Our map of data analytics

I have a question about my data

→ Compute a Statistic

Single Variable questions

"What is the average mass of a penguin?"

"What is the probability C. Clark gets > 20 asst's?"

How certain am I that this answer is close to the population statistic?

→ estimate uncertainty with Bootstrapping (or classic stats)

"Joint" questions ("and")

"What fraction of games does Caitlin Clark get > 10 Ast, > 10 pts and > 10 Reb?"

"What is the mean bill len and flipper len.?"

What if my data is not representative enough of the population?

→ Fit a probability model

Conditional questions ("given")

"What is the mean mass of Adelie penguins?"

"What is probability of a penguin that weighs 5 kg being a Gentoo?"

Our map of data analytics

I have a question about my data

→ Compute a Statistic

How certain am I that this answer is close to the population statistic?

→ estimate uncertainty with Bootstrapping (or classic stats)

What if my data is not representative enough of the population?

→ Fit a probability model

Single Variable questions

"What is the average mass of a penguin?"

"What is the probability C. Clark gets > 20 asst's?"

"Joint" questions ("and")

"What fraction of games does Caitlin Clark get > 10 Ast, > 10 pts and > 10 Reb?"

"What is the mean bill len and flipper len..?"

Conditional questions ("given")

"What is the mean mass of Adelie penguins?"

"What is probability of a penguin that weighs 5 kg being a Gentoo?"

Our map of data analytics

I have a question about my data

→ Compute a Statistic

Single Variable questions

"What is the average mass of a penguin?"

"What is the probability C. Clark gets > 20 asst's?"

How certain am I that this answer is close to the population statistic?

→ estimate uncertainty with Bootstrapping (or classic stats)

"Joint" questions ("and")

"What fraction of games does Caitlin Clark get > 10 Ast, > 10 pts and > 10 Reb?"

"What is the mean bill len and flipper len.?"

What if my data is not representative enough of the population?

→ Fit a probability model ↔

Conditional questions ("given")

"What is the mean mass of Adelie penguins?"

"What is probability of a penguin that weighs 5 kg being a Gentoo?"

Our map of data analytics

I have a question about my data

→ Compute a Statistic

Single Variable questions

"What is the average mass of a penguin?"

"What is the probability C. Clark gets > 20 asst's?"

How certain am I that this answer is close to the population statistic?

→ estimate uncertainty with Bootstrapping (or classic stats)

"Joint" questions ("and")

"What fraction of games does Caitlin Clark get > 10 Ast, > 10 pts and > 10 Reb?"

"What is the mean bill len and flipper len.?"

What if my data is not representative enough of the population?

→ Fit a probability model

Conditional questions ("given")

"What is the mean mass of Adelie penguins?"

"What is probability of a penguin that weighs 5 kg being a Gentoo?"

Variable types and models

Type	Subtypes	Distributions
Nominal	Categorical	Bernoulli, Categorical
	General →	?
Ordinal →	Ordinal →	?
Quantitative	Integer: Positive	Poisson, Negative Binomial
	Integer: Bounded →	Binomial
	Real: Positive →	Exponential
	Real: Bounded →	Uniform
	Real: Unbounded	Normal, Student's t

Probability Notation & terminology

Random Variable: X → A value we haven't yet seen

e.g. The number of ast. Caitlin Clark gets in her next game

Support: What possible values could X take?

Distribution: How probable is each value of X?

PDF/pmf: $p(x) = \text{Prob}[X = x]$

function that defines these probabilities

"True distribution": What we would get if we picked an observation at random from the population (unknown)

Model: An approximation to the true distribution that makes some simplifying assumptions (estimated)

Probability Notation & terminology

mean of a sample $\bar{X} = \sum_{i=1}^{N} \frac{1}{N} X_i$

Average of observations

mean of a distribution $E[X] = \sum P(x)\,x$

[Expectation of a random variable] or $\int P(x)\,x\,dx$

Weighted average of all possible observations

Distributions w/ Mutiple Variables

Data set w/ 2 variables PTS and AST

Our next observation has 2 unknown values

X: number of PTS Y: number of AST

Joint distribution: How probable is every combination?

Joint pdf/pmf: $p(x, y) = Prob[X = x, Y = y]$

$p(10, 12) \rightarrow$ "What is the probability of 10 pts and 12 Ast?"

Distributions w/ Mutiple Variables

Marginal distribution: How probable is each value of X ignoring Y

Marginal Pdf/pmf: $P(x) = \text{Prob}[X = x]$

Independence: Condition that $p(x,y) = p(x)p(y)$ for all combinations

The values are unrelated
e.g. rolling 2 dice

Distributions w/ Mutiple Variables

Conditional distribution: how probable is each value of X if we know Y?

e.g. "If we know caitlin has 30 pts, we know she is playing well and may have more assists"

"If we know the penguin we're looking at is a gentoo, we know it's more likely to have a higher mass even before we weigh it"

Conditional pdf/pmf: $P(X|y) = Prob[X = x | Y = y]$

e.g. $P(5000|Gentoo)$ = What is the probability of our penguin being 5kg if we know it's Gentoo?

Probability Rules

$$P(X, Y) = P(X|Y)\,P(Y)$$

$$P(X) = \sum_{Y} P(X, Y) \quad \text{(discrete)} \qquad \int P(X, Y)\,dY \quad \text{(cont.)}$$

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

Categorical distribution

If we want a distribution over a categorical variable, we need a prob. for each value

$$p(c) = p_c \quad \text{or} \quad P(x) = \sum_{c=1}^{K} \mathbb{I}(x=c)\, p_c$$

$$\text{MLE: } p_c = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(x_i = c)$$

For 2 variables:

$$p(c,d) = p_{cd} \quad \text{or} \quad P(x,y) = \sum_{c=1}^{K} \sum_{d=1}^{K} \mathbb{I}(x=c)\mathbb{I}(y=d)\, p_{cd}$$

$$\text{MLE: } p_{cd} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(x_i = c)\, \mathbb{I}(y_i = d)$$

Bernoulli distribution

Categorical with $\underline{2}$ possible values $\{0, 1\}$

→ only need 1 probability

$$P(1) = p \qquad P(0) = 1 - p$$

$$P(x) = p\,I(x=1) + (1-p)\,I(x=0) = p^x (1-p)^{1-x}$$

Conditional distribution

$$P(X|y) =$$

Compute probability for every subset

MLE: $$P_{cld} = \frac{\sum_{i=1}^{N} I(x_i = c) I(x_i = d)}{\sum_{i=1}^{N} I(y_i = d)}$$

Subset w/ $y = d$

In Pandas or

Not really any distinction from Sample Prof. (yet!)

Computed Variables

Our "variables" don't need to be columns of our original dataframe

→ we can compute new ones

E.g. Quantitative variable:
Caitlin Clark's assists

→ Categorical variable:
Caitlin Clark's assists ≥ 10

Space of Conditional Models

$$P(y \mid X)$$

Dependent Variable

Independent Variable

y \ X	Categorical X	Quantitative X
Categorical y	Categorical y, Categorical X	Categorical y, Quantitative X
Quantitative y	Quantitative y, Categorical X	Quantitative y, Quantitative X

Ordinal later!

Space of Conditional Models

$$P(y \mid X)$$

Dependent Variable

Independent Variable

y \ X	Categorical X	Quantitative X
Categorical y	✓	Ordinal Logistic
Quantitative y	→	Quantitative

P(Quantitative | Categorical)

Straight forward!

- Choose appropriate distribution for quantitative variable
- Fit on each subset

E.g. Model Caitlin Clark's assists conditioned on Home Vs. Away

$$P(\text{AST} \mid \text{home}) = \text{Poisson}(\text{AST}; \hat{\lambda}), \quad \hat{\lambda} = \frac{\sum_{i=1}^{N} \mathbb{I}(x_i = \text{home})\, y_i}{\sum_{i=1}^{N} \mathbb{I}(x_i = \text{home})}$$

Space of Conditional Models

$$P(y \mid X)$$

Dependent Variable

Independent Variable

y \ X	Categorical X	Quantitative X
Categorical y	✓	Categorical y ← ordinal later!
Quantitative y	✓ Categorical X	Quantitative X

P(Categorical | Quantitative)

E.g. What's the probability the fever won the game given how many points Caitlin Clark scored?

"Compute the probability on every subset..."

P(won|0), P(won|1), P(won|2), ...

Not easy!

P(won | 20) = 1

P(won | 22) = 0 ?

P(Categorical | Quantitative)

Same problem that motivated models in the first place! → We need to make more assumptions

Lets consider 2 possibilities:

1) Assume $P(X \mid y)$ is easy to model

e.g. $P(\text{Pts} \mid \text{won}) = \text{Poisson}(\lambda_w)$

$P(\text{pts} \mid \text{loss}) = \text{Poisson}(\lambda_l)$

2) Assume $P(y \mid x)$ changes predictably as x changes

e.g.

$$P(y \mid x) = \frac{e^{ax+b}}{1+e^{ax+b}}$$

P(Categorical | Quantitative)

How does modeling P(x|y) help us?

Bayes rule!

Quantitative | Categorical ✓ — easy to estimate!*

↓

Marginal

$$P(y|x) = \frac{P(x|y)\,P(y)}{P(x)}$$

$$P(x) = \sum_{y} P(x|y)\,P(y)$$

$P(\text{Categorical} \mid \text{Quantitative})$

Ex.

$$P(\text{win} \mid \text{pts}) = \frac{P(\text{pts} \mid \text{win})\, P(\text{win})}{P(\text{pts} \mid \text{win})P(\text{win}) + P(\text{pts} \mid \text{loss})P(\text{loss})}$$

$$P(\text{pts} \mid \text{win}) = \text{Poisson}(\hat{\lambda}_w), \quad \hat{\lambda}_w = \frac{\sum_{i=1}^{N} x_i \, I(y_i = \text{win})}{\sum_{i=1}^{N} I(y_i = \text{win})}$$

$x_i = \text{pts}$, $y_i = \text{win/loss}$

$$P(\text{pts} \mid \text{loss}) = \text{Poisson}(\hat{\lambda}_l)$$

$$P(\text{win}) = \frac{\sum_{i=1}^{N} I(y_i = \text{win})}{N} \qquad P(\text{loss}) = \frac{\sum_{i=1}^{N} I(y_i = \text{loss})}{N}$$

P(Categorical | Quantitative)

Wait! This is just a mixture model, But we know the assignments!

M-Step: Fit $\bar{P}(x|y)$, $P(y)$

Get probabilties for unknown y

E-Step: find $p(y|x)$

P(Categorical | Quantitative)

Assume P(Y|X) changes predictably with X

Logistic regression! → For 2 outcomes

Reminder: Let's define $P(y=1 \mid x=0) = b$

assume as x increases by 1, P(y|x) always increases by a

So, $P(y|x) = ax + b$, But

$P(y|x) \in [0, 1]$ or bad things happen ...

P(Categorical | Quantitative)

So, apply a function that maps $\mathbb{R} \to [0,1]$

e.g. Sigmoid $= \sigma(x) = \frac{e^x}{1+e^x}$

Logistic regression

$$P(y|x) = \frac{e^{ax+b}}{1+e^{ax+b}}$$ ← Usually the better choice

Probit regression

$$P(y|x) = \Phi(ax+b)$$

Φ Normal cdf

P(Categorical | Quantitative)

How do we chose a and b?

Maximum likelihood!

$$\underset{a,b}{\operatorname{argmax}} \prod_{i=1}^{N} P(y_i | x_i) = \underset{a,b}{\operatorname{argmax}} \sum_{i=1}^{N} \log P(y_i | x_i)$$

$$= \operatorname{argmax} \sum_{i=1}^{N} \log \frac{e^{ax+b}}{1+e^{ax+b}}$$

No easy solution ☹ → Gradient descent

P(Categorical | Quantitative)

What Assumptions are we making for each?

Bayesian Model

P(y|x) determined by ratio of distance to each mean

Not monotonic in x

Logistic Regression

P(y|x) determined by magnitude of x

monotonic in x

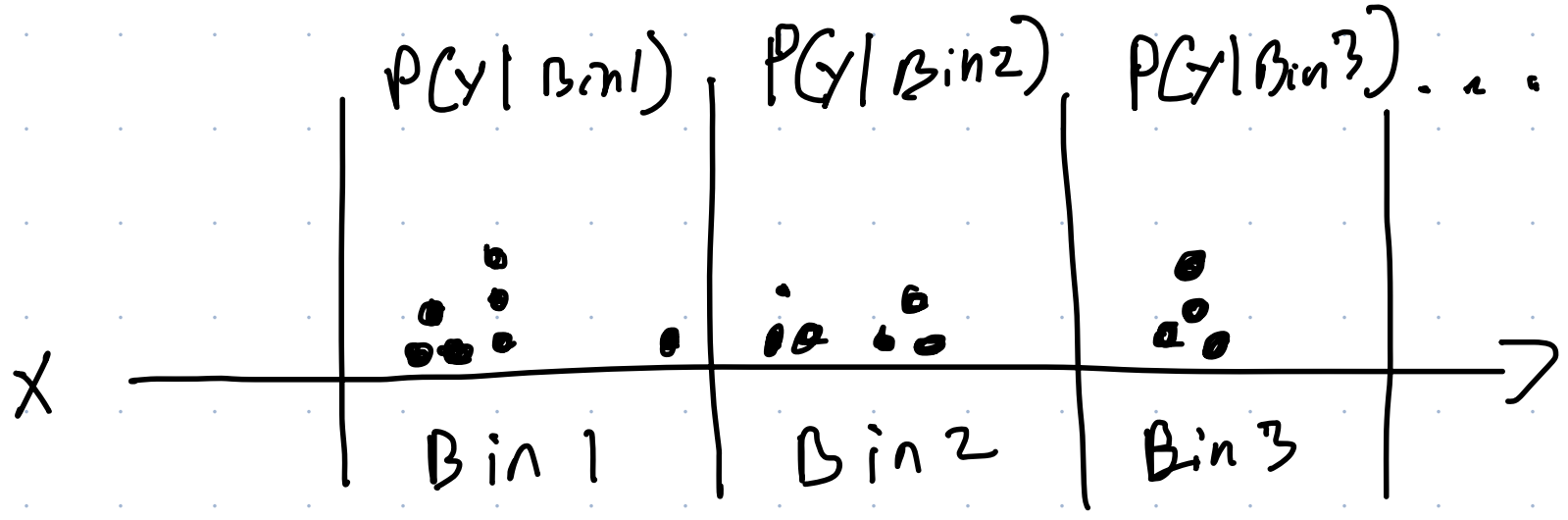
Which is better? Depends on the data!

Can compare likelihoods

P(Categorical | Quantitative)

Other alternatives?

What about a histogram style approach?



→ Decision trees

(Maybe) Later in this course ...

Space of Conditional Models

$$P(y \mid X)$$

Dependent Variable

Independent Variable

y \ X	Categorical X	Quantitative X
Categorical y	✓	✓
Quantitative y	✓	✗

Ordinal later!

P(Quantitative | Quantitative)

Does our Bayesian method work?

No ☹

$$P(y|x) = \frac{P(x|y)\,P(y)}{P(x)}$$

Still would need to enumerate every y

P(Quantitative | Quantitative)

Linear regression:

$P(Y|X)$ is Normal $\mathcal{N}(\mu, \sigma)$

Let mean (μ) of $P(Y|X=0) = b$

Assume that as x increases by 1
μ increases by $a \rightarrow \mu = ax + b$
[σ is Constant]

$$P(y|x) = \mathcal{N}(ax+b, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{y-ax+b}{\sigma}\right)^2}$$

P(Quantitative | Quantitative)

Picking a, b and σ

Maximum likelihood!

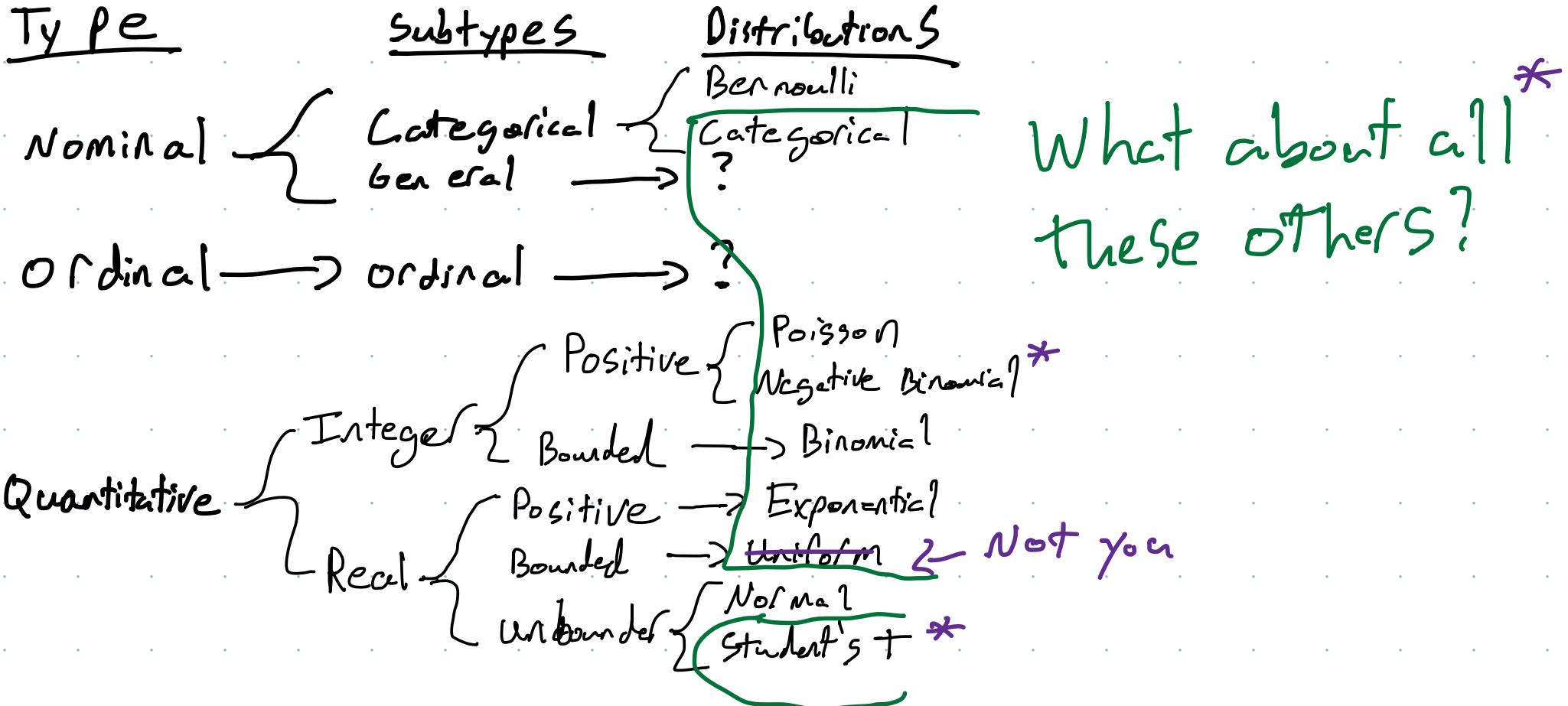
$$\underset{a,b}{\operatorname{argmax}} \prod_{i=1}^{N} P(Y_i \mid X_i) = \underset{a,b}{\operatorname{argmax}} \sum_{i=1}^{N} \log P(Y_i \mid X_i)$$

$$= \operatorname{argmax} \sum_{i=1}^{N} \log \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y_i - aX_i + b}{\sigma}\right)^2}$$

Here there is an "easy" solution

P(Quantitative | Quantitative)

Linear regression lets us assume $p(y|x)$ is Normal. Logistic regression assumed Bernoulli



Generalized linear models

Mean of a Normal distribution $E[x] = \mu$

Linear regression $E[y|x] = \mu = ax+b$

Mean of a Bernoulli distribution $E[x] = p$

Logistic regression $E[y|x] = p = \sigma(ax+b)$

General Distribution

obey conditions for mean

Assume $E[y|x] = f(ax+b)$

Generalized linear models

What if we want $p(y|x)$ to be a poisson?

Unconditional poisson $p(y) = \frac{\lambda^y e^{-\lambda}}{y!}$

mean: $E[Y] = \lambda$

Poisson regression:

Let $\lambda = ax+b$? No! $\lambda \geq 0$

$\lambda = e^{ax+b}$

$f(x) = e^x$

$p(y|x) = \frac{(e^{ax+b})^y e^{-e^{ax+b}}}{y!}$

Generalized linear models

Recipe:

Distribution: Normal, Bernoulli, etc.

Structure: $\beta_1 x + \beta_0$

Link function: $g(\ )$

$$E[Y|x] = g^{-1}(\beta_1 x + \beta_0)$$

Categorical Regression

What if we want $p(y|x)$ to be categorical w/ more than 2 outcomes?

e.g. $y \in \{\text{"a"}, \text{"b"}, \text{"c"}\}$, no ordering

Categorical distribution

$p(a) = p_a = .5 \quad p(b) = p_b = .3 \quad p(c) = p_c = .2$

$$E[y] = .5a + .3b + .2c = ?$$

"Dummy" Variables

Already saw we can compute new variables, compute 1 for each value!

$$Y \in \{a, b, c\}$$

$$Y \longrightarrow \begin{cases} A = I(Y=a) \in \{0,1\} \\ B = I(Y=b) \in \{0,1\} \\ C = I(Y=c) \in \{0,1\} \end{cases}$$

$$E[A] = P_a \quad E[B] = P_b \quad E[C] = P_c$$