Bayesian Modeling Workshop

Daniel Lee

Zelus Analytics

dlee@zelusanalytics.com

Thank You!!!

SSAC Organizers

- Chris Couch
- Zach Wang
- Tal Gilad
- Jessica Gelman

Support From

- Kiran Gauthier
- Evan Miyakawa
- Alex Franks

Stan Case Study

- Andrew Gelman
- Mark Broadie
- https://mc-stan.org/users/documentation/case-studies/golf.html

Everyone here!

Preamble

Welcome!

Goal: everyone learns something

- Hands-on.
- Light on math.

- This is your time. Please ask questions.

Workshop materials: https://github.com/bayesianops/ssac24

Outline

1. Data
2. Statistical Models
3. Hands-on work with Stan
4. Q&A

Workshop materials: https://github.com/bayesianops/ssac24

Workshop Materials

- Slides
- Data
- R script
- Stan programs

Workshop materials: https://github.com/bayesianops/ssac24

If you have your computer

- Git clone or download:
  https://github.com/bayesianops/ssac24
- R:
  - Install CmdStanR

Workshop materials: https://github.com/bayesianops/ssac24

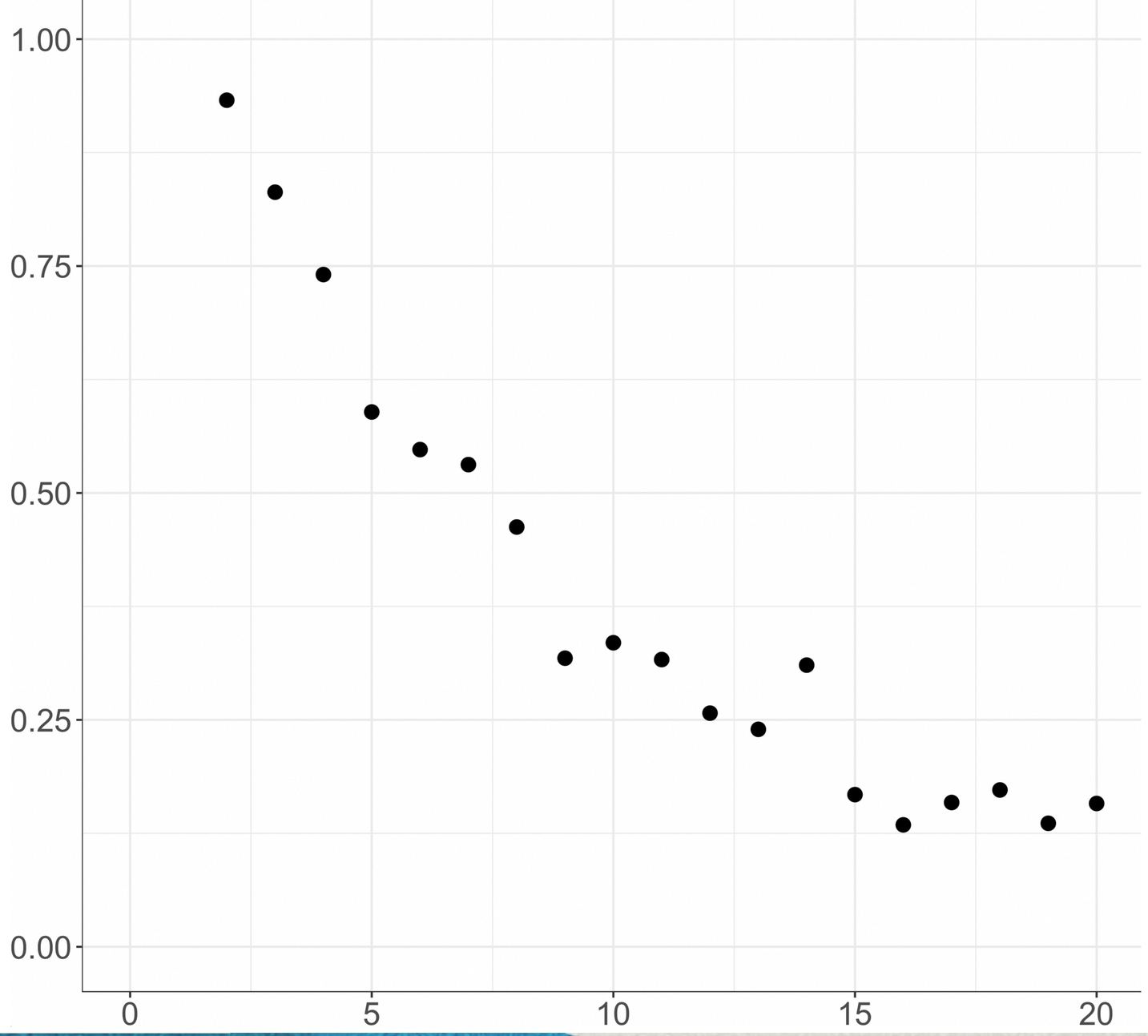
Data

Data

- What do you notice?

- Data:

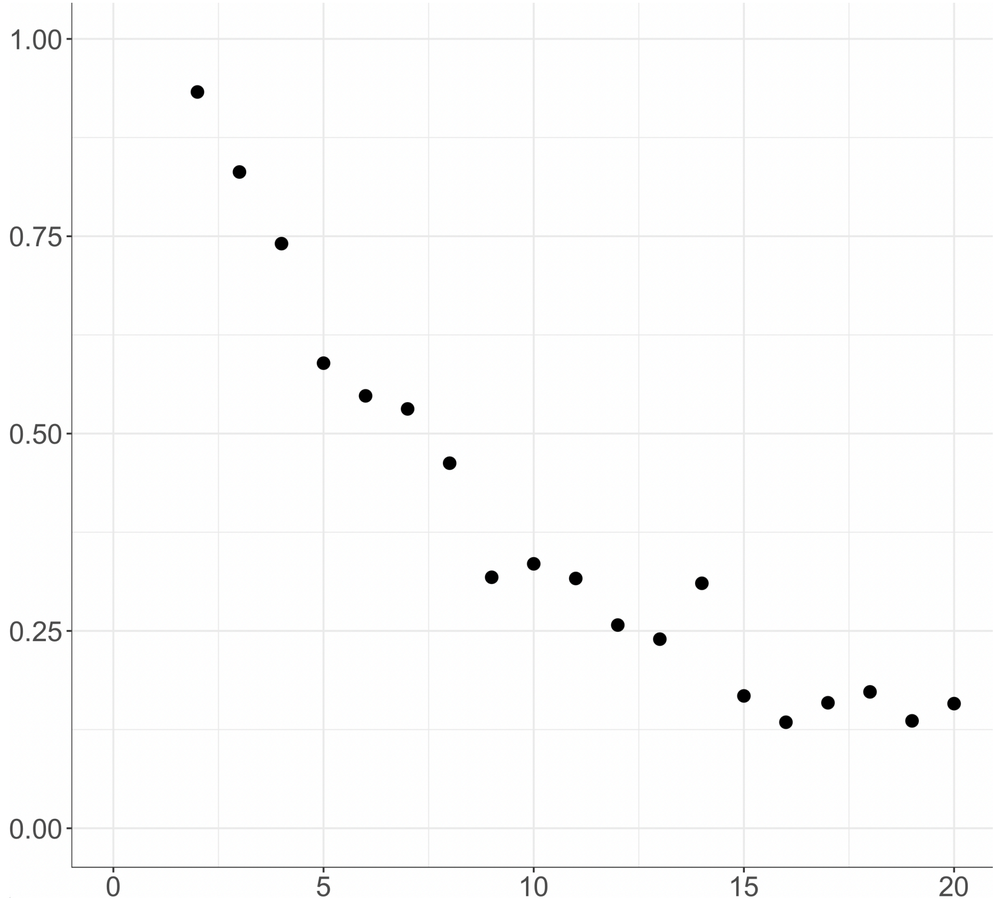
```
 x,    y
 2, 0.93
 3, 0.83
 4, 0.74
 5, 0.59
 6, 0.55
 7, 0.53
 8, 0.46
...
```



Statistical Model

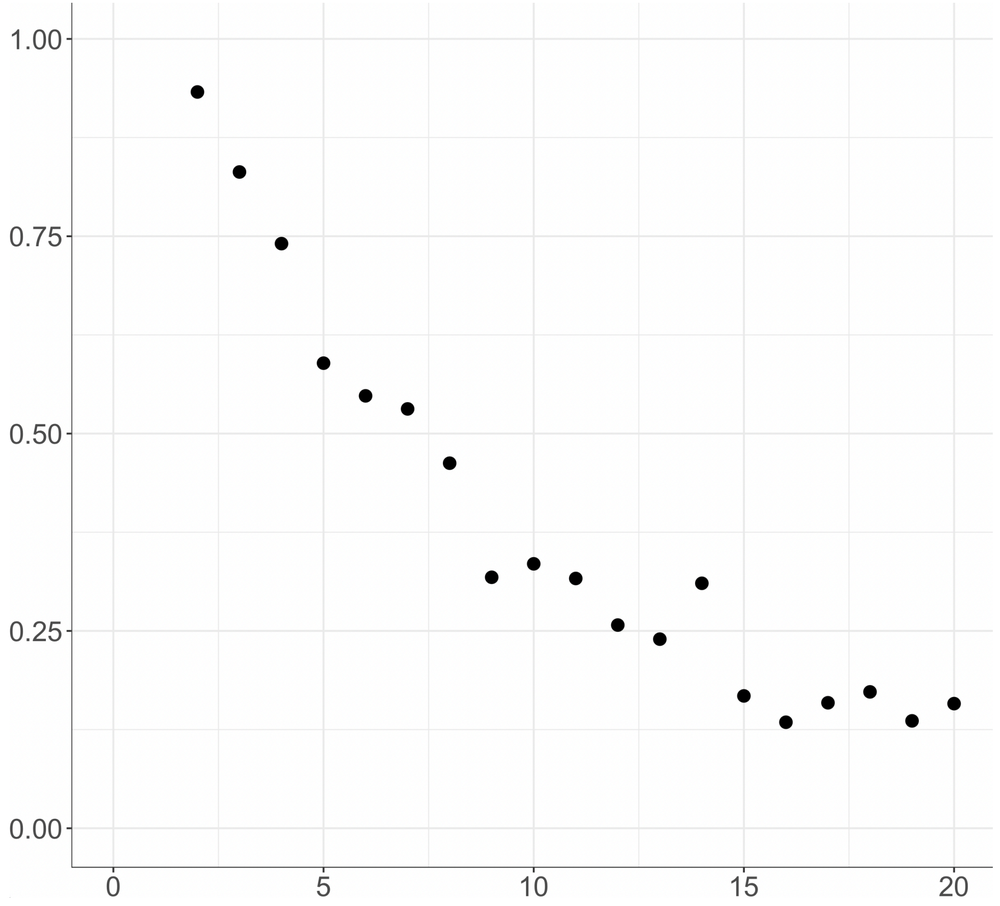
What is a statistical model?

- Set of **assumptions** that explain the data



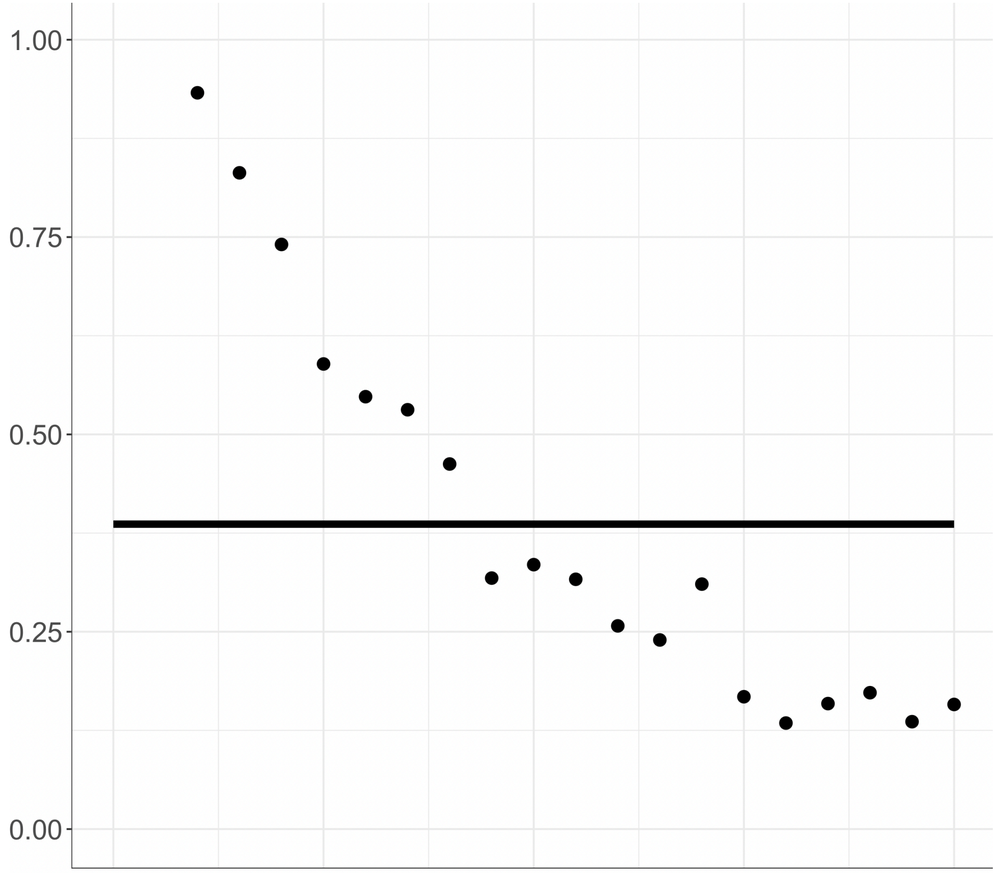
What is a statistical model?

- Set of **assumptions** that explain the data
  - Shape
  - Range
  - Data collection



What is a statistical model?

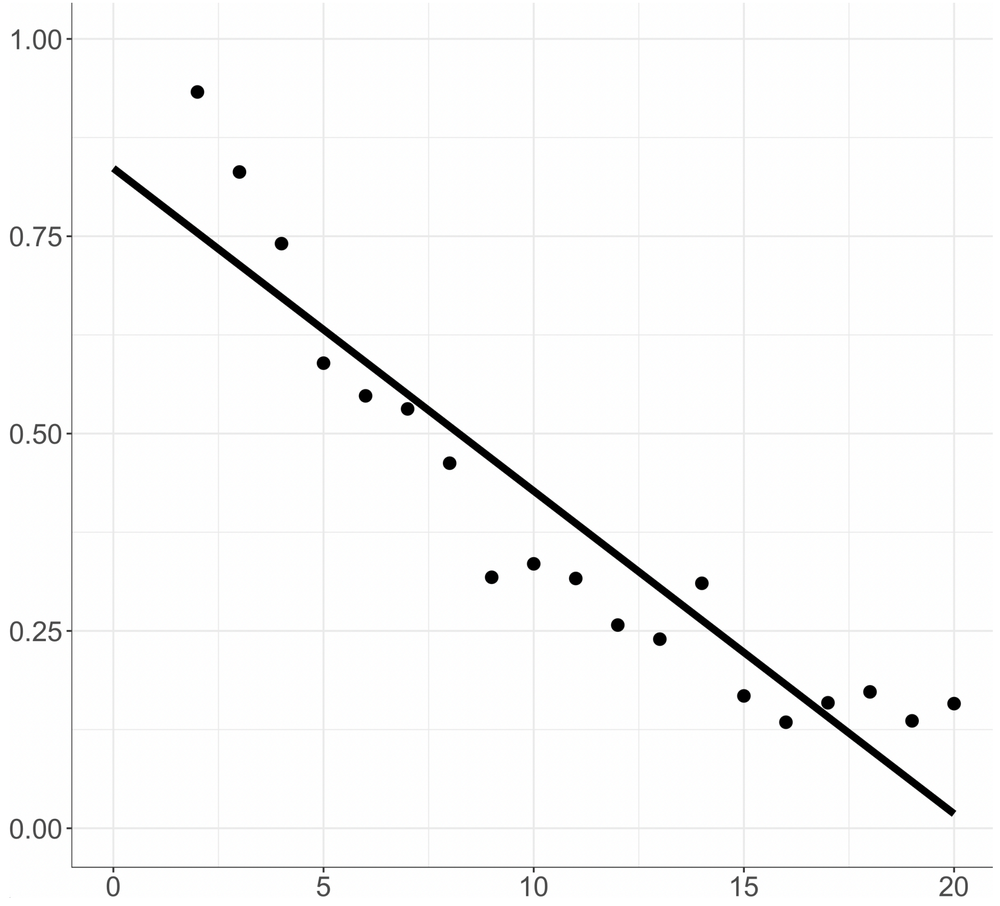
- Set of **assumptions** that explain the data
  - Average of the data
  - y = a + error
  - error ~ normal(**0**, sigma)



What is a statistical model?

Linear model

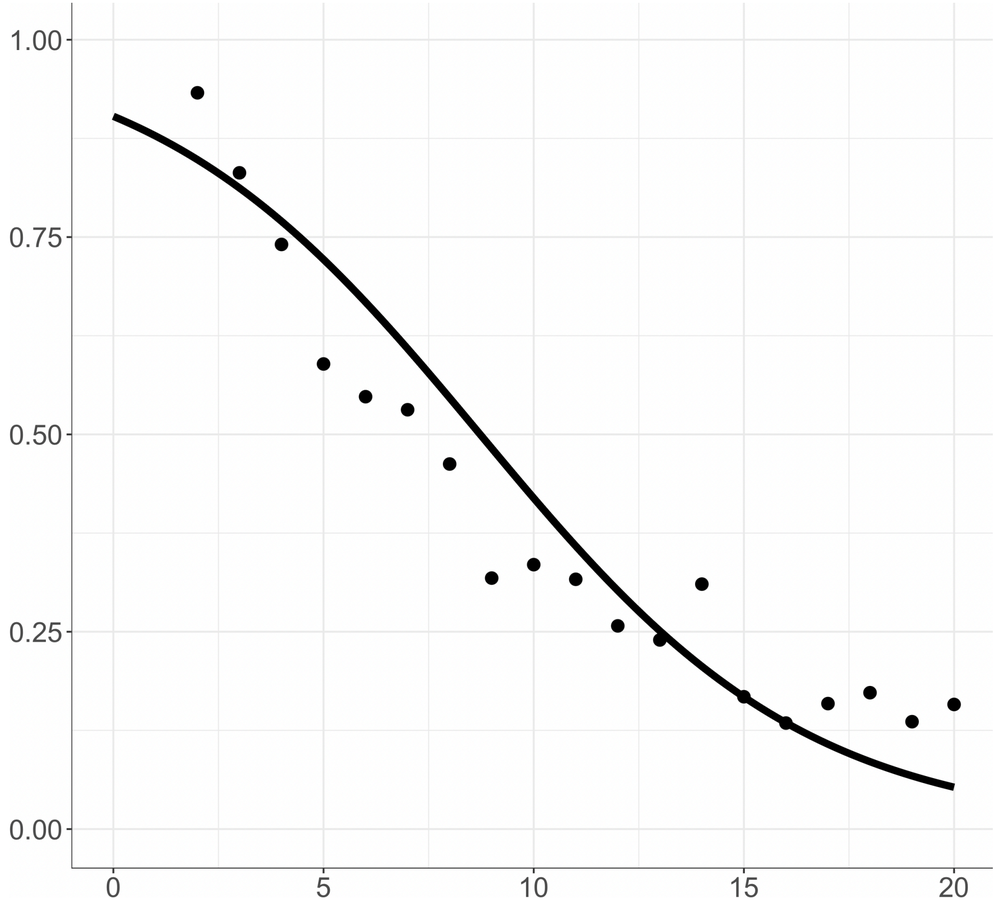
- Set of **assumptions** that explain the data
  - Straight line
  - y = a + b * x + error
  - error ~ normal(**0**, sigma)



What is a statistical model?

Logistic model

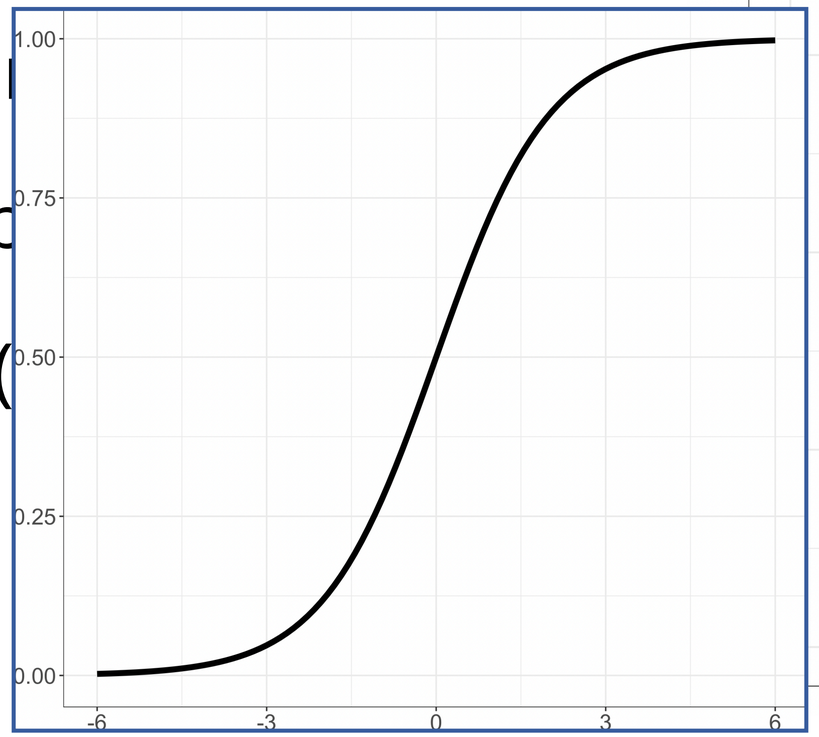
- Set of **assumptions** that explain the data
  - y is between 0 and 1
  - y = inv.logit(a + b * x) + error
  - inv.logit(x) = 1 / (1 + exp(x))

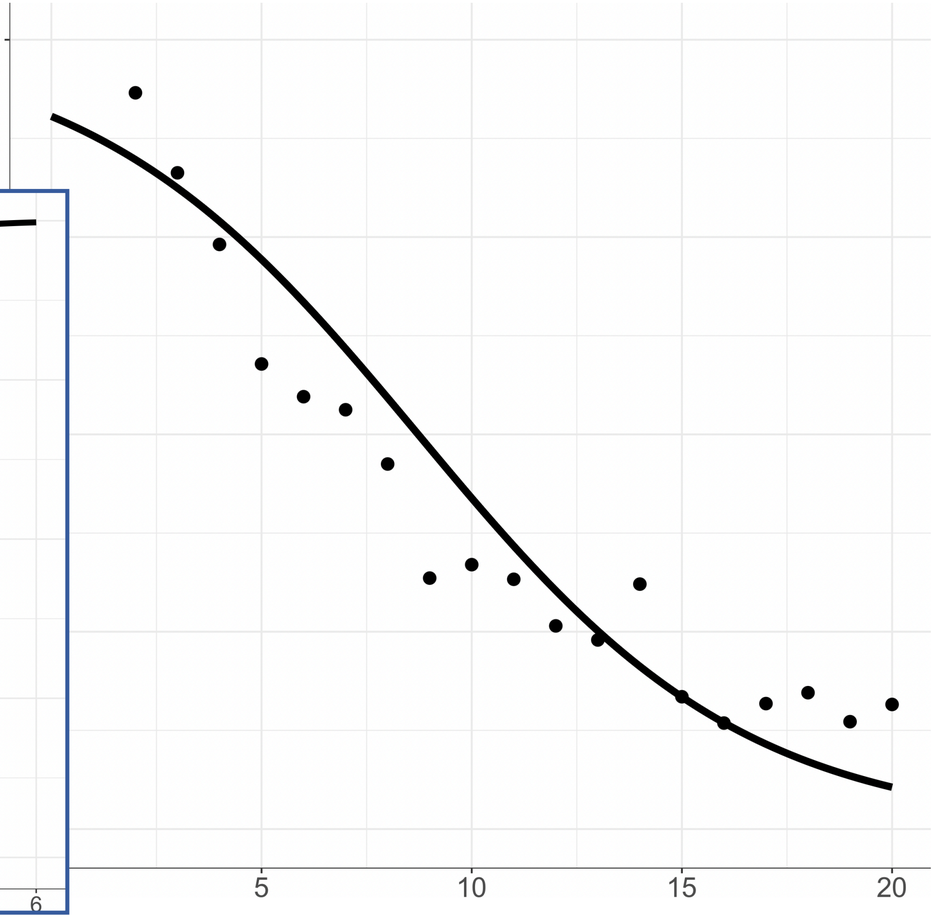


What is a statistical model?

Logistic model

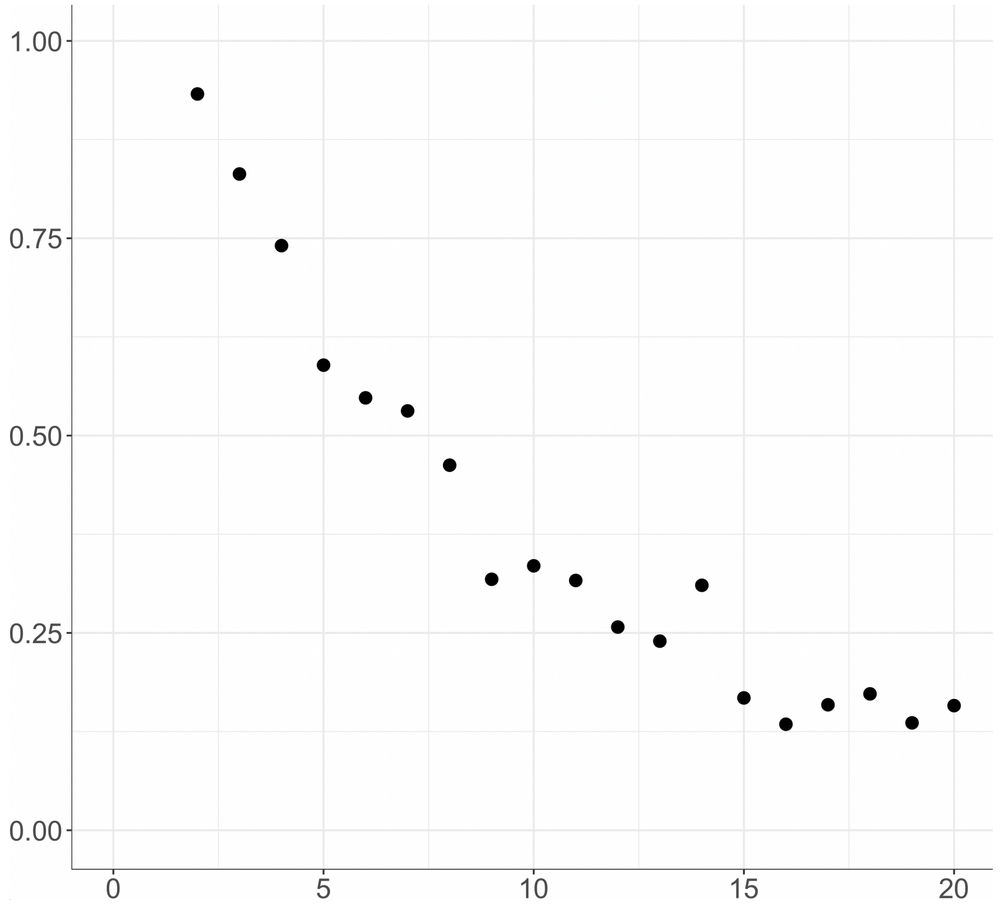
- Set of **assumptions** that explain the data
  - y is between 0 a
  - y = inv.logit(a + b
  - inv.logit(x) = 1 / (





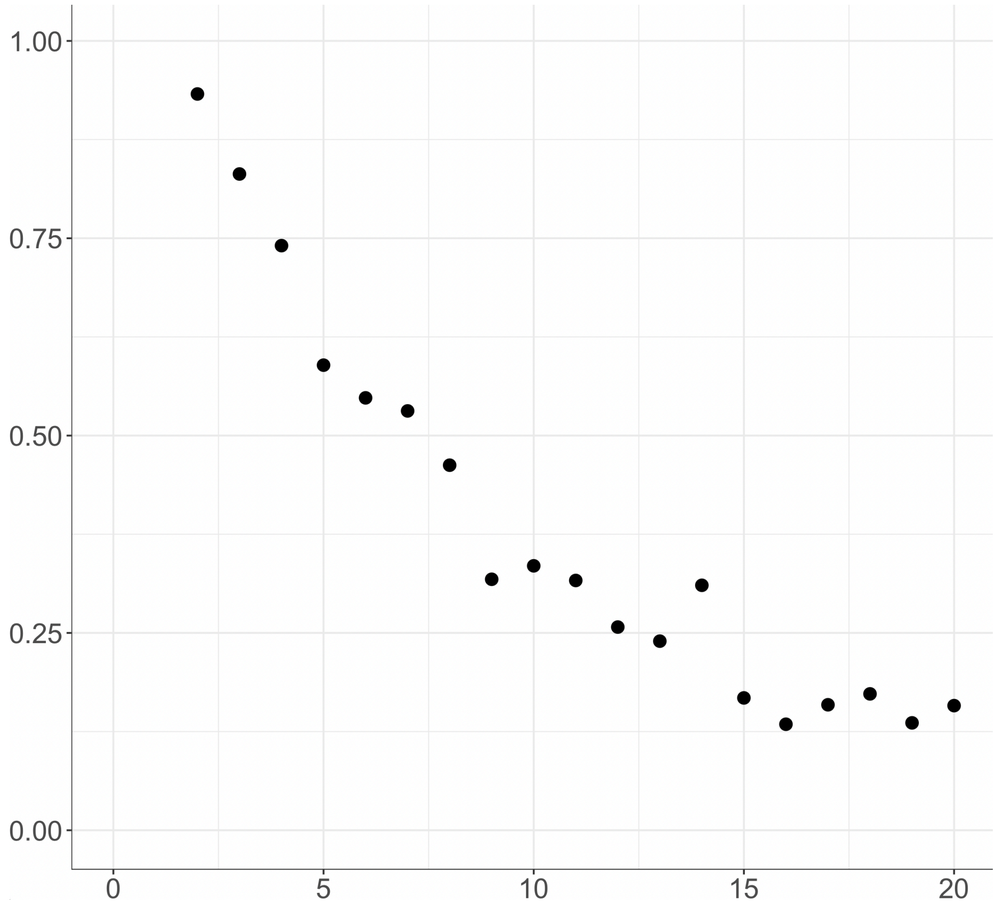
Are these models appropriate?

- Depends
  - What do we know about the data?



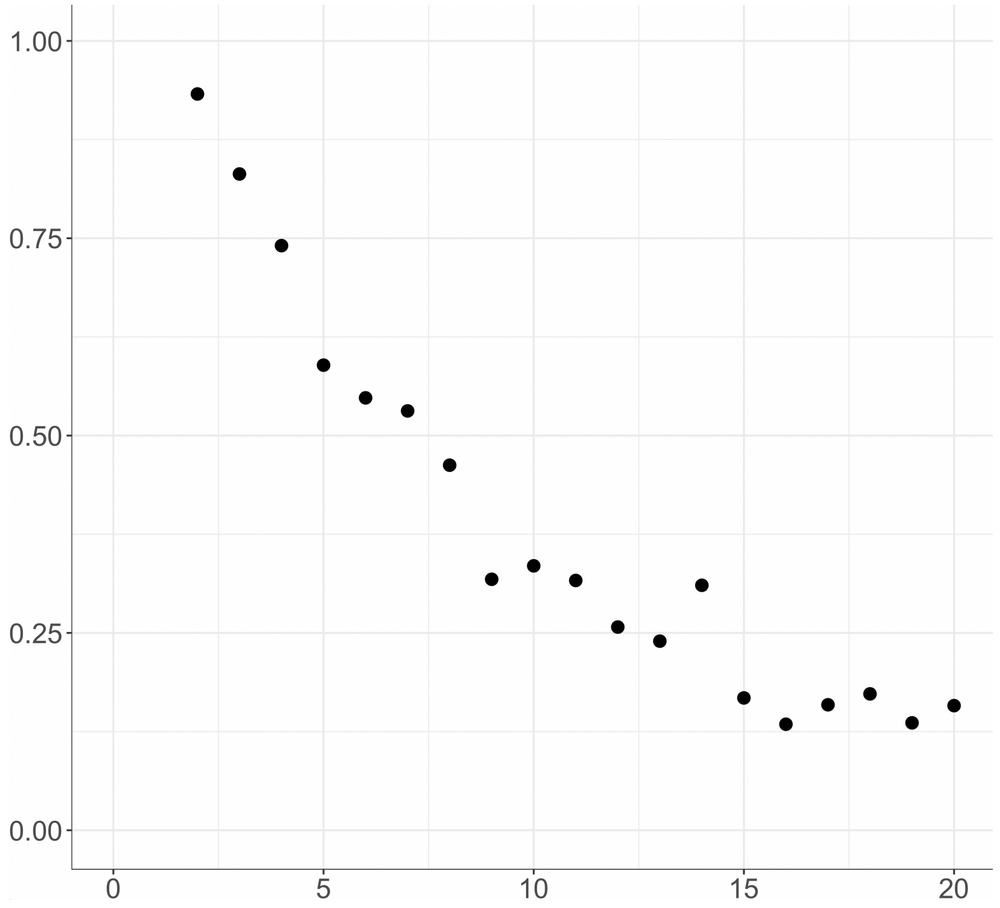
Are these models appropriate?

- Depends
  - What do we know about the data?
  - Do you think I generated this data?



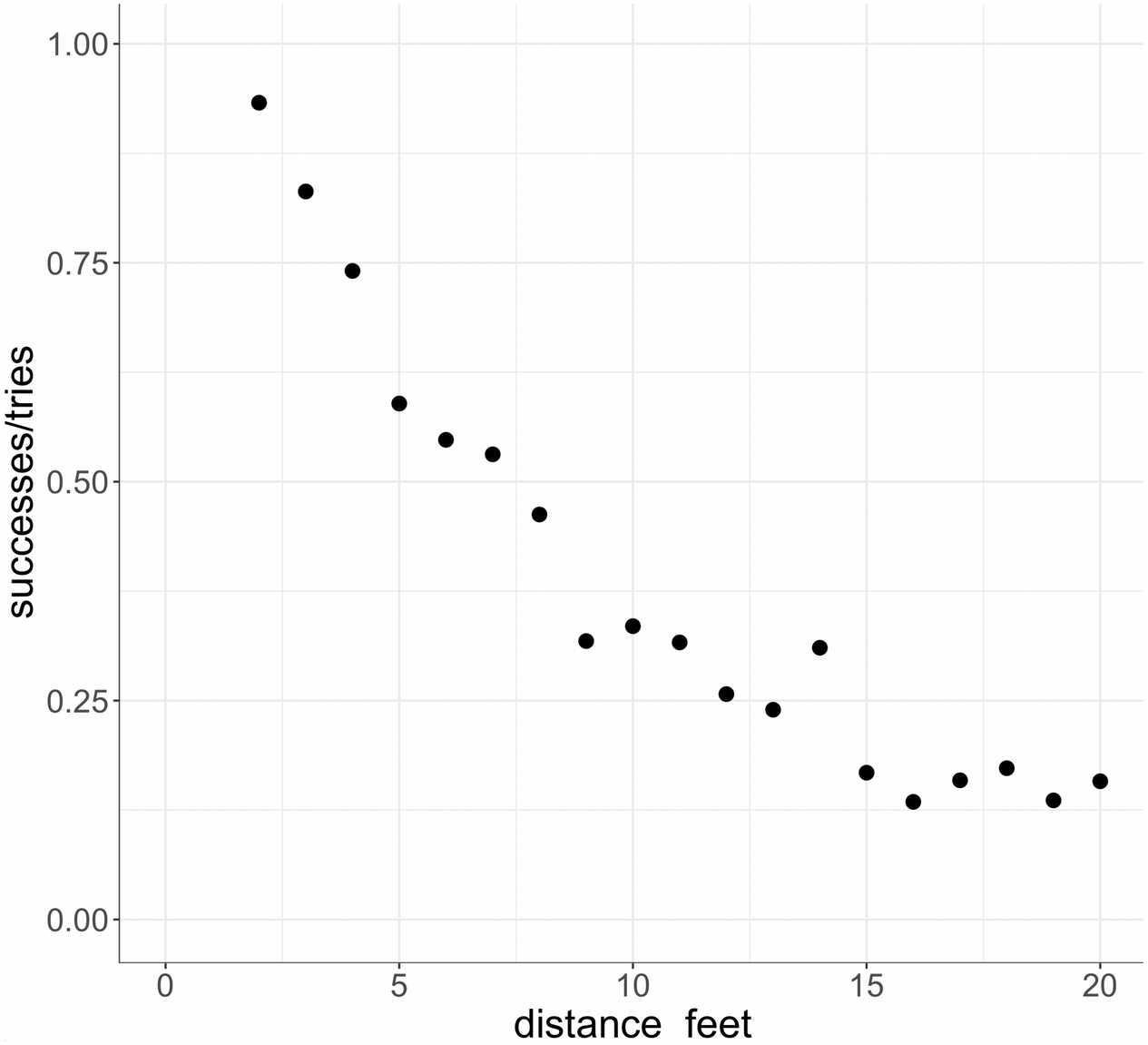
Are these models appropriate?

- Depends
  - What do we know about the data?
  - Do you think I generated this data?
  - What sport does this data come from?



Data

- PGA putting data
  - **Summary data**
  - x: distance in feet
  - y: frequency of makes at that distance
    - successes, attempts



Who makes the assumptions?

Who makes these assumptions?

- You. Me.
- Tool to build statistical models: Stan

Who makes these assumptions?

- You. Me.
- Tool to build statistical models: Stan 
- Other PPLs: PyMC, TensorFlow Probability, NumPyro, …

Inference

Notation

- Data: x
- Statistical model: $p(x, \theta)$
  - Parameters: θ (theta)

Notation

- Data: x
- Statistical model: $p(x, \theta)$
  - Parameters: θ (theta)
  - Joint probability distribution function: $p(x, \theta)$

Notation

- Data: x
- Statistical model: $p(x, \theta)$
  - Parameters: θ (theta)
  - Joint probability distribution function: $p(x, \theta)$

    Usually: $p(x, \theta) = p(\theta) \times p(x \mid \theta)$
  - Prior $p(\theta)$

Notation

- Data: x
- Statistical model: $p(x, \theta)$
  - Parameters: θ (theta)
  - Joint probability distribution function: $p(x, \theta)$

    Usually: $p(x, \theta) = p(\theta) \times p(x \mid \theta)$
  - Prior $p(\theta)$
  - Likelihood $p(x \mid \theta)$

Inference Types

- Frequentist, optimization:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(x, \theta)$$

- Bayesian inference: $p(\theta \mid x)$

approximated with $p(\theta \mid x) \approx \{\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots\}$

- Approximate Inference: replace problem with easier. Example: Variational Inference

$$p(\theta \mid x) \approx q(\hat{\phi})$$

where $$\hat{\phi} = \underset{\phi}{\operatorname{argmax}}\ \operatorname{KL}(q(\phi) \mid\mid p(x, \theta))$$

Hands-On: Linear model

Load Data; Convert for Stan

```
data <- read.csv("golf_data.csv")
stan_data <-
  list(N = nrow(data),
       distance_feet = data$distance_feet,
       tries = data$tries,
       successes = data$successes)
```

Data Block: golf_1.stan

```
data {
  int<lower = 0> N;
  array[N] real<lower = 0> distance_feet;
  array[N] int<lower = 0> tries;
  array[N] int<lower = 0> successes;
}
transformed data {
  vector[N] y = to_vector(successes) ./ to_vector(tries);
}
```

Parameters / Model: golf_1.stan

```
data {...}
parameters {
  real a;
  real b;
  real<lower = 0> sigma;
}
model {
  y ~ normal(a + b * distance_feet, sigma);
}
```

Fit the model

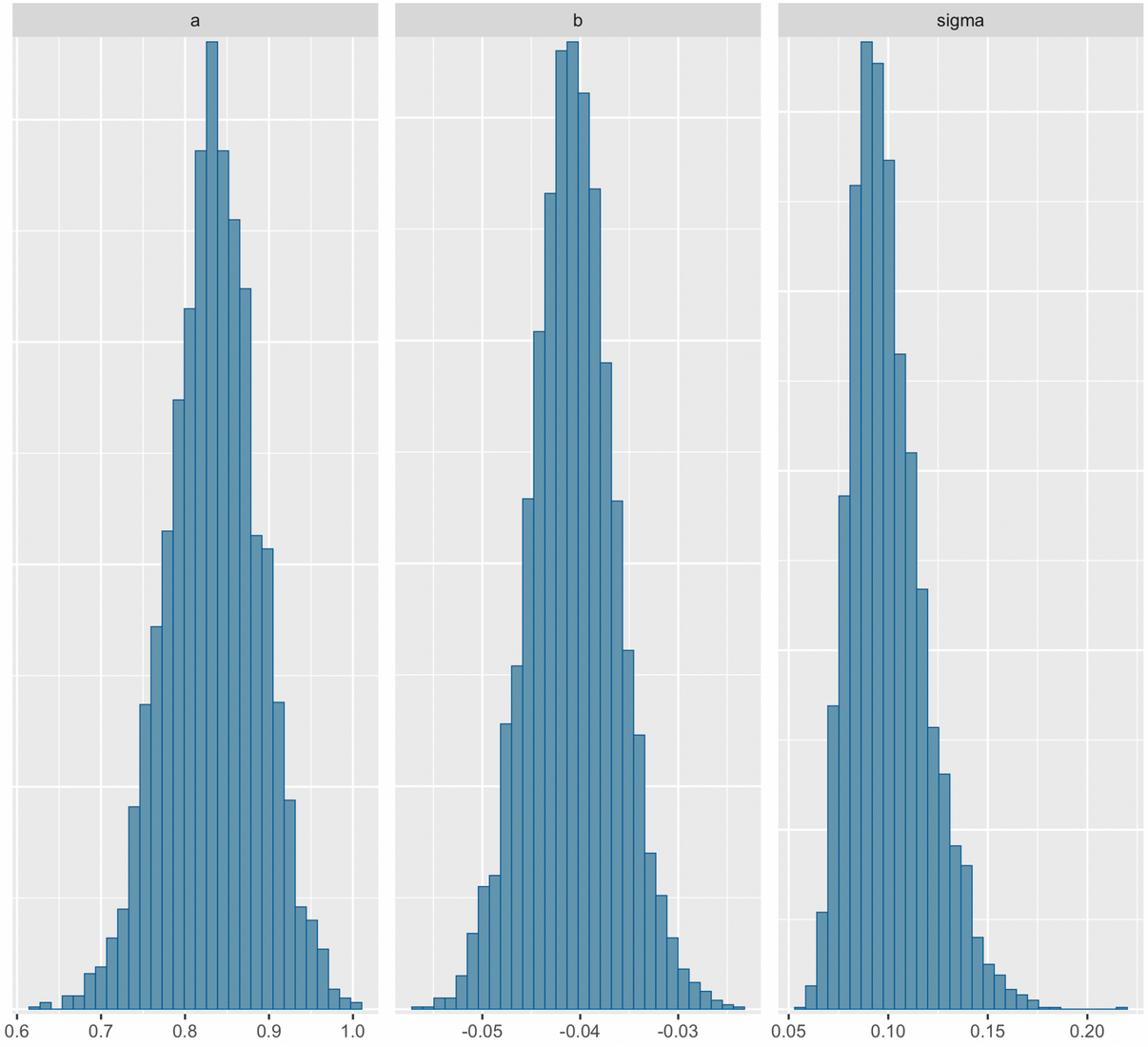
```
model <- cmdstanr::cmdstan_model(“golf_1.stan")
fit_mcmc <- model$sample(data = stan_data)
fit_optim <- model$optimize(data = stan_data)
fit_advi <- model$variational(data = stan_data)
```

What does it look like?

```
bayesplot::mcmc_hist(fit_mcmc$draws(c('a',
'b', ‘sigma')))
```

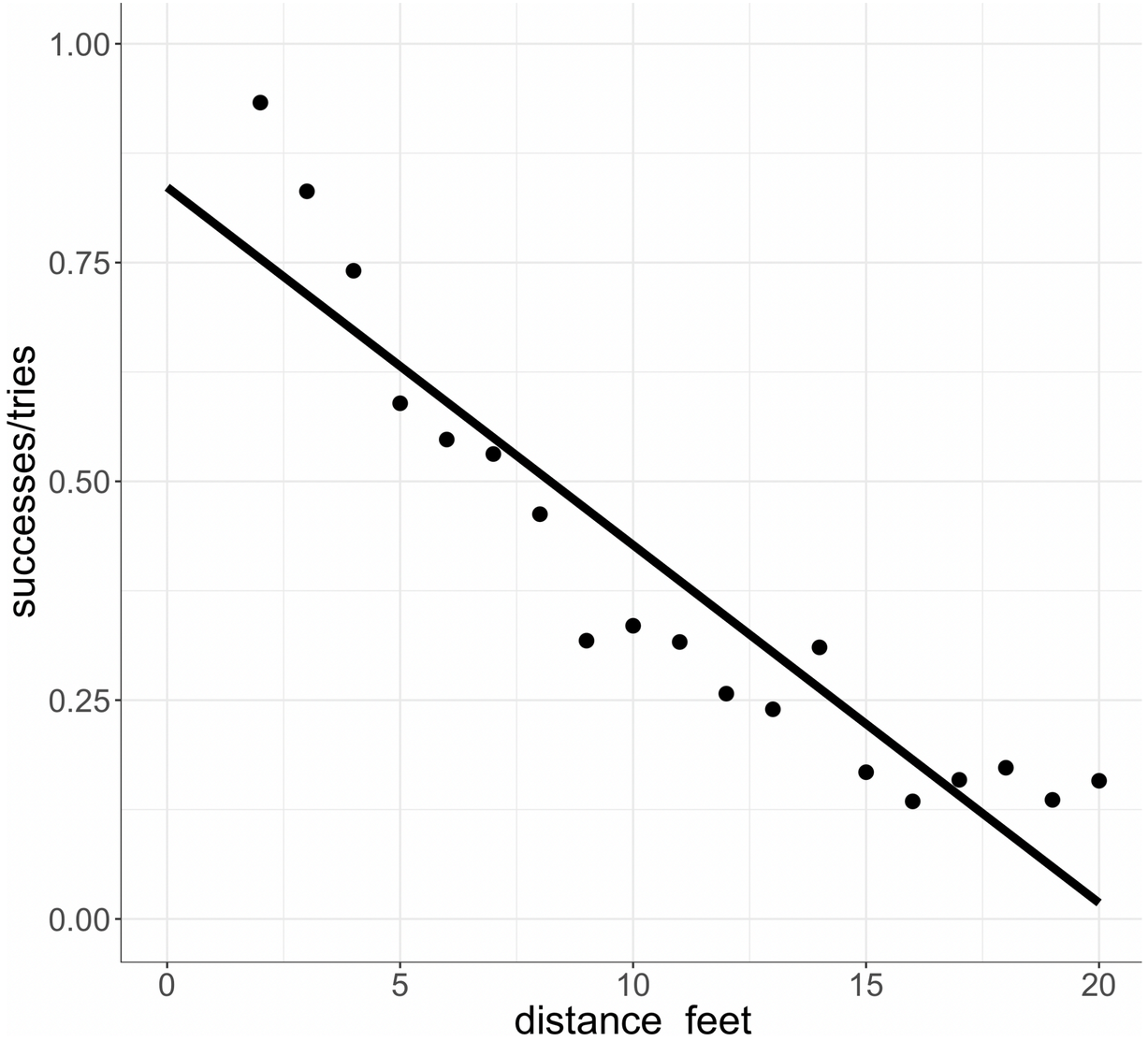
```
fit_optim$mle(c('a', 'b', ‘sigma'))
```

```
         a           b       sigma
0.8360870 -0.0408878  0.0867286
```



What does it look like?

```
(ggplot(data, aes(x = distance_feet, y = successes / tries))
  + geom_point(size = 3)
  + stat_function(fun = function(x) fit_optim$mle()['a'] +
fit_optim$mle()['b'] * x, linewidth = 2)
  + xlim(0, 20) + ylim(0, 1)
  + theme_bw()
  + theme(plot.margin = margin(t = 0, b = 0),
          text = element_text(size = 20))
  + theme(axis.title.x = element_blank(),
  axis.title.y = element_blank())
)
```



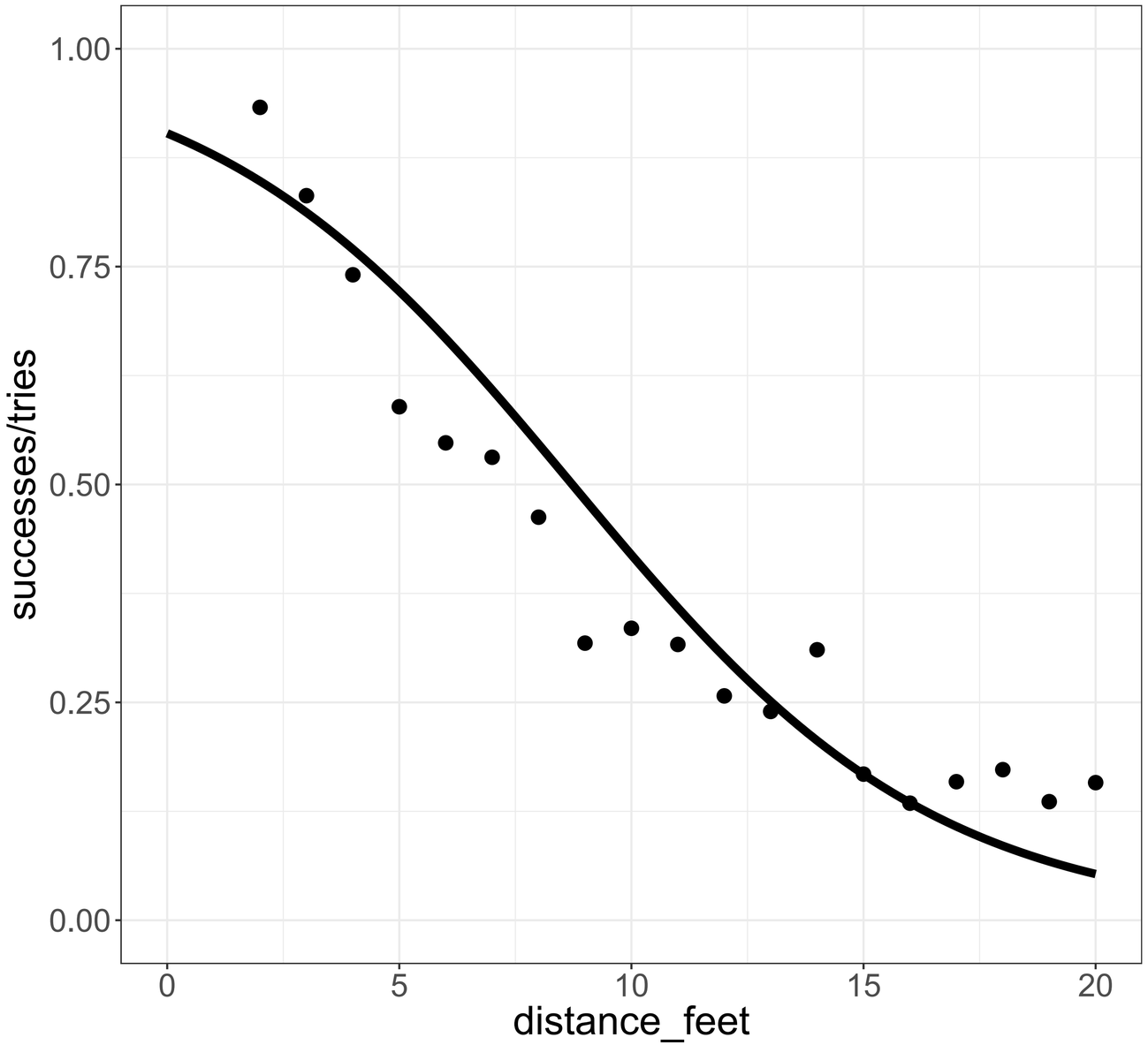
Hands-On:
Logistic model

Let’s go quick

```
data {
  int<lower = 1> N;
  vector<lower = 0>[N] distance_feet;
  array[N] int tries;
  array[N] int successes;
}
parameters {
  real a;
  real b;
}
model {
  successes ~ binomial_logit(tries, a + b * distance_feet);
}
```

Let's go quick

```
data {
  int<lower = 1> N;
  vector<lower = 0>[N] distance_feet;
  array[N] int tries;
  array[N] int successes;
}
parameters {
  real a;
  real b;
}
model {
  successes ~ binomial_logit(tries, a + b *
distance_feet);
}
```



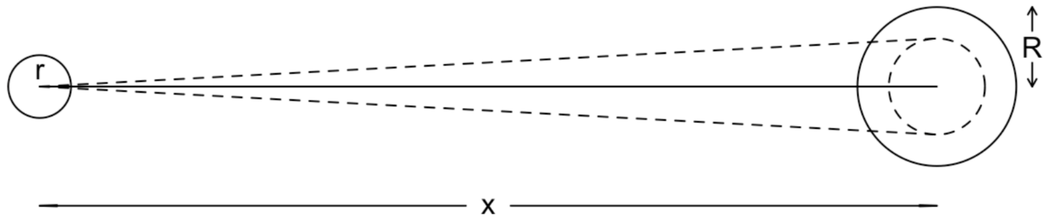
Can we do better?

Geometry

r: radius of the golf ball

R: radius of the hole

x: distance to hole



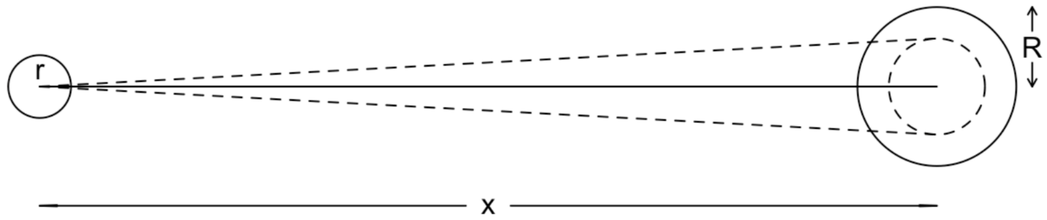
https://mc-stan.org/users/documentation/case-studies/golf.html

Geometry

r: radius of the golf ball

R: radius of the hole

x: distance to hole



If angle is within a threshold, the ball goes in.

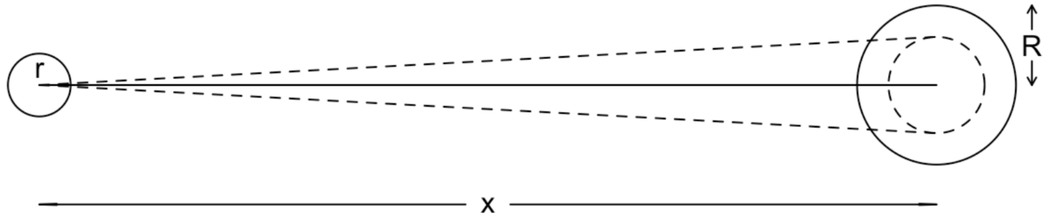
Threshold angle (radians): arcsin((R - r) / x)

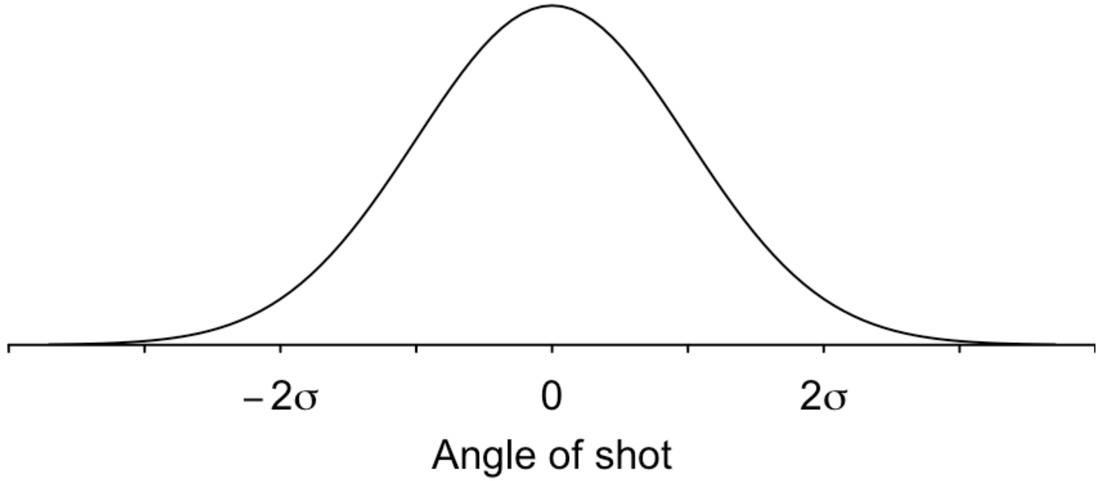
https://mc-stan.org/users/documentation/case-studies/golf.html

Modeling golfer's ability

Assume: Golfer wants to hit it straight.

Assume: angle error is normally distributed: mean = 0, standard deviation = sigma





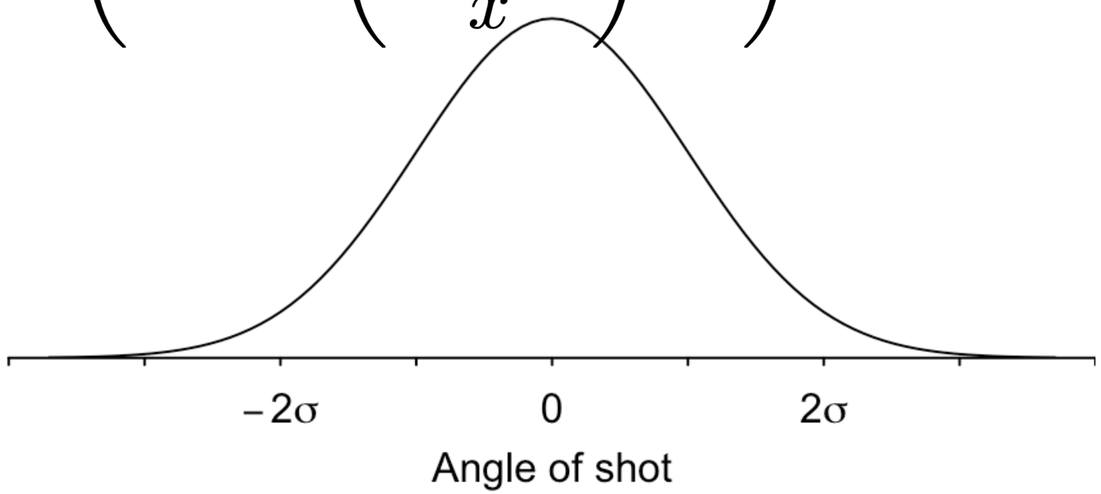
https://mc-stan.org/users/documentation/case-studies/golf.html

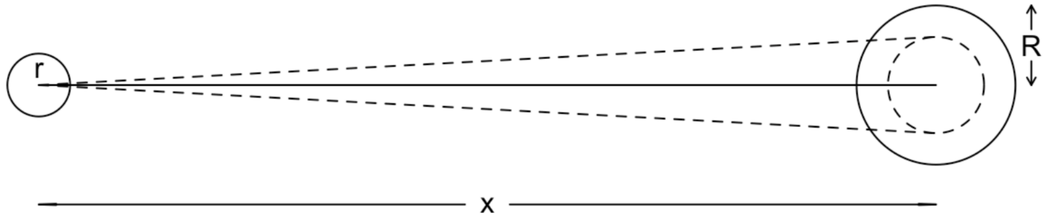
Probability of ball going in from distance x?

Pr(ball goes in | x) = Pr(angle < arcsin(...) AND angle > arcsin(...))

$$= \left[\Phi\left(\arcsin\left(\frac{R-r}{x}\right)/\sigma\right) - 0.5\right] + \left[0.5 - \Phi\left(\arcsin\left(\frac{R-r}{x}\right)\sigma\right)\right]$$

$$= 2 * \Phi\left(\arcsin\left(\frac{R-r}{x}\right)/\sigma\right) - 1$$





One parameter model

```
transformed data {
  real r = 1.68 / 12 / 2;
  real R = 4.25 / 12 / 2;
  vector[N] threshold_angle = asin((R - r) ./ x);
}
parameters {
  real<lower = 0> sigma;
}
```

https://mc-stan.org/users/documentation/case-studies/golf.html

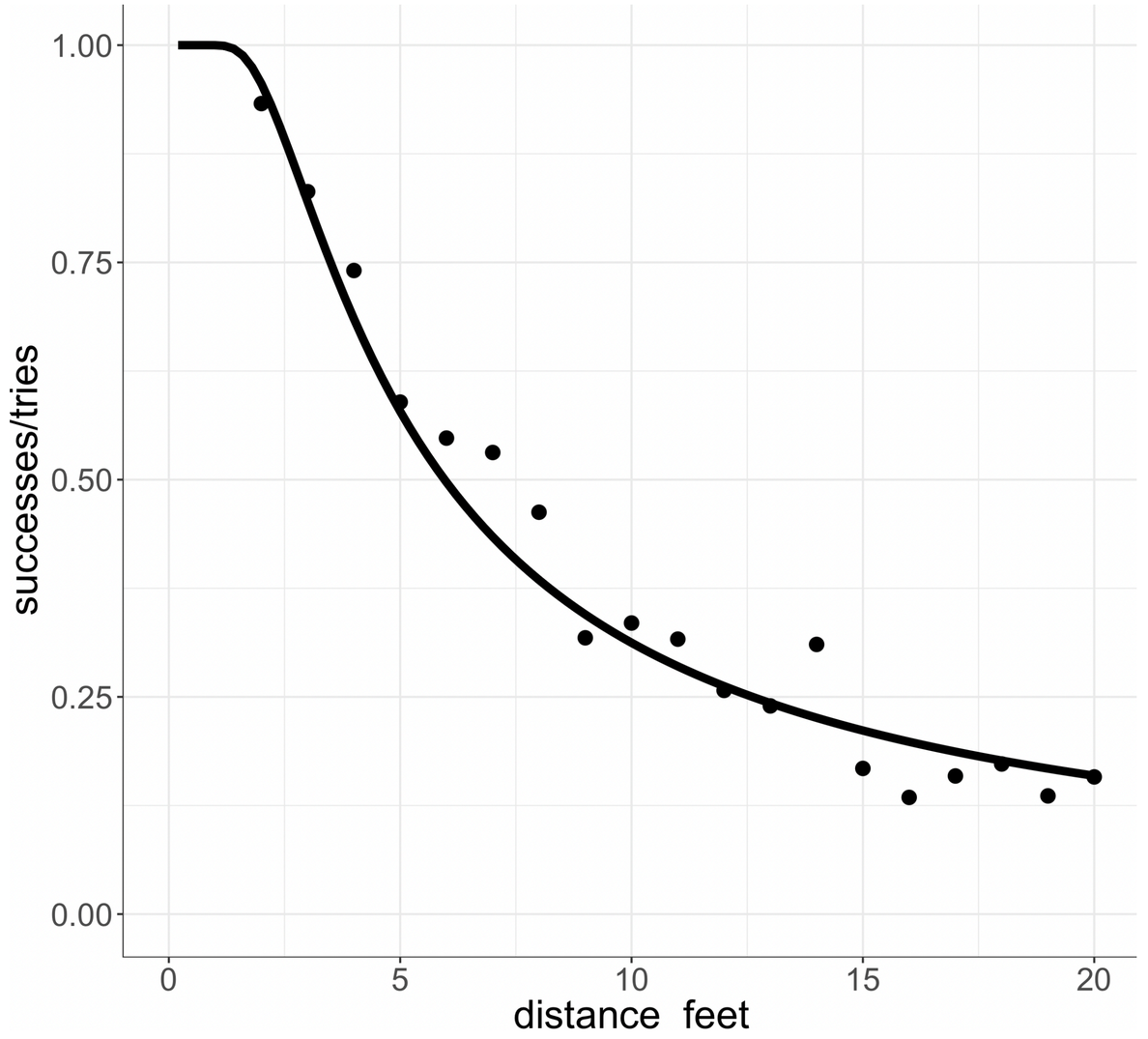
One parameter model

```
model {
  vector[N] p = 2 * Phi(threshold_angle / sigma) - 1;
  y ~ binomial(n, p);
}
generated quantities {
  real sigma_degrees = sigma * 180 / pi();
}
```

https://mc-stan.org/users/documentation/case-studies/golf.html

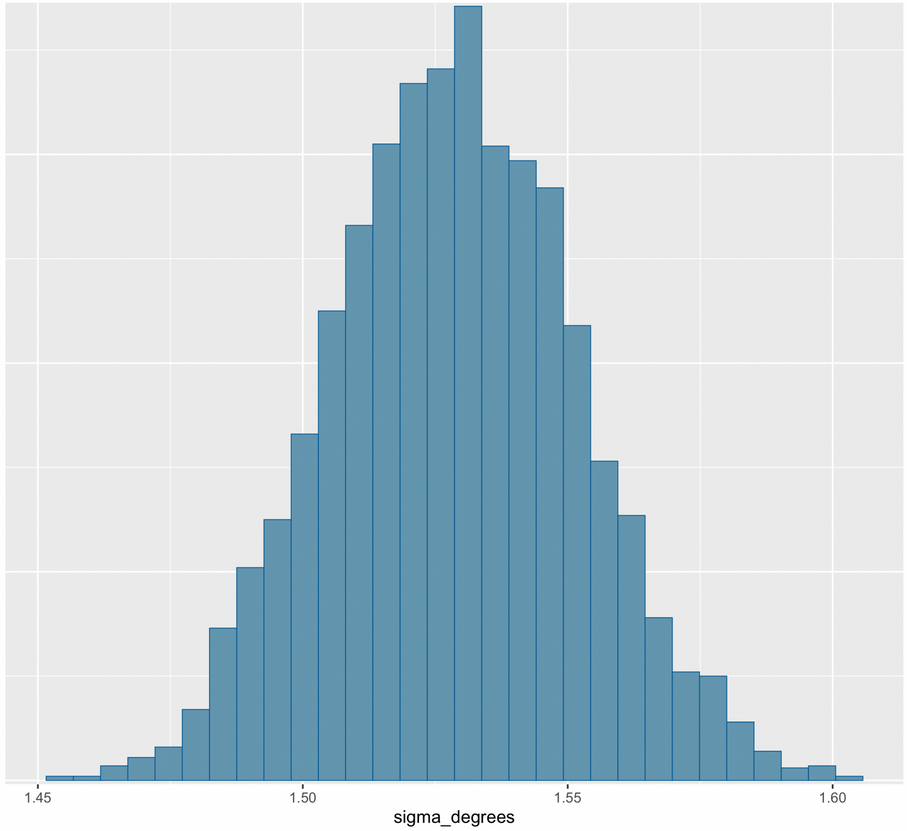
Look at the output

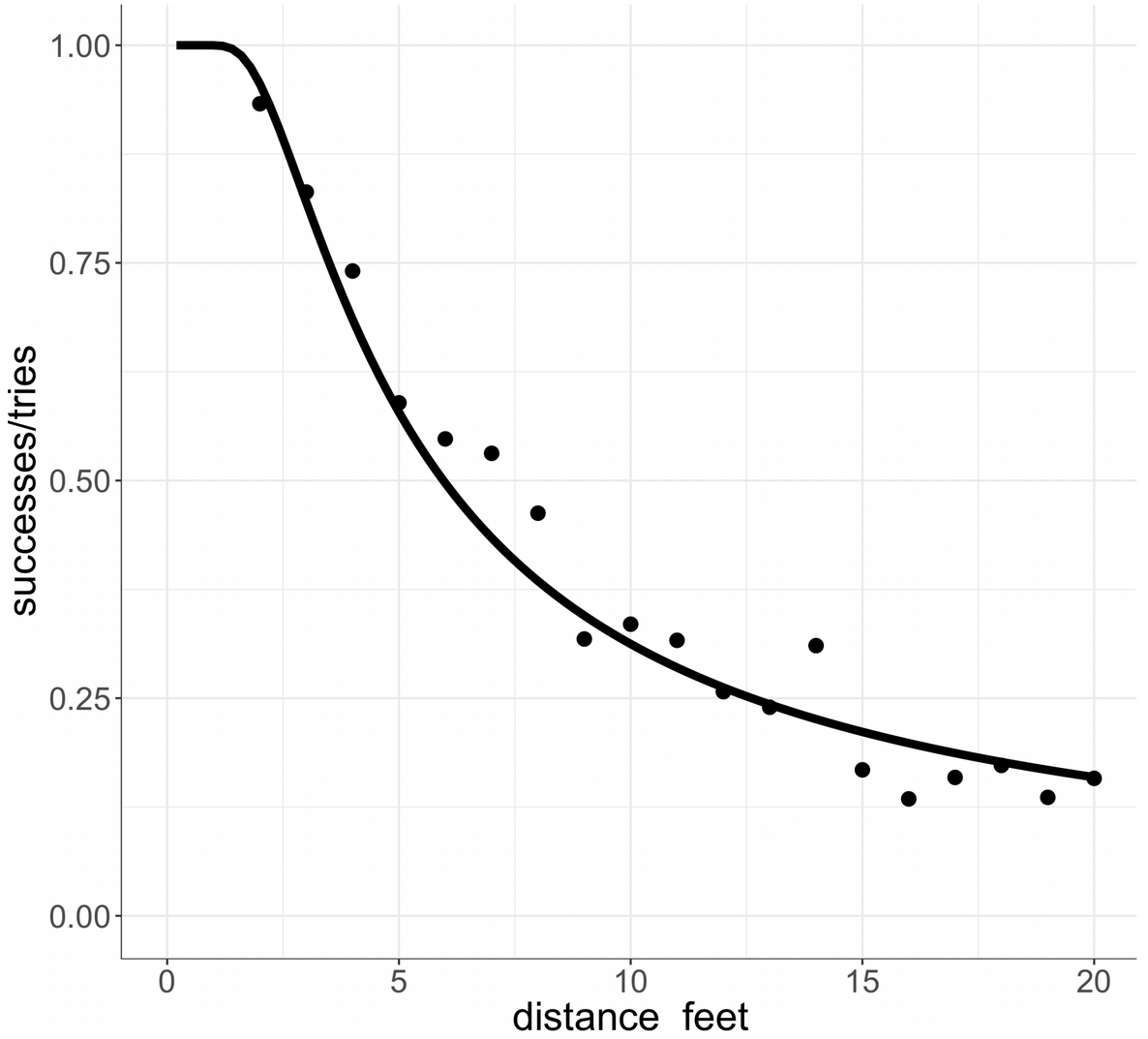
Do we like it?



Look at the output

Do we like it?





What are some of the assumptions?

- Geometry can explain the data
- Angle error is the only thing that matters

- What are we assuming doesn't have an effect?
  - Distance
  - Individual golfer ability (need finer data)
  - Green topography; course difficulty
  - Pressure; fatigue

https://mc-stan.org/users/documentation/case-studies/golf.html

When to use Bayesian Inference?

Recap

- Statistical model encodes a set of assumptions
- We have tools to write custom (and non-custom) models

When to use Bayesian inference?

- Low data relative to model complexity
  - Optimization is over-confident.
- Hierarchical models
  - Parameters are nested within groups
  - E.g. players within teams, players within roles
  - Optimization fails.

Thank you!!!

Q&A Time

Connect on LinkedIn

Daniel Lee

https://www.linkedin.com/in/syclik/

Kiran Gauthier

https://www.linkedin.com/in/kiran-gauthier/