dashed")
```

What is the uncertainty in our estimate?



2.6 true sample

Better Estimands

- Better estimands would be:
  - The fraction of people who go fishing, p
  - The rate at which fishers catch fish, λ
- How do we estimate p and λ?

Maximum Likelihood in the Mixture Model

Method of moments

```
library(tidymodels)

get_lambda_hat <- \(y) {
  ytrunc <- y[y != 0]
  mean(ytrunc) - mean(ytrunc) * exp(-mean(ytrunc))
}

get_pi_hat <- \(y, lambda_hat) {
  1 - mean(y) / lambda_hat
}

get_lambda_hat(y)
```

```
[1] 3.309259
```

```
get_pi_hat(y, get_lambda_hat(y))
```

```
[1] 0.2143257
```

Quantifying Uncertainty

```r
bootstrap_data <- tibble(fish=y) |> bootstraps(times=1000) |>
  mutate(lambda_hat = map_dbl(splits, \(df) get_lambda_hat(as_tibble(df)$fi
  mutate(pi_hat = map2_dbl(splits, lambda_hat, \(df, l) get_pi_hat(as_tibbl

lambda_hat_uncertainty <- bootstrap_data |> ggplot() + geom_histogram(aes(x
  geom_vline(aes(xintercept=3), col="red", linetype="dashed")

pi_hat_uncertainty <- bootstrap_data |> ggplot() + geom_histogram(aes(x=pi_
  geom_vline(aes(xintercept=0.2), col="red", linetype="dashed")

lambda_hat_uncertainty + pi_hat_uncertainty
```

Quantifying Uncertainty