PSYCH 201B

Statistical Intuitions for Social Scientists

Modeling data

You can download these slides:
course website > Week 5 > Overview

02/03/2025

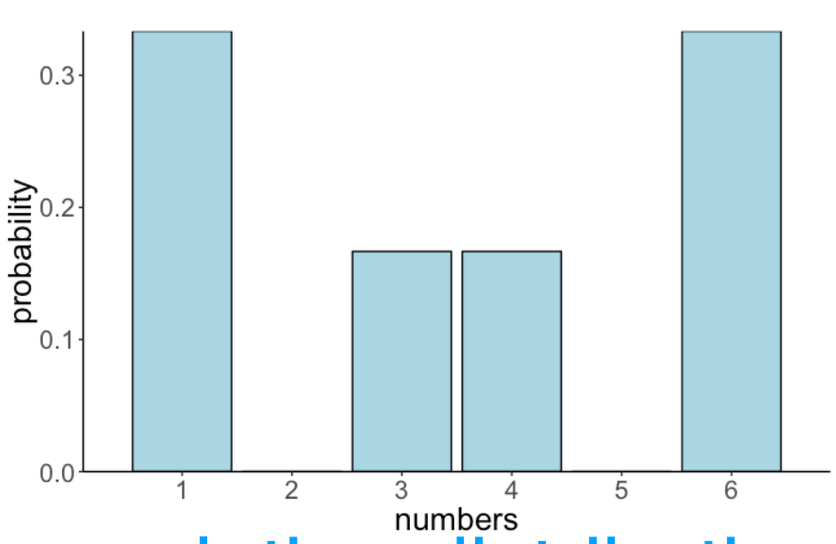
Announcements

- Keep working on HW2: **due in 1 week**
- *More* cheatsheets on the website
  - Common formulas
- This week
  - Give you a break from new Python libraries (until next week)
  - Back to slides (today)
    - A little bit of review
    - Introduction to **modeling data**

Our journey so far...

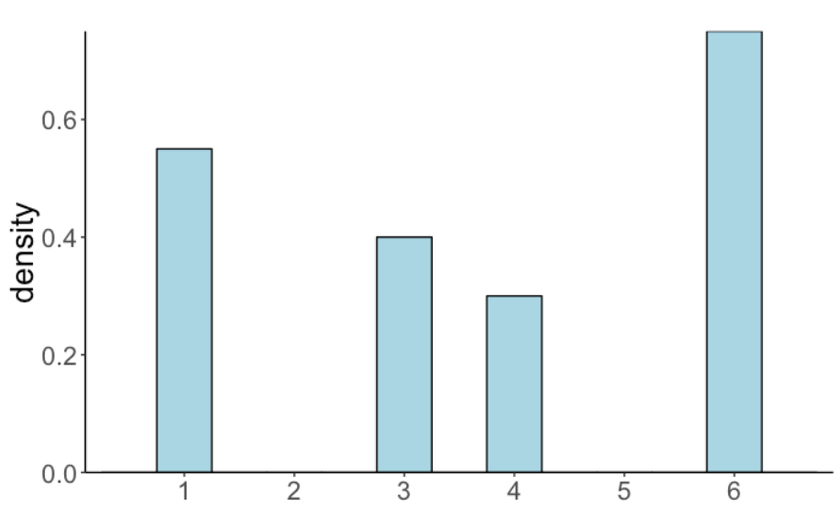
Statistical inference via **Central Limit Theorem**

The process of making claims about a population based on information from a sample.



sampling

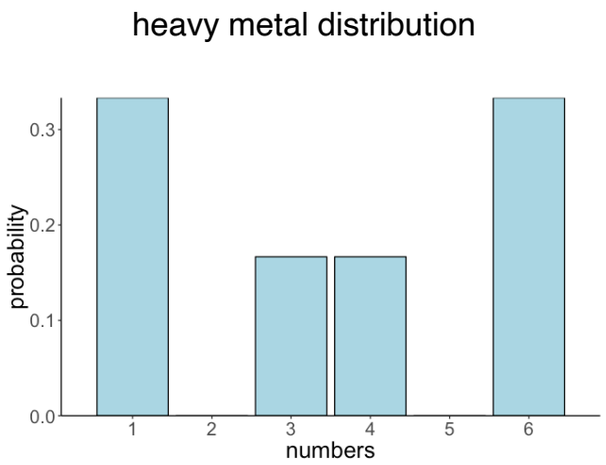
inference



population distribution

our sample

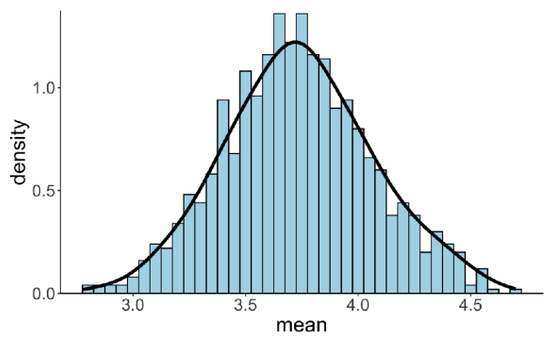
3 distributions in statistical inference



population

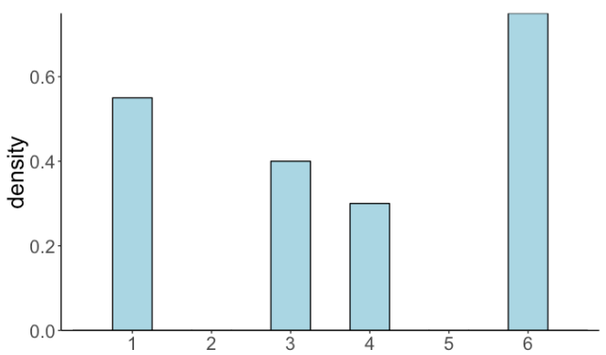
- unknown
- our target for inference
- e.g. we might be interested in the mean of the population distribution

---



sampling distribution

- bridge between sample and population
- derived mathematically / computationally
- asymptotic distribution theory or resampling approaches
- shows how test statistic varies between samples



sample distribution

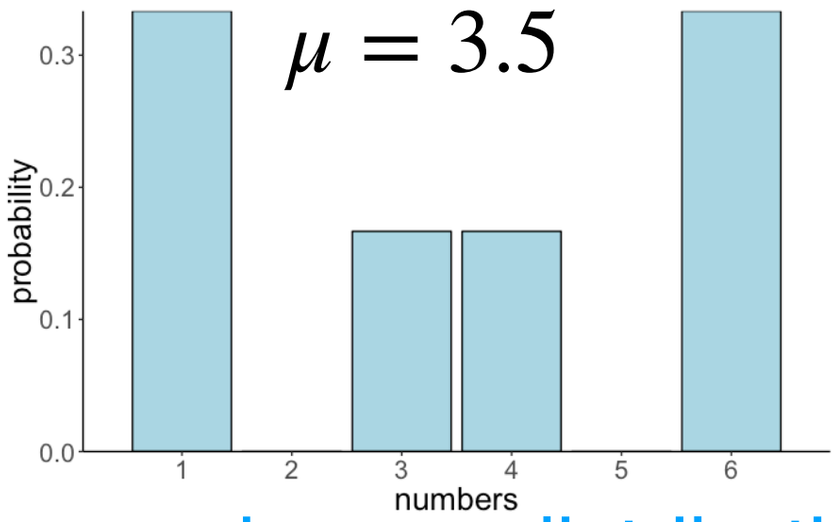
- our observed sample
- we compute statistics of interest (mean, variance, correlation, ...)
- make an inference about the population via the sampling distribution

Statistical inference

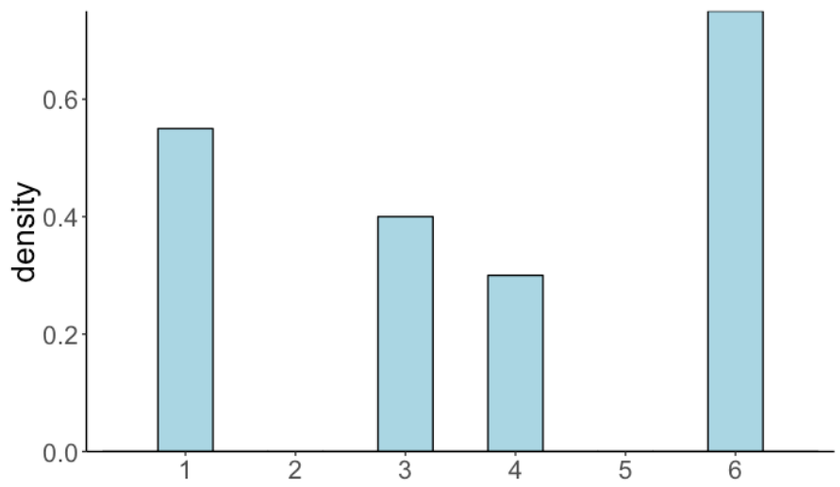
what's the population mean?

```
sample mean = 3.725
standard deviation = 2.05
n = 40
```

heavy metal distribution

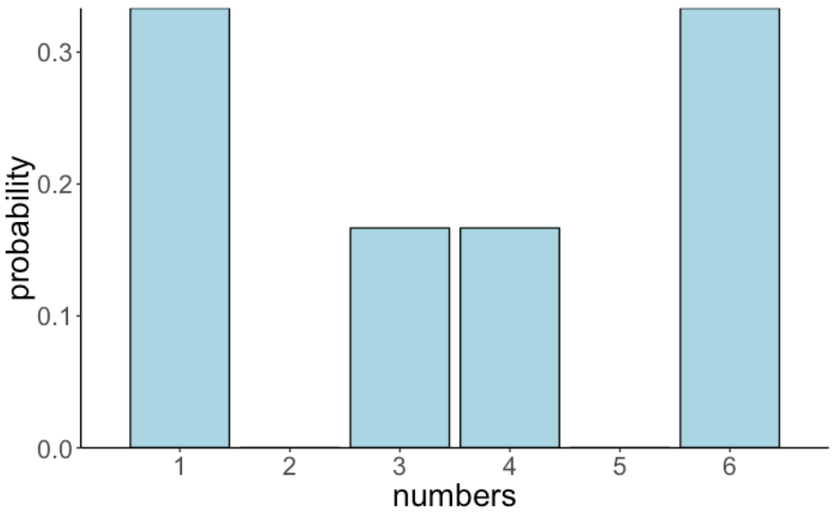


true unknown distribution

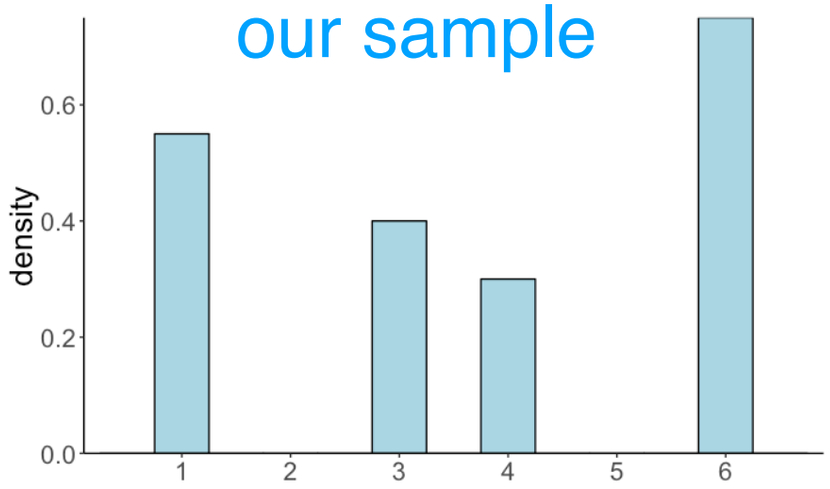


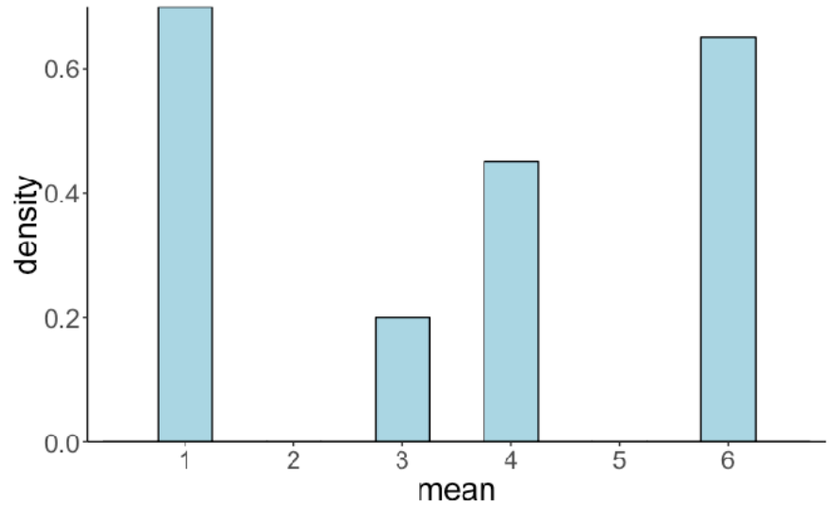
our sample

Sampling variation

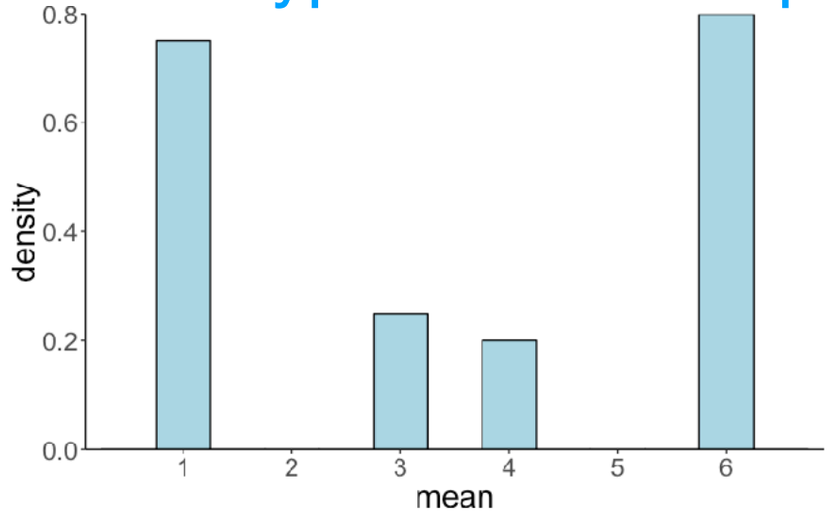


population distribution



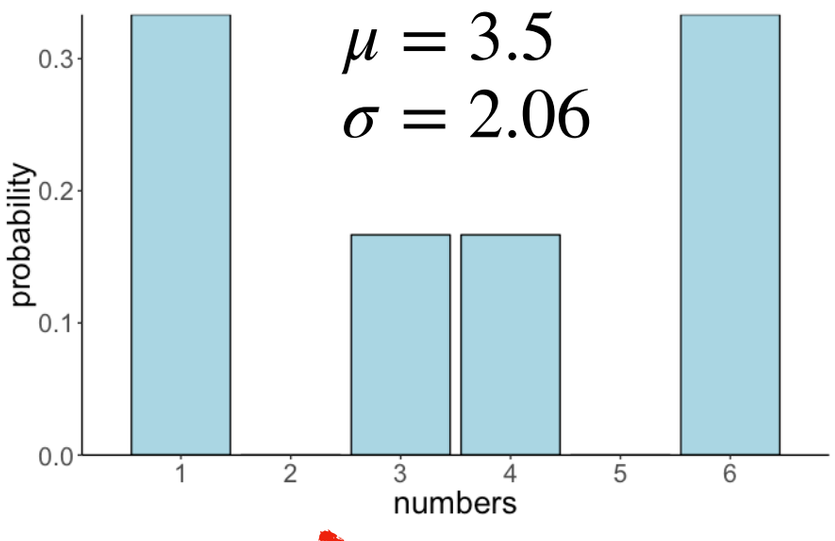


another hypothetical sample

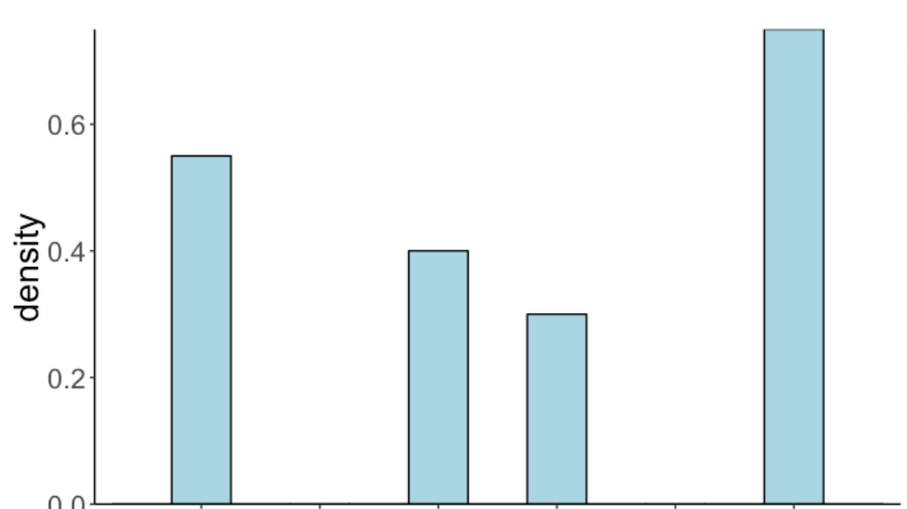


Sampling distribution

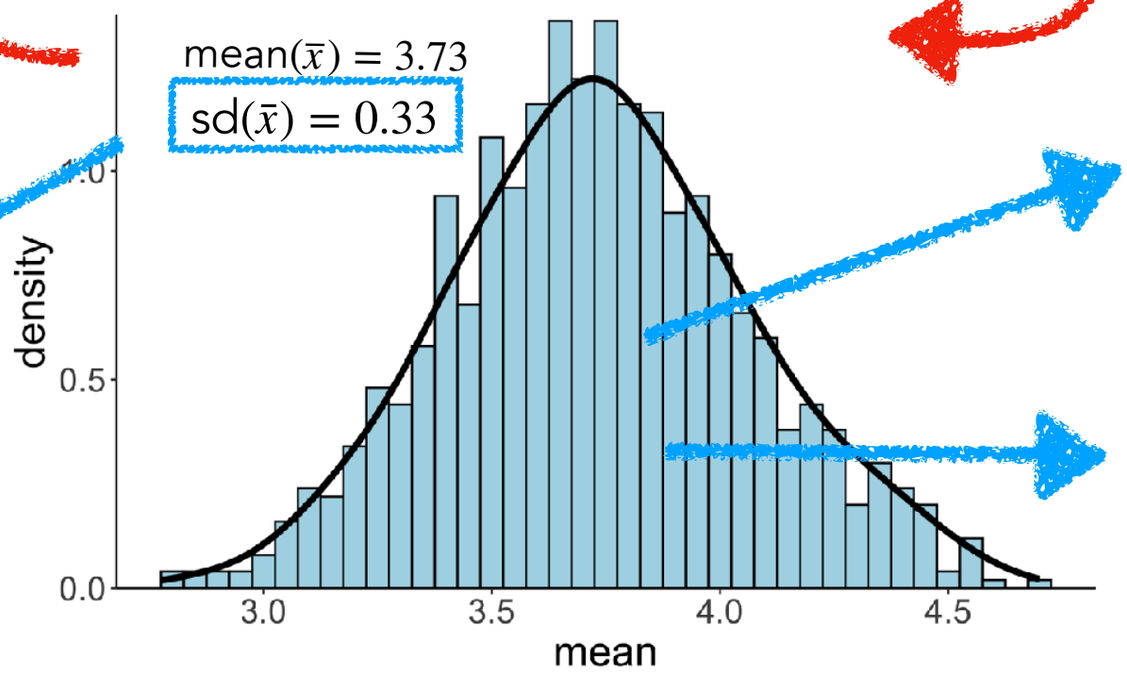
population distribution



our sample



standard error

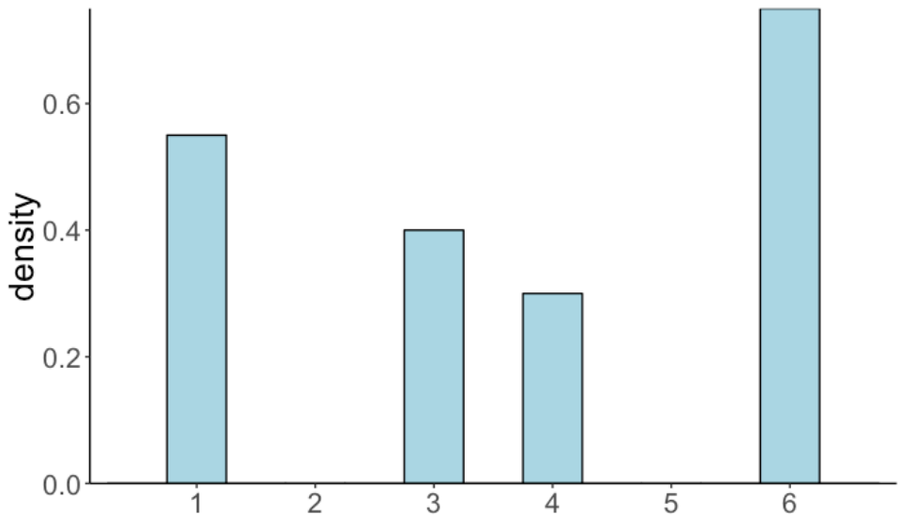


p-values

confidence intervals

sampling distribution of the mean

our sample

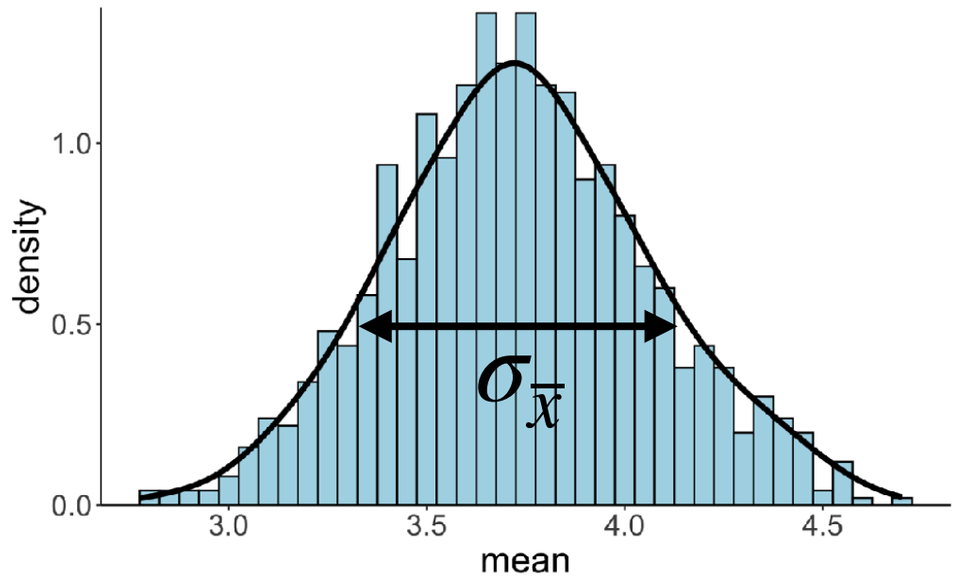


standard deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N - 1}}$$

gives a sense for how well the mean summarizes the data

sampling distribution



standard error

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

the standard deviation of the sampling distribution

how much variation would we expect between the means of different samples

how likely is it that our sample mean is representative of the population mean?

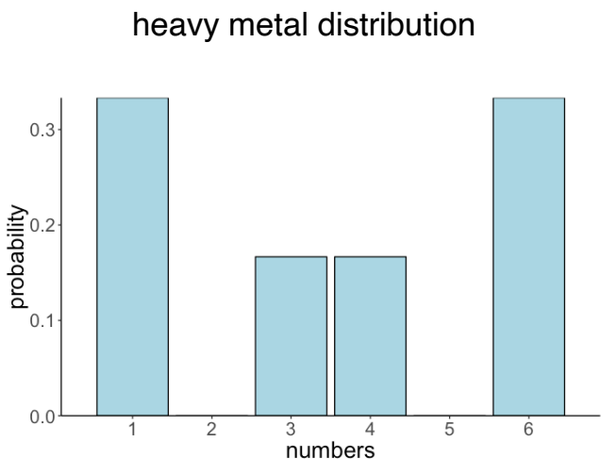
Fundamental concepts of statistics

- **Aggregation**
  - describe and *compress* the data into a summary
- **Sampling**
  - facilitate *generalizing* to unseen data
- **Uncertainty**
  - quantifying *trust* in our estimates based on different sources of error
- **Learning**
  - using data to *update* our estimates

Fundamental concepts of statistics

- **Aggregation**
  - describe and *compress* the data into a summary
- ***Re*-Sampling**
  - facilitate *generalizing* to unseen data
- **Uncertainty**
  - quantifying *trust* in our estimates based on different sources of error
- **Learning**
  - using data to *update* our estimates

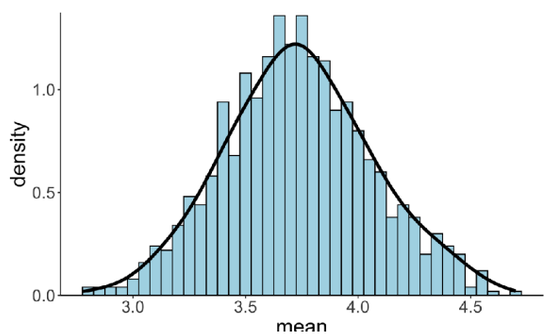
3 distributions in statistical inference



population

- unknown
- our target for inference
- e.g. we might be interested in the *mean* of the population distribution

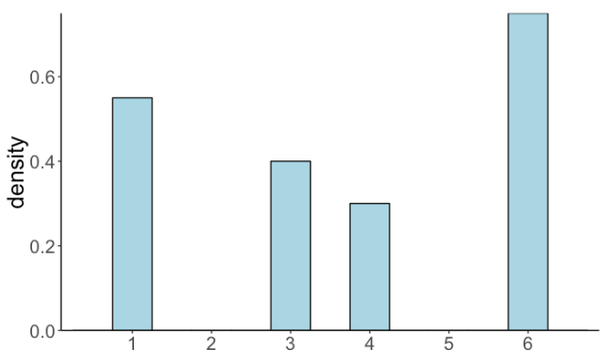
---



re-sampled distribution

- bridge between sample and population
- derived mathematically / ***computationally***
- asymptotic distribution theory or ***resampling*** approaches

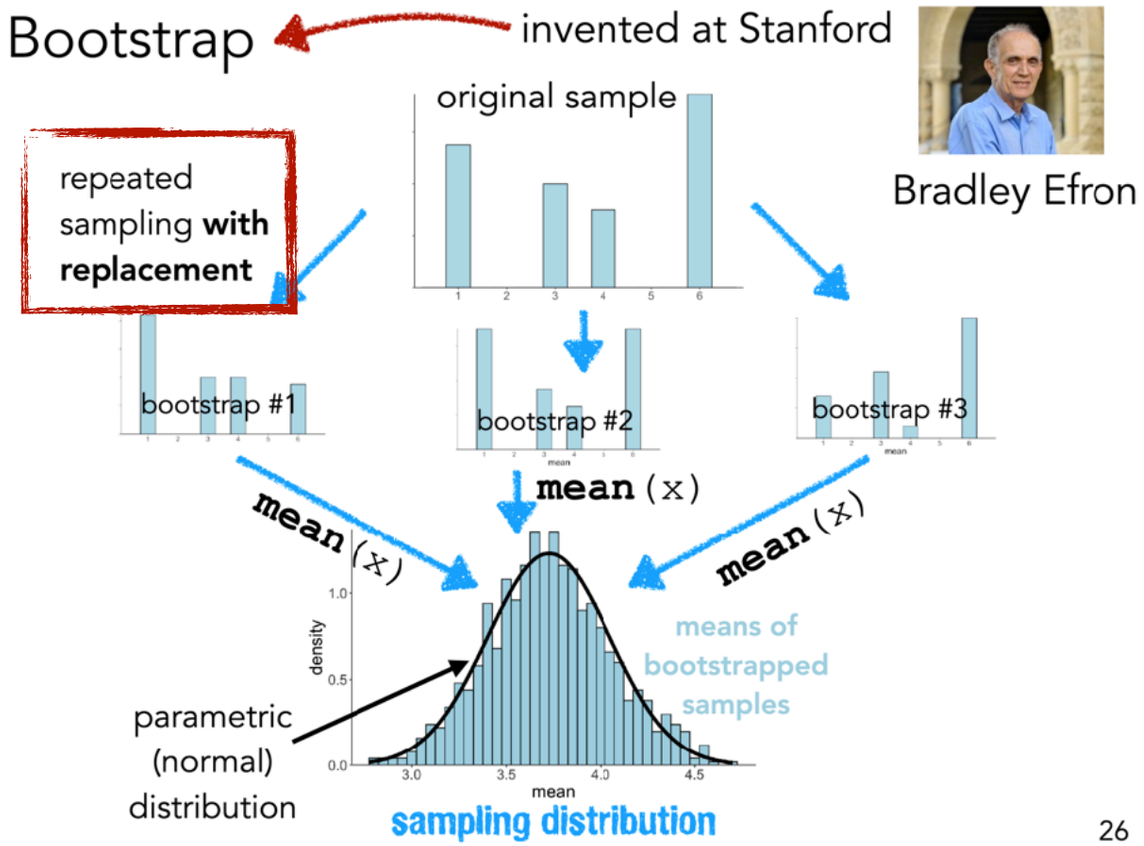
  shows how test statistic varies between samples



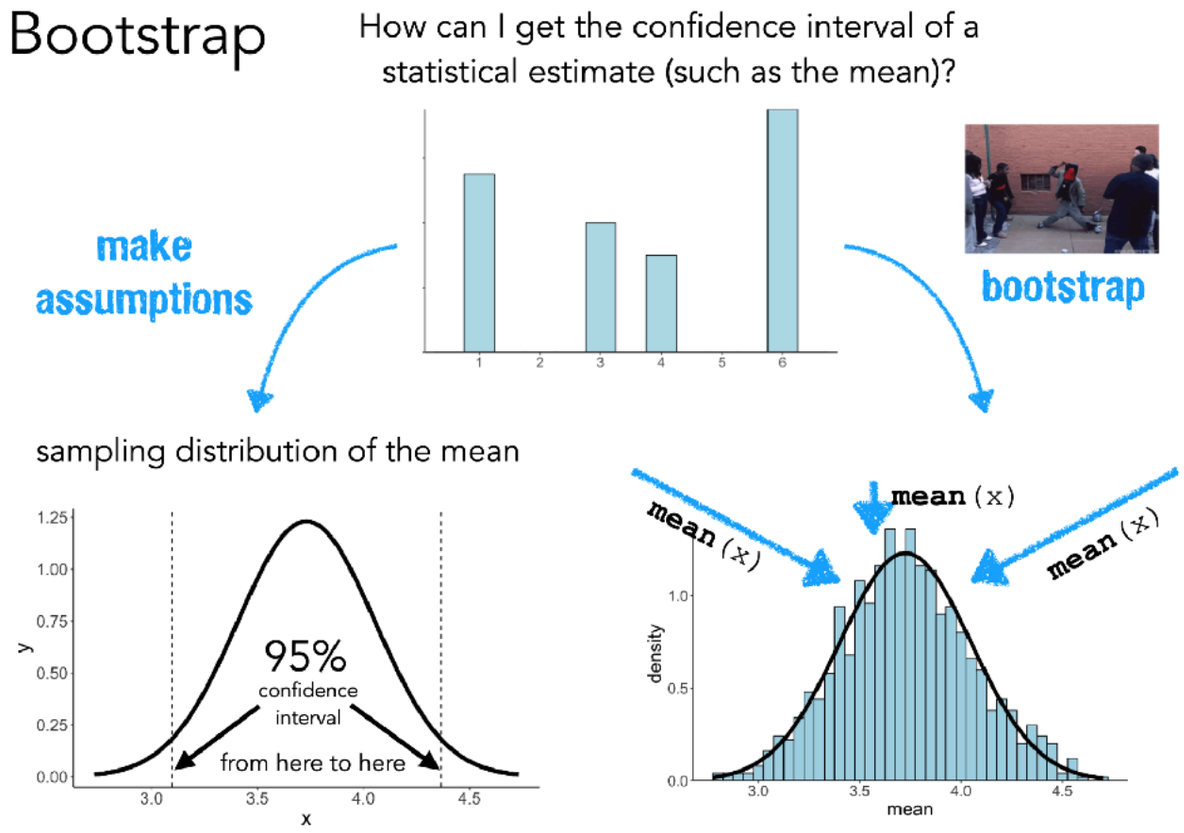
sample distribution

- our observed sample
- we compute statistics of interest (mean, variance, correlation, ...)
- make an inference about the population via the sampling distribution

Bootstrapping: resample *with* replacement



26



SciPy API > Statistical functions (`scipy.stats`) > bootstrap

scipy.stats.

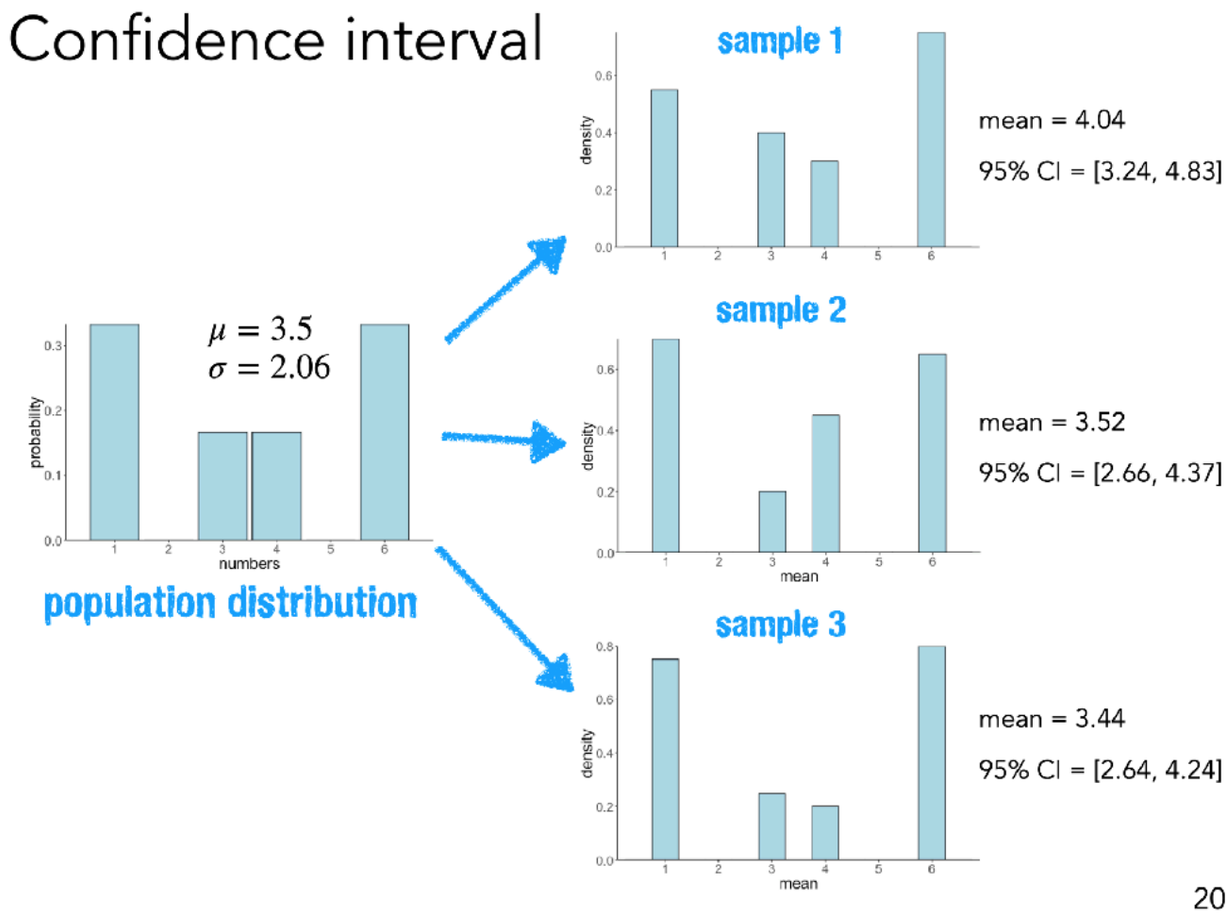
bootstrap

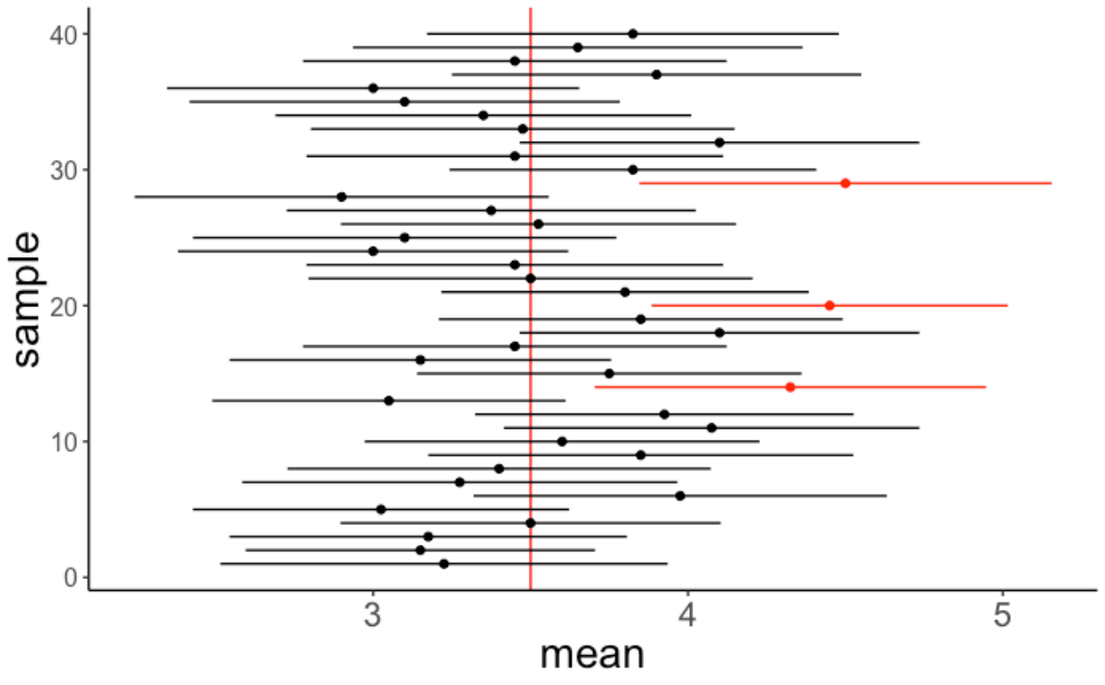
```
bootstrap(data, statistic, *, n_resamples=9999, batch=None,
vectorized=None, paired=False, axis=0, confidence_level=0.95,
alternative='two-sided', method='BCa', bootstrap_result=None, rng=None)
```

Compute a two-sided bootstrap confidence interval of a statistic. [source]

Reminder: Confidence intervals

"If we were to repeat the experiment over and over, then 95% of the time the confidence intervals contain the estimate of interest."





Permutation: resample ***without*** replacement

Permutation = Shuffling

observed data

participant	condition	performance
1	control	4.25
2	control	5.87
3	control	3.83
4	control	8.69
5	control	6.16
26	experimental	4.42
27	experimental	4.27
28	experimental	2.29
29	experimental	3.78
30	experimental	5.13

1

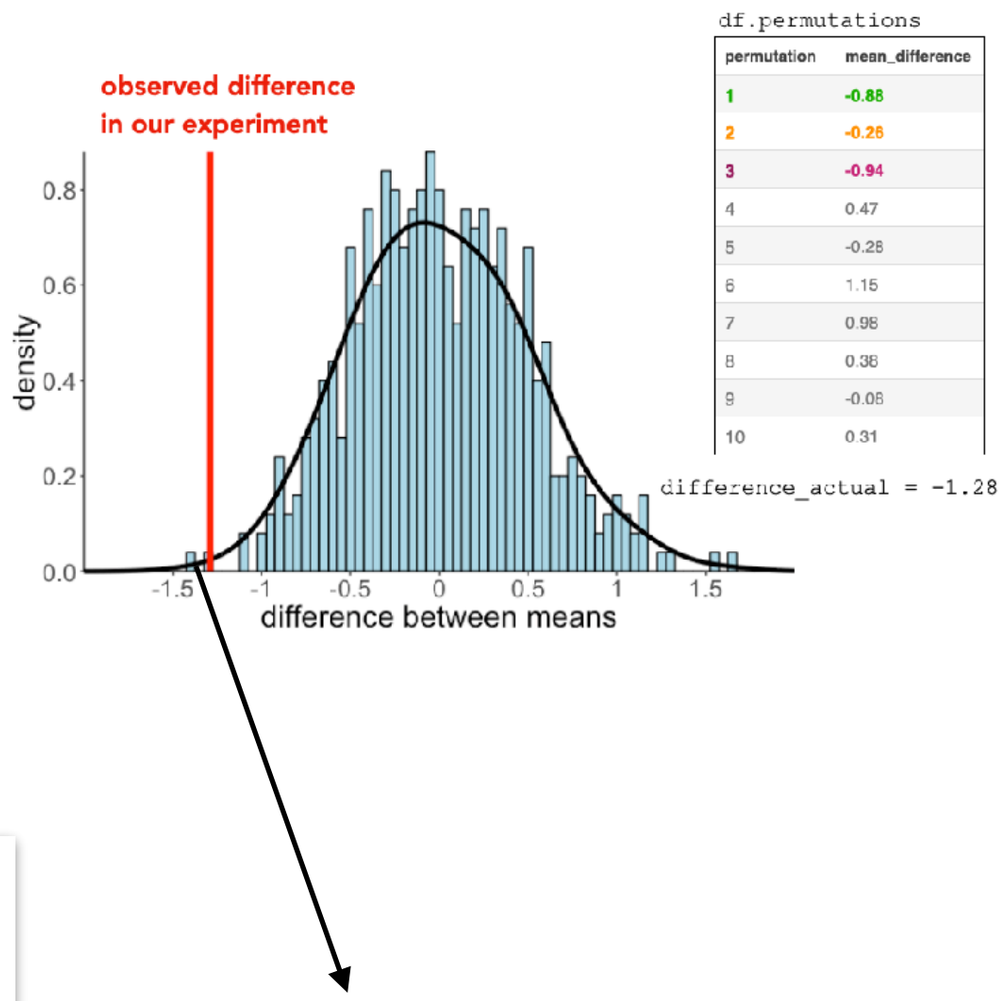
2

3

⋮

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

Statistic after shuffling =
resampled null distribution



df.permutations

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

difference_actual = -1.28

59

SciPy API > Statistical functions (`scipy.stats`) > permutation_test

scipy.stats.

permutation_test

permutation_test(*data*, *statistic*, *, *permutation_type*='independent', *vectorized*=None, *n_resamples*=9999, *batch*=None, *alternative*='two-sided', *axis*=0, *rng*=None) [source]

Performs a permutation test of a given statistic on provided data.

For independent sample statistics, the null hypothesis is that the data are randomly sampled from the same distribution. For paired sample statistics, two null hypothesis can be tested: that the data are paired at random or that the data are assigned to samples at random.

What is a p-value?

The **p-value** is the probability of finding the observed (or more extreme) results when the null hypothesis (H_0) is true.

$$p(\text{test statistic} \geq \text{observed value} \,|\, H_0 = \text{true})$$

Fundamental concepts of statistics

- **Aggregation**
  - describe and *compress* the data into a summary
- ***Re*-Sampling**
  - facilitate *generalizing* to unseen data
- **Uncertainty**
  - quantifying *trust* in our estimates based on different sources of error
- **Learning**
  - using data to *update* our estimates

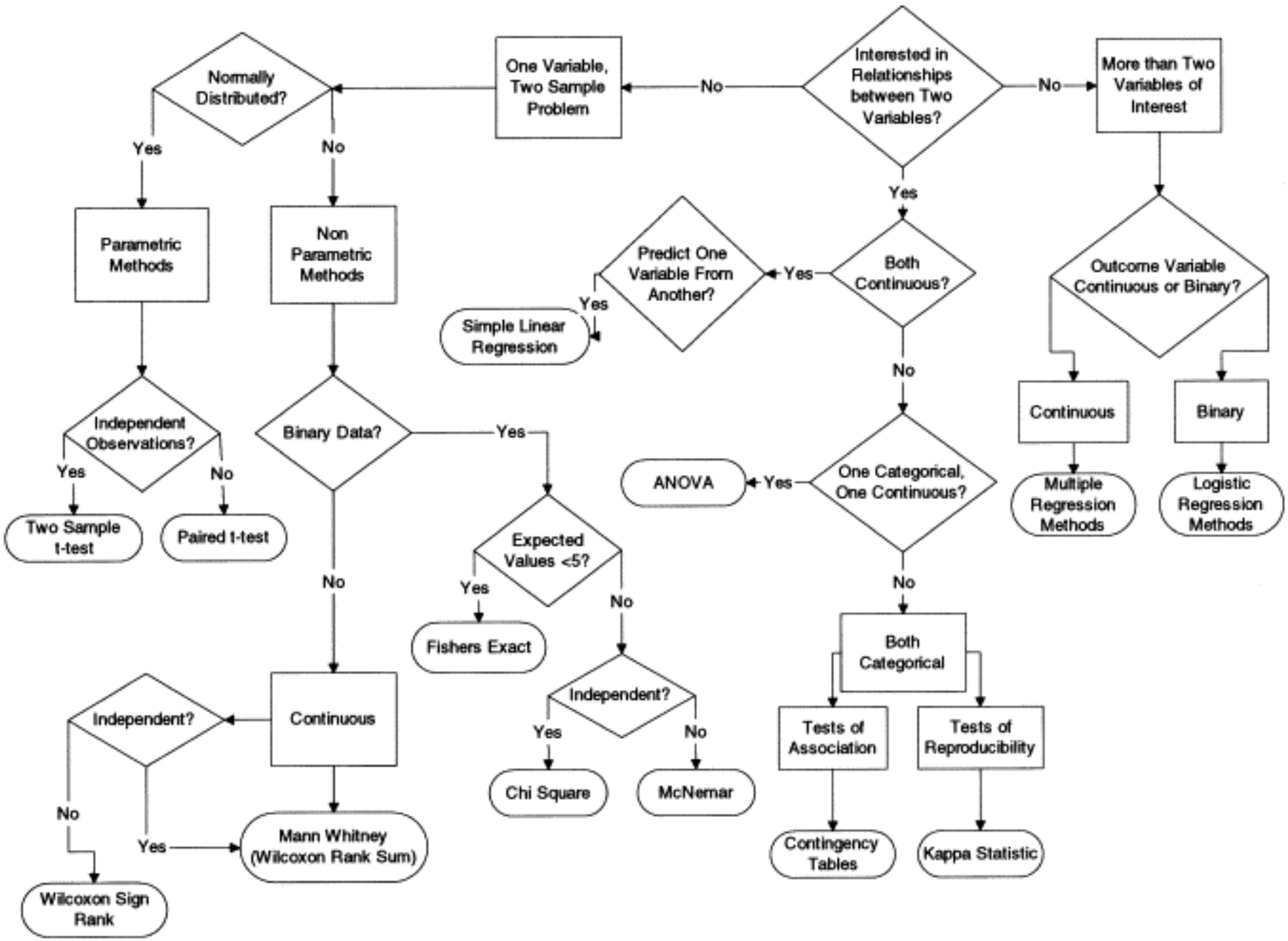
Fundamental concepts of statistics

- **Aggregation**
  - describe and *compress* the data into a summary
- ***Re*-Sampling**
  - facilitate *generalizing* to unseen data
- **Uncertainty**
  - quantifying *trust* in our estimates based on different sources of error
- **Learning**
  - using data to *update* our estimates

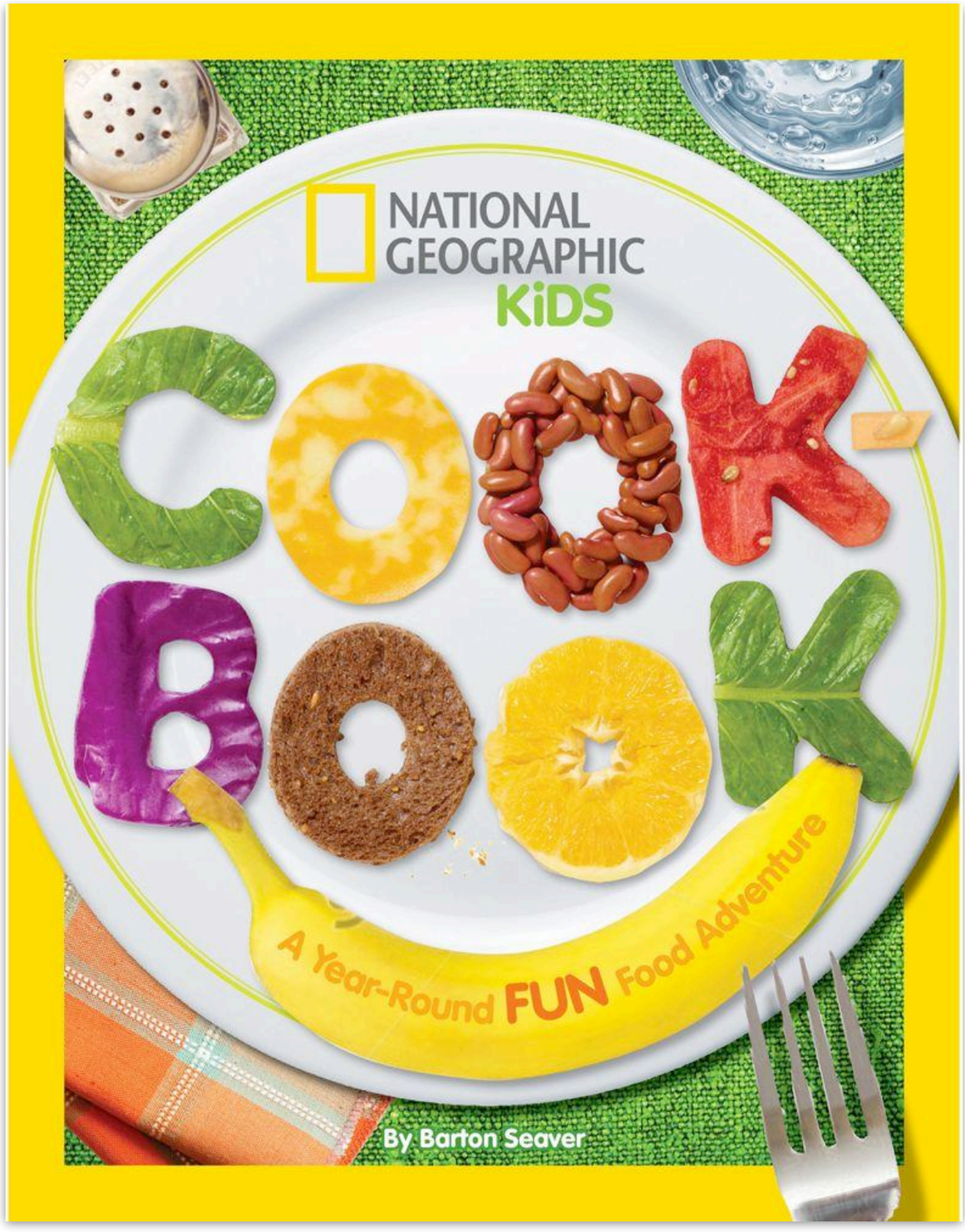
Modeling data

Cookbook vs. Model Comparison

The cookbook approach

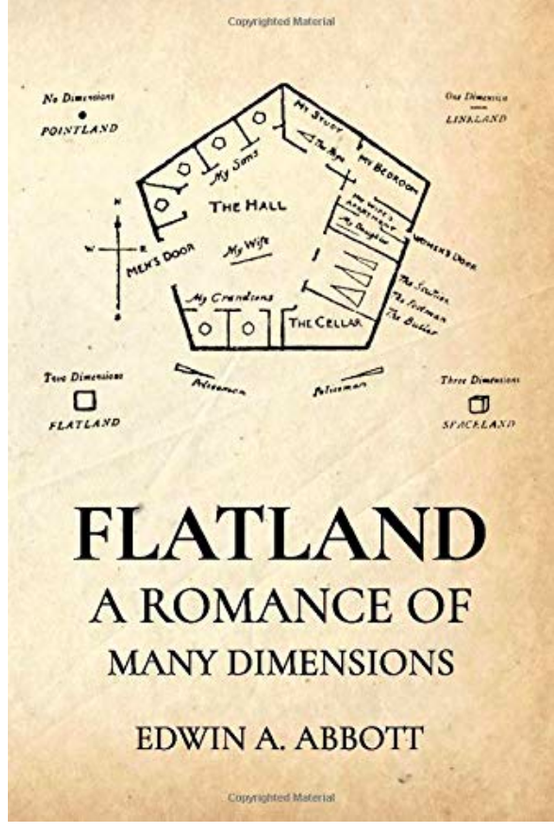


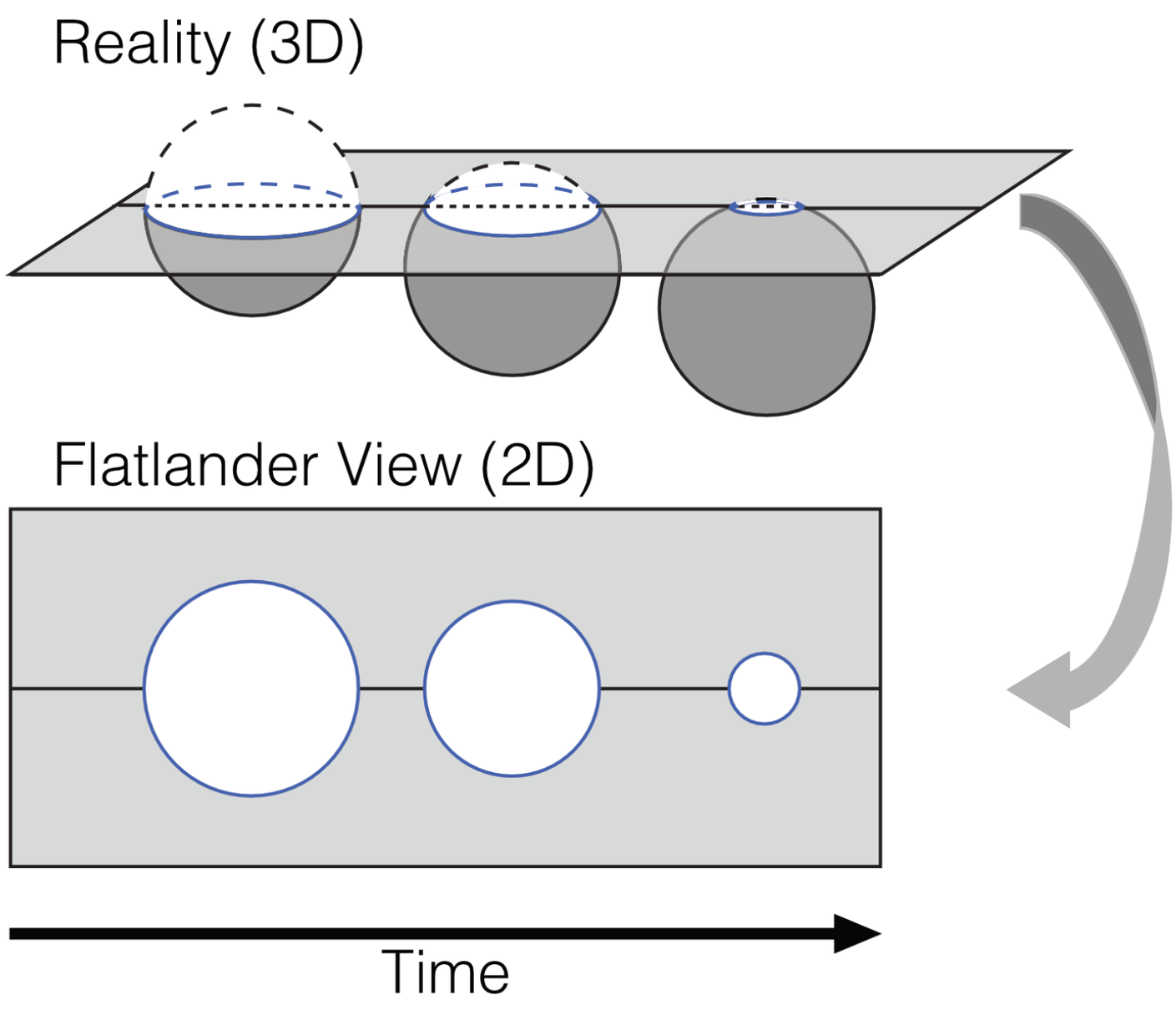
The cookbook approach



- many statistics textbooks are organized in this way
- works reasonably well if what we want to cook is in the book
- leaves us with **no idea what to do if we can't find a recipe**

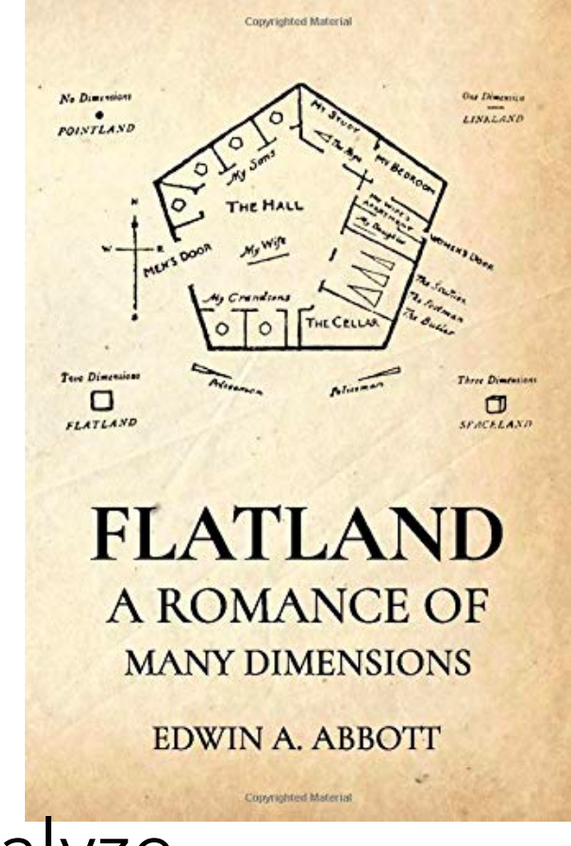
Remember? The Flatland Fallacy





Jolly & Chang, 2019

Remember? The Flatland Fallacy



If we're **limited by recipes** we've memorized

↓

We're **limited** to designing **experiments** we can analyze

↓

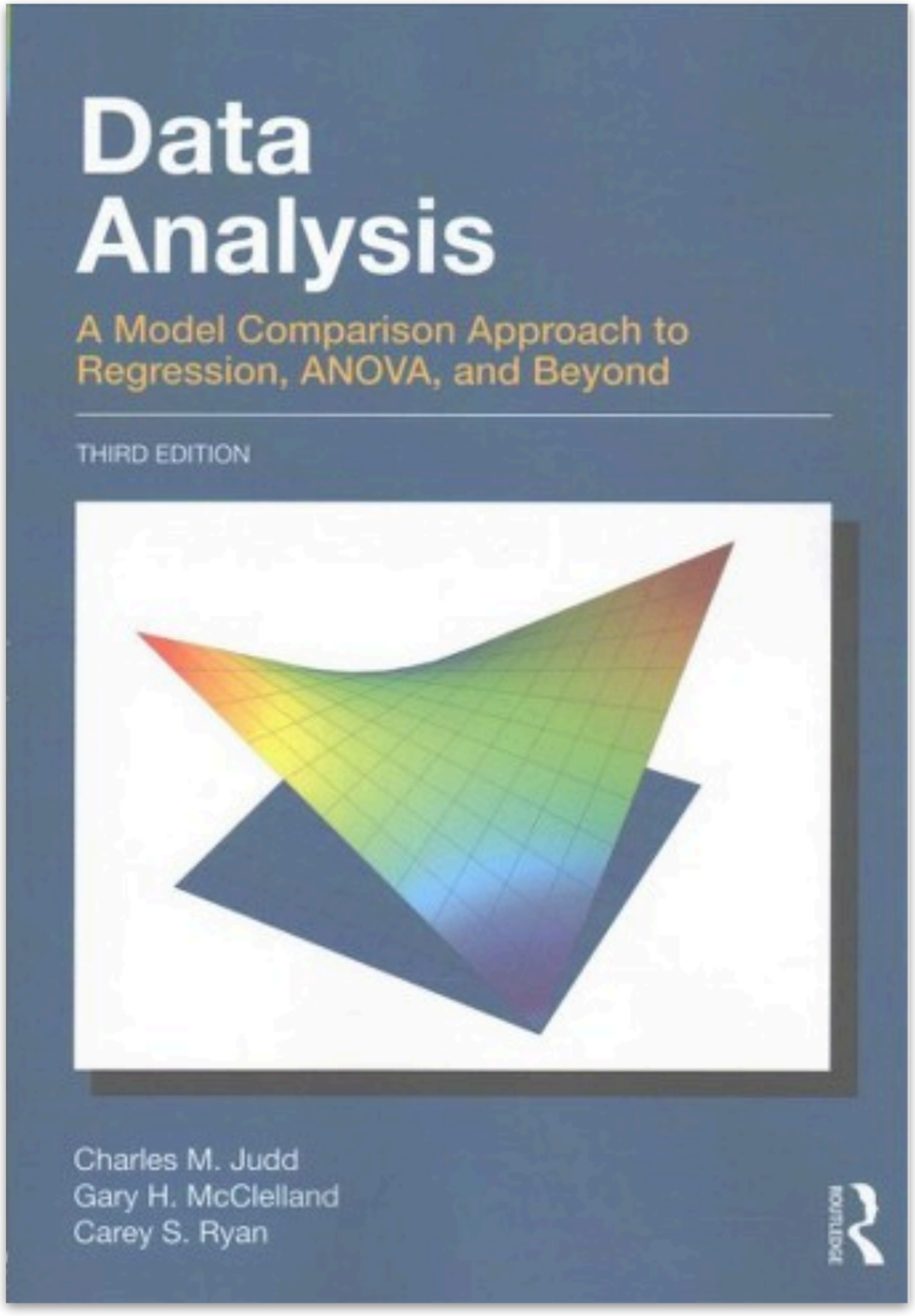
We're **limiting** the **answers** to our scientific questions

↓

What are we even studying?

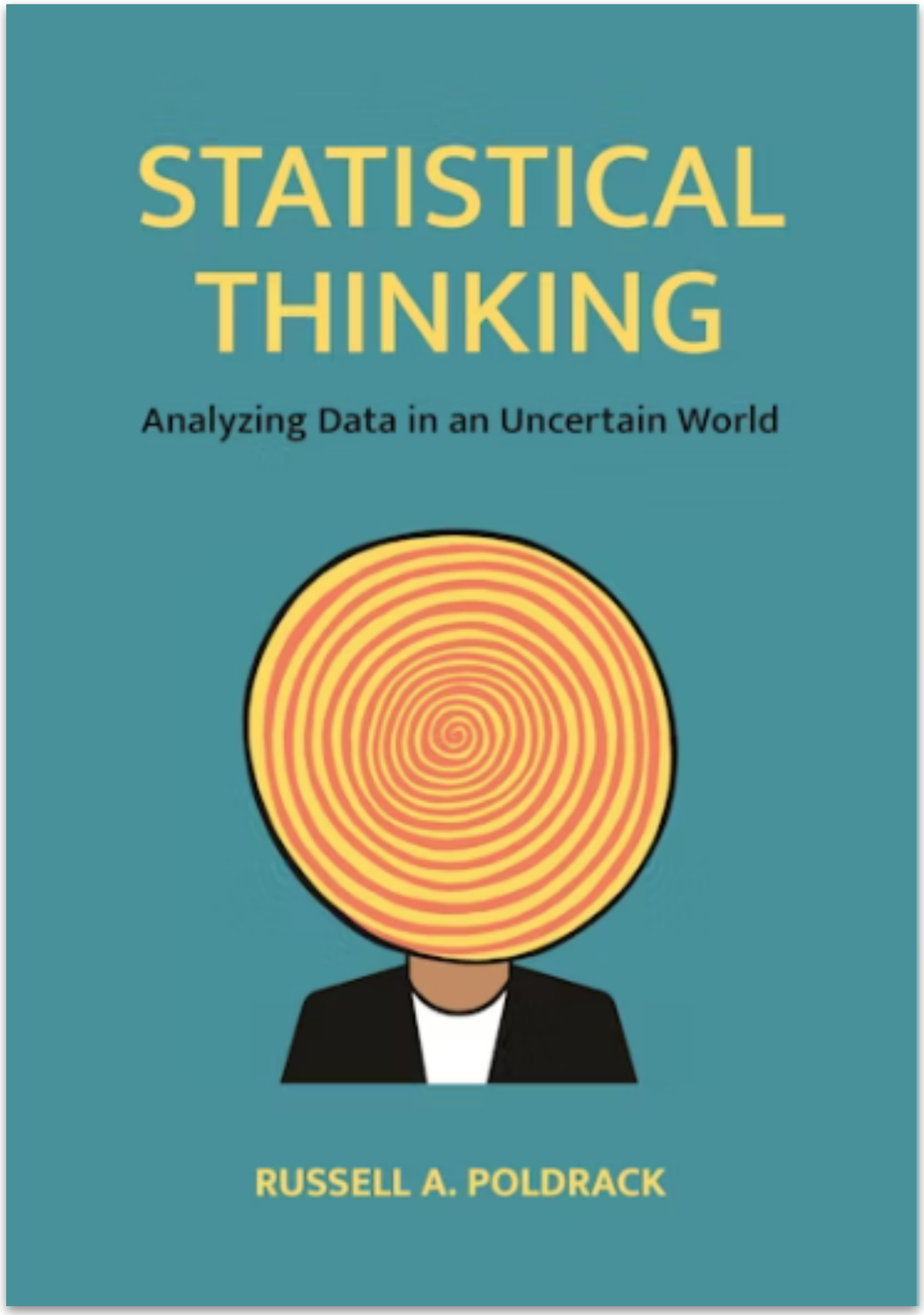
Jolly & Chang, 2019

Model comparison approach



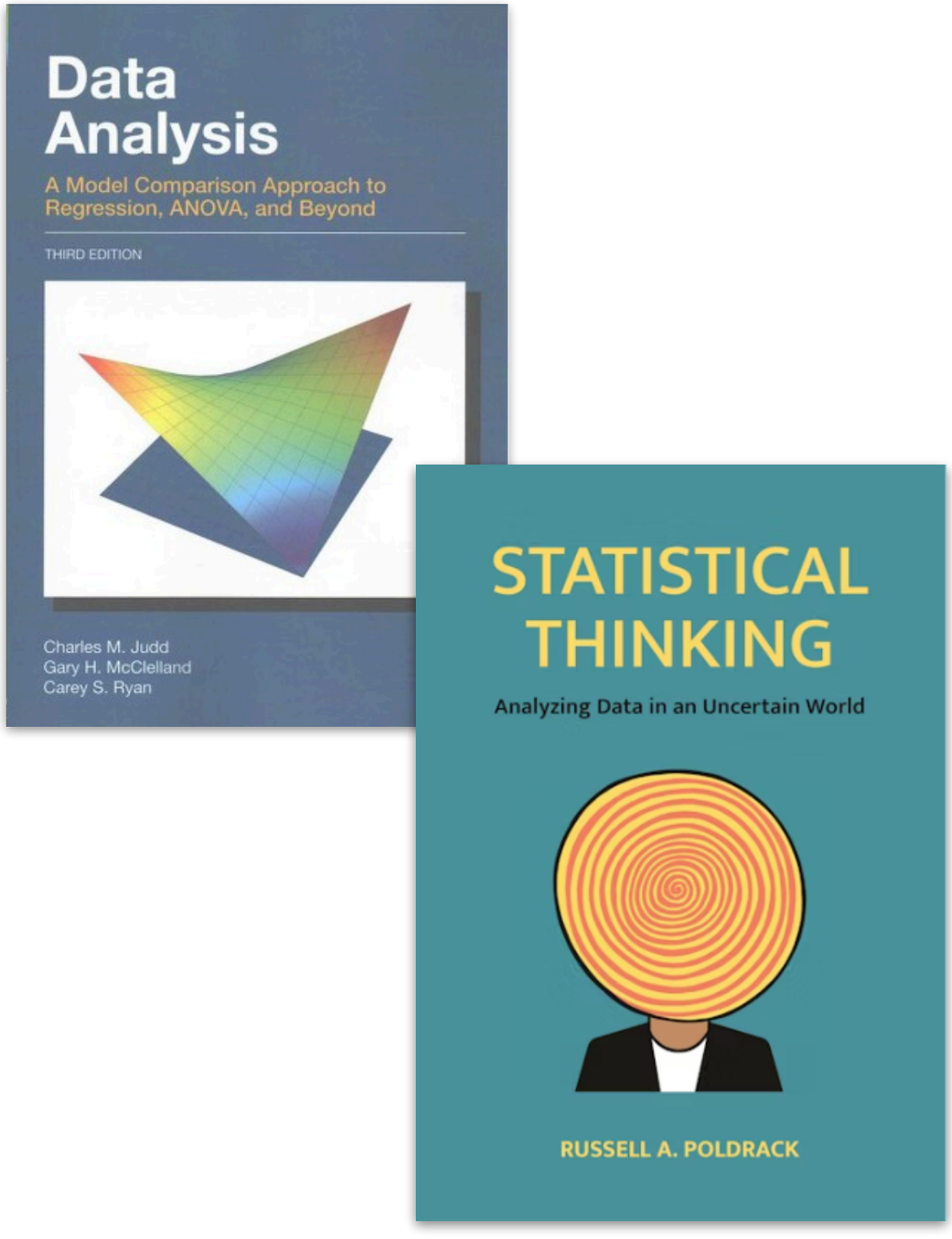
Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach*. Routledge.

Model comparison approach



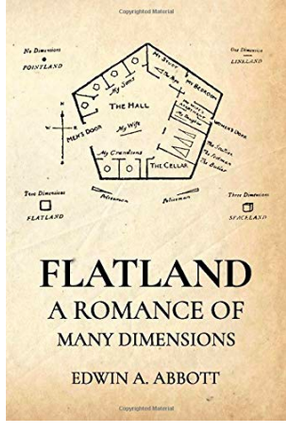
Model comparison approach: why?

- more flexible approach
- hopefully generates better insight
- thinking of statistical analysis as modeling
- allows for a smoother transition into more advanced topics (e.g. Bayesian, deep-learning)



What is a model?

What is a model?



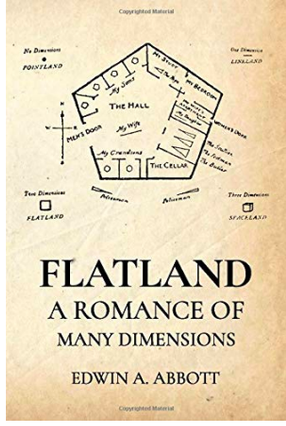
General: A general mathematical **function** that transforms *inputs* into *outputs*

```
output = f(input)

happiness = f(chocolate)

y = f(x)

y = f(x1, x2,…)
```

We care about this

Jolly & Chang, 2019

What is a model?



General: A general mathematical **function** that transforms *inputs* into *outputs*

```
output = f(input)

happiness = f(chocolate)

y = f(x)

y = f(x1, x2,…)
```

It’s our theory in *math*

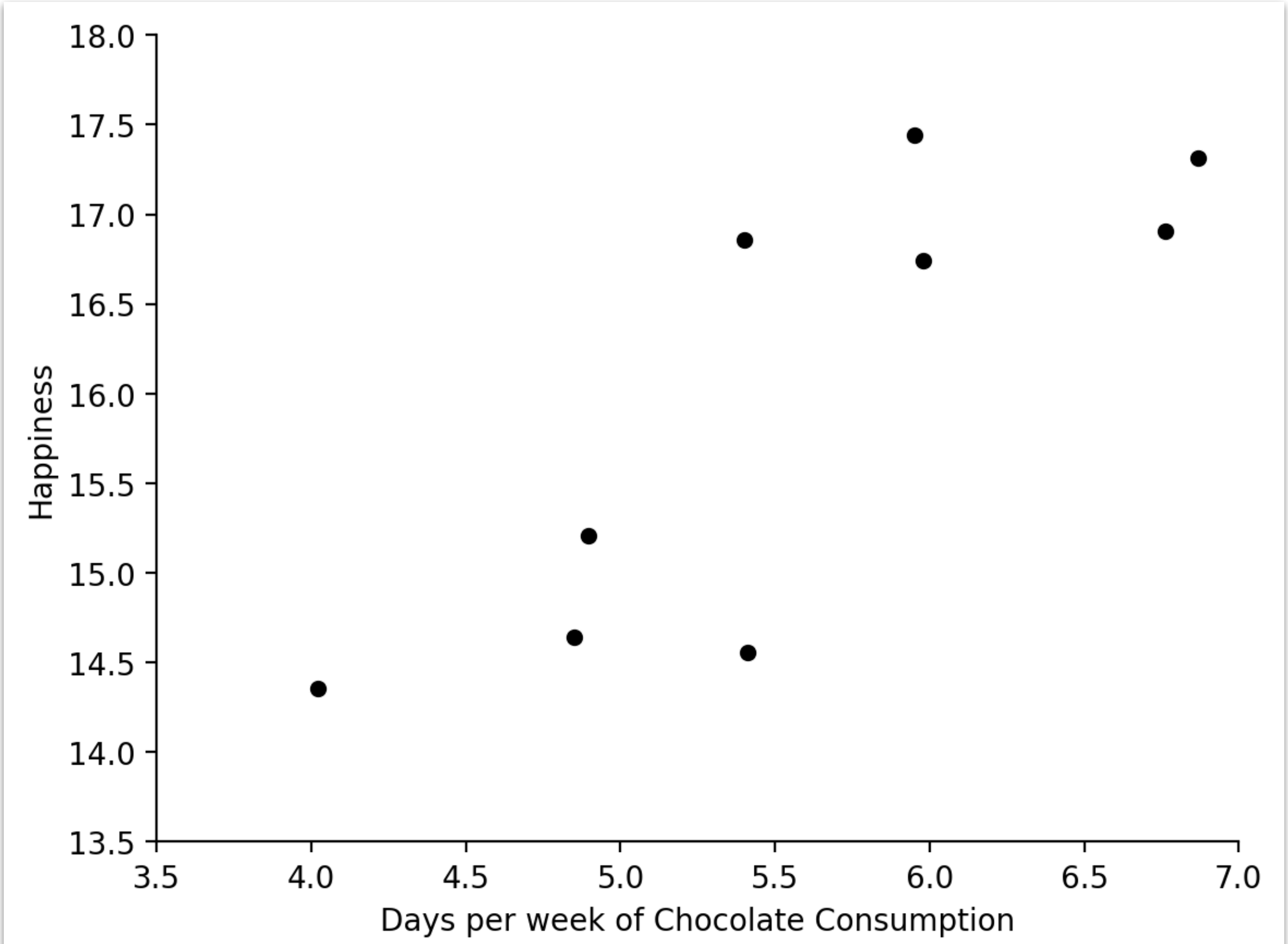
Jolly & Chang, 2019

What is a **statistical** model?

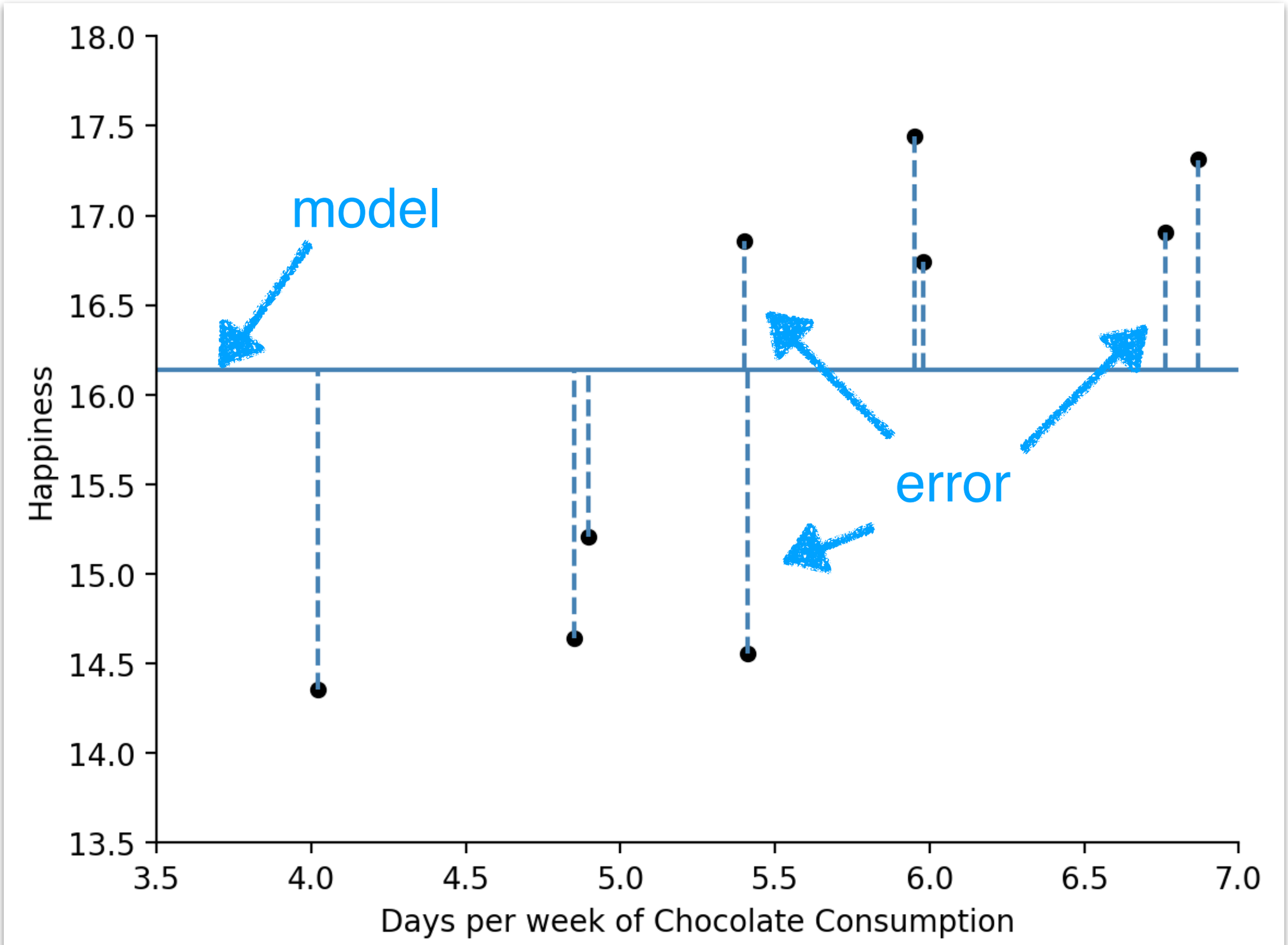
A **theory** of how **observed data** were **generated**

$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

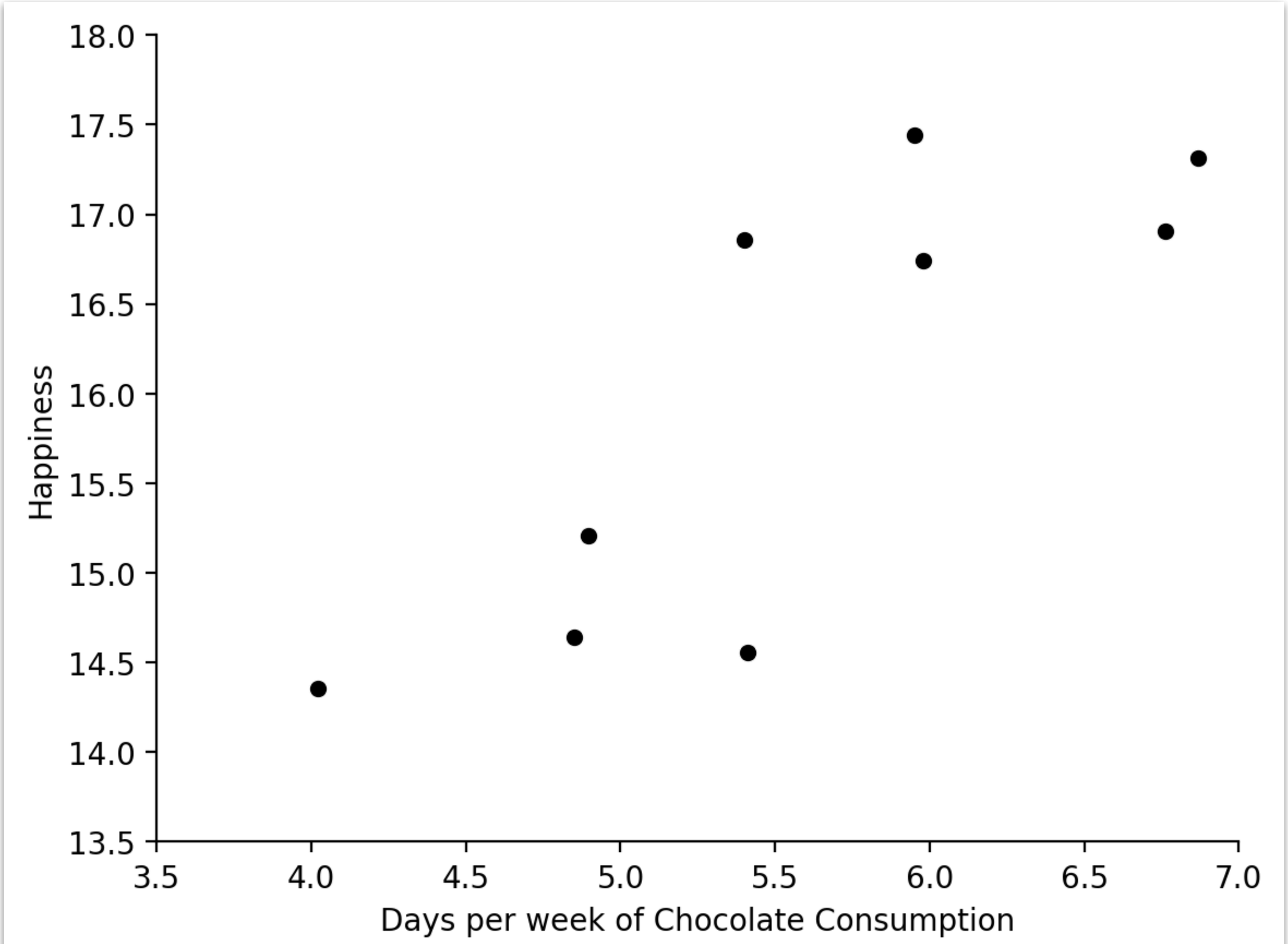
Data = Model + Error



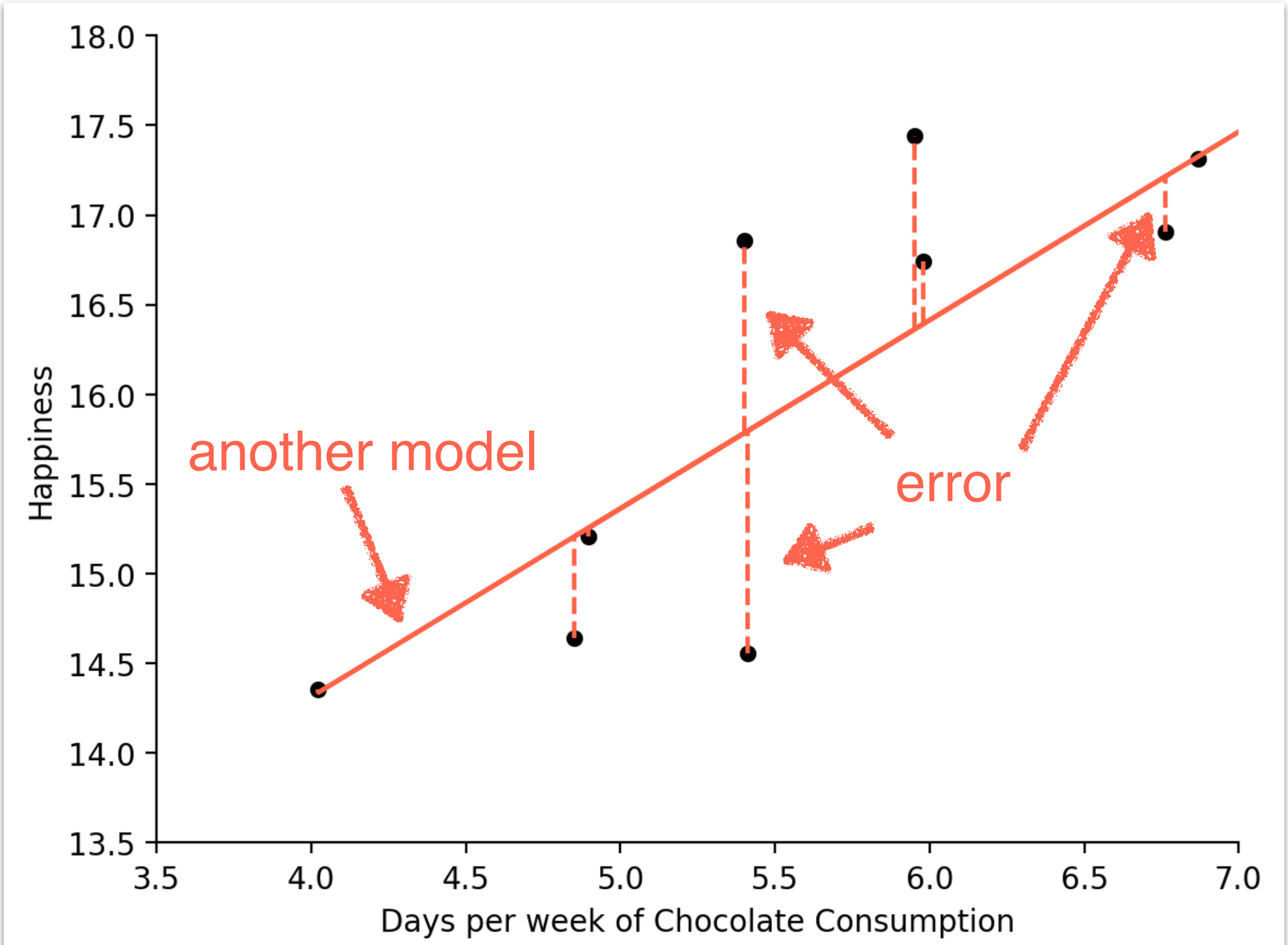
Data = Model + Error



Data = Model + Error



Data = Model + Error



What is a model?

A **theory** of how **observed data** were **generated**

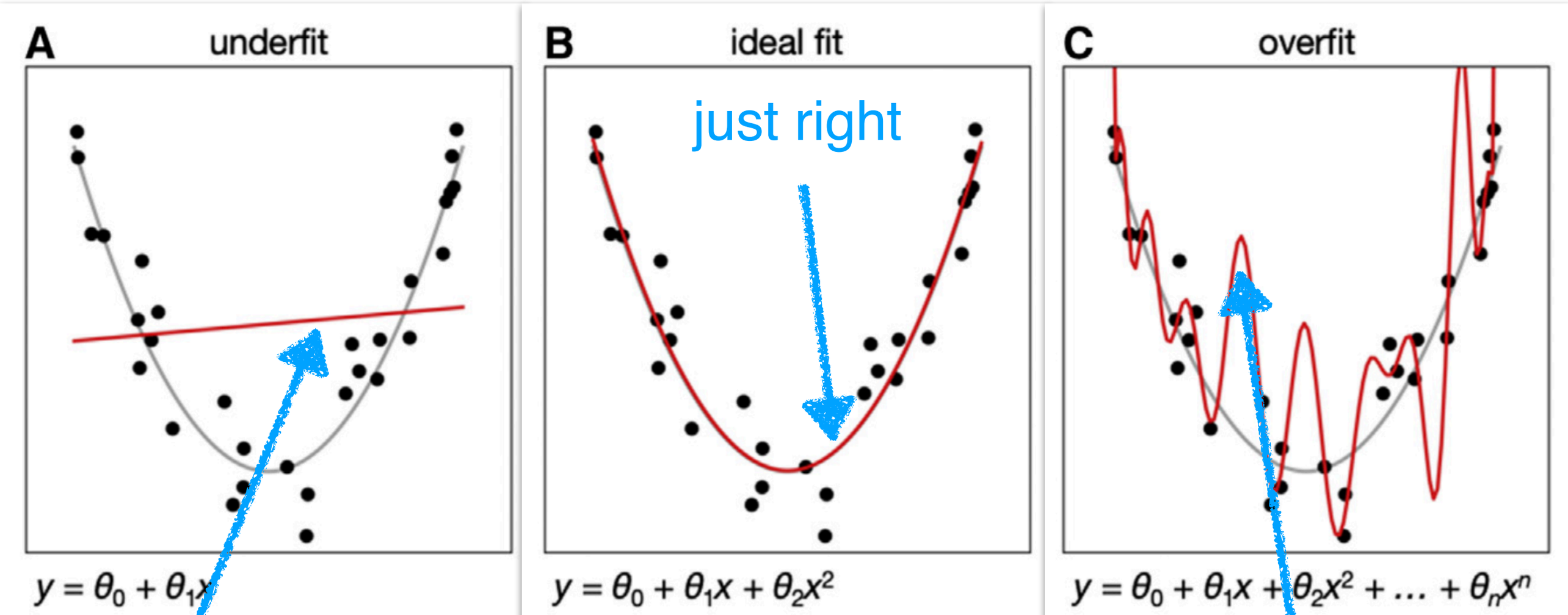
$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

what's a good model?

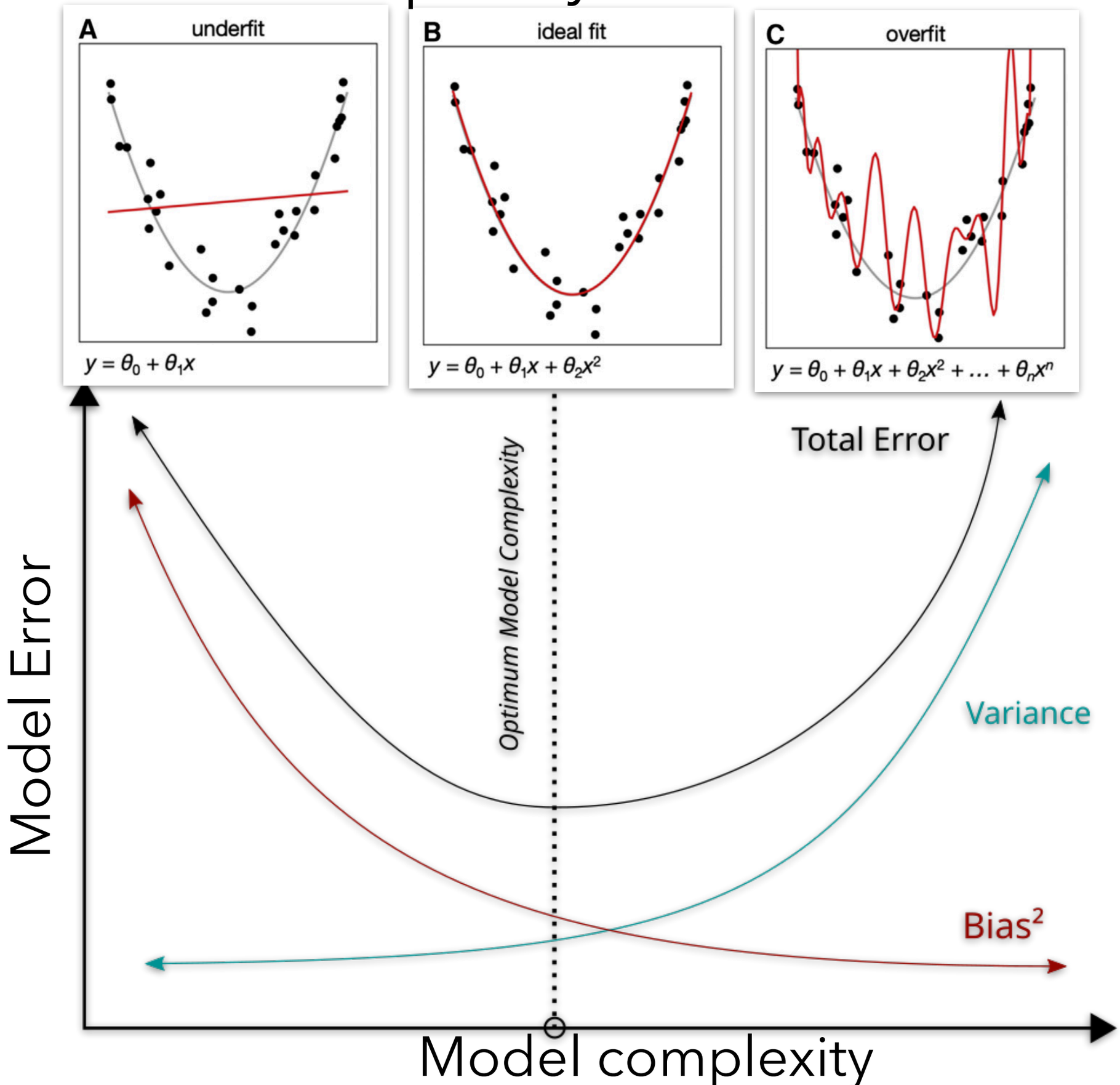
a good model balances

fit and **complexity**

Remember? Fit vs Complexity -> Bias-Variance trade-off



Review: Fit vs Complexity -> Bias-Variance trade-off



What is a model?

A **theory** of how **observed data** were **generated**

$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

how shall we define this?

Residual: the part that's left over after we have used the model to predict/explain the data

$$\text{Error} = \text{Data} - \text{Model}$$

Residual: the part that's left over after we have used the model to predict/explain the data

$$\text{Error} = \text{Data} - \hat{Data}$$

Model output

"hats" on variables are *estimated* not measured quantities

Residual: the part that's left over after we have used the model to predict/explain the data

$$\text{Error} = \text{Data} - \text{Model}$$

To reduce error we can:

improve data quality

e.g. run good experiments

improve the model

e.g. make predictions using additional information

Error = Data - Model

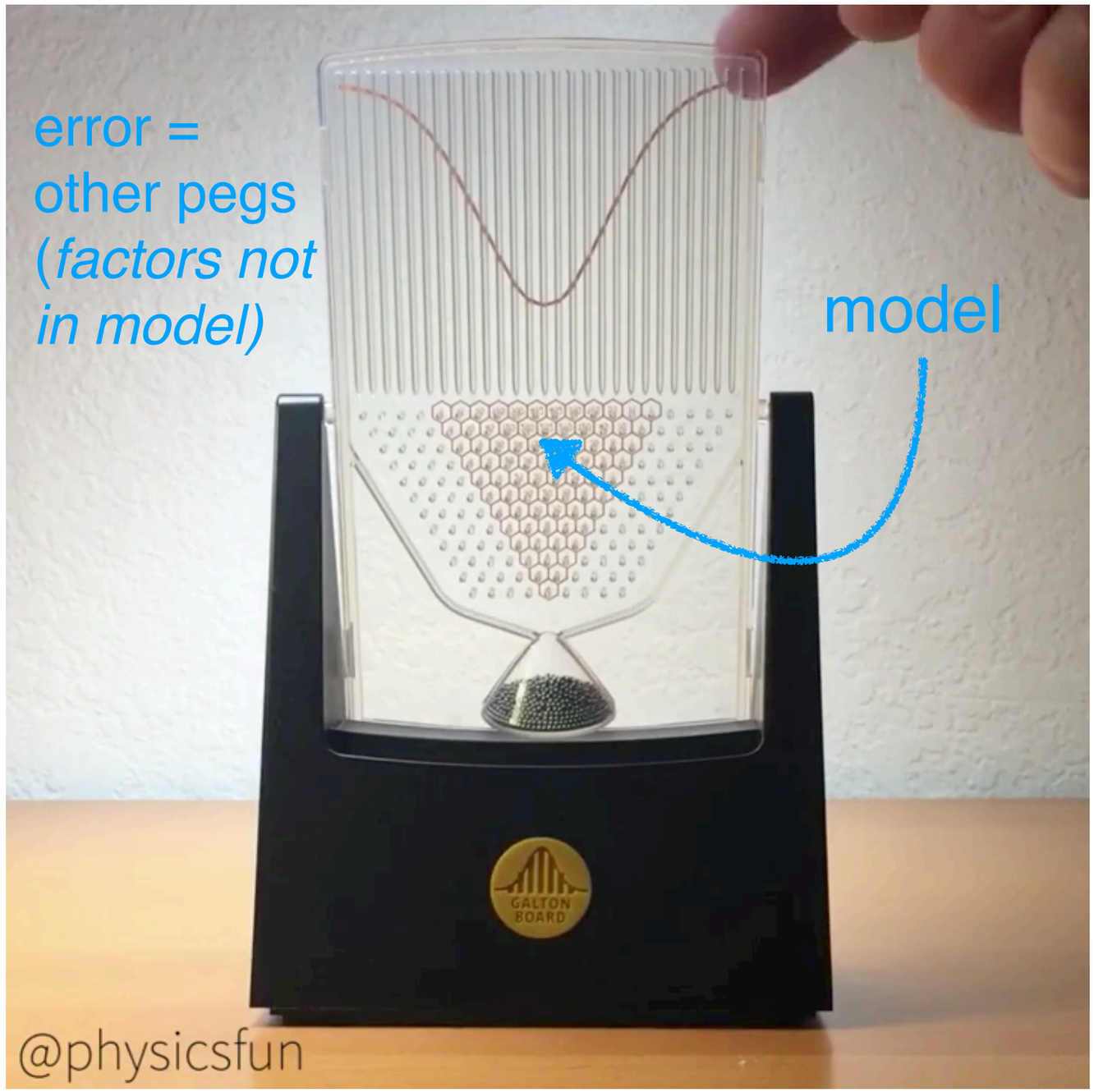
1. We **assume** that the **errors** are due to (*a potentially large number of*) factors that we didn't take into account.

2. We **assume** that each of these factors influences the data in **an additive way** (some pulling in one, others pulling in another direction).

Error = Data - Model

1. We **assume** that the **errors** are due to (*a potentially large number of*) factors that we didn't take into account.

2. We assume that each of these factors influences the data in **an additive way** (some pulling in one, others pulling in another direction).



Result: **Normally Distributed Errors**

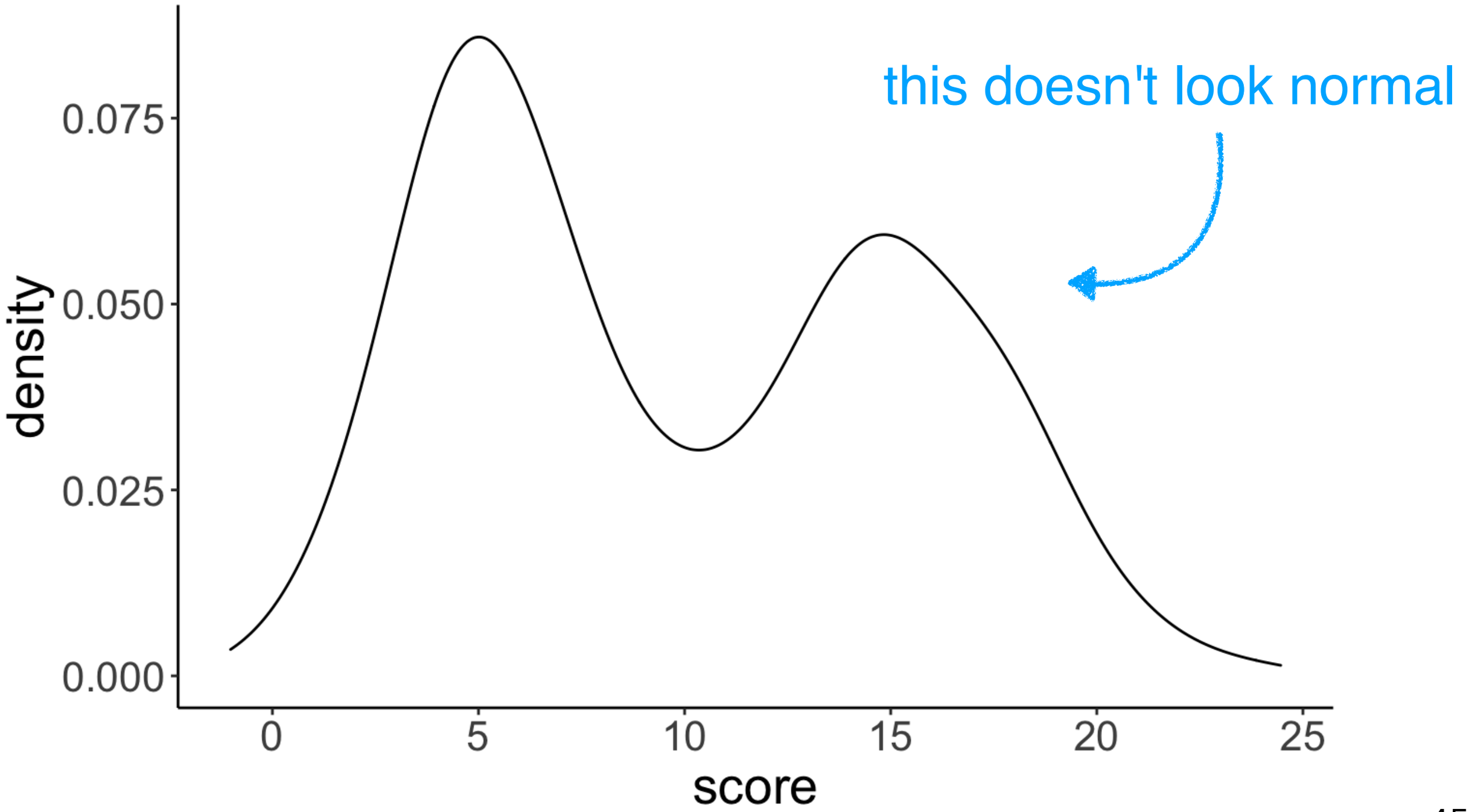
$$\text{Error} = \text{Data} - \text{Model}$$

assumed to be normally distributed

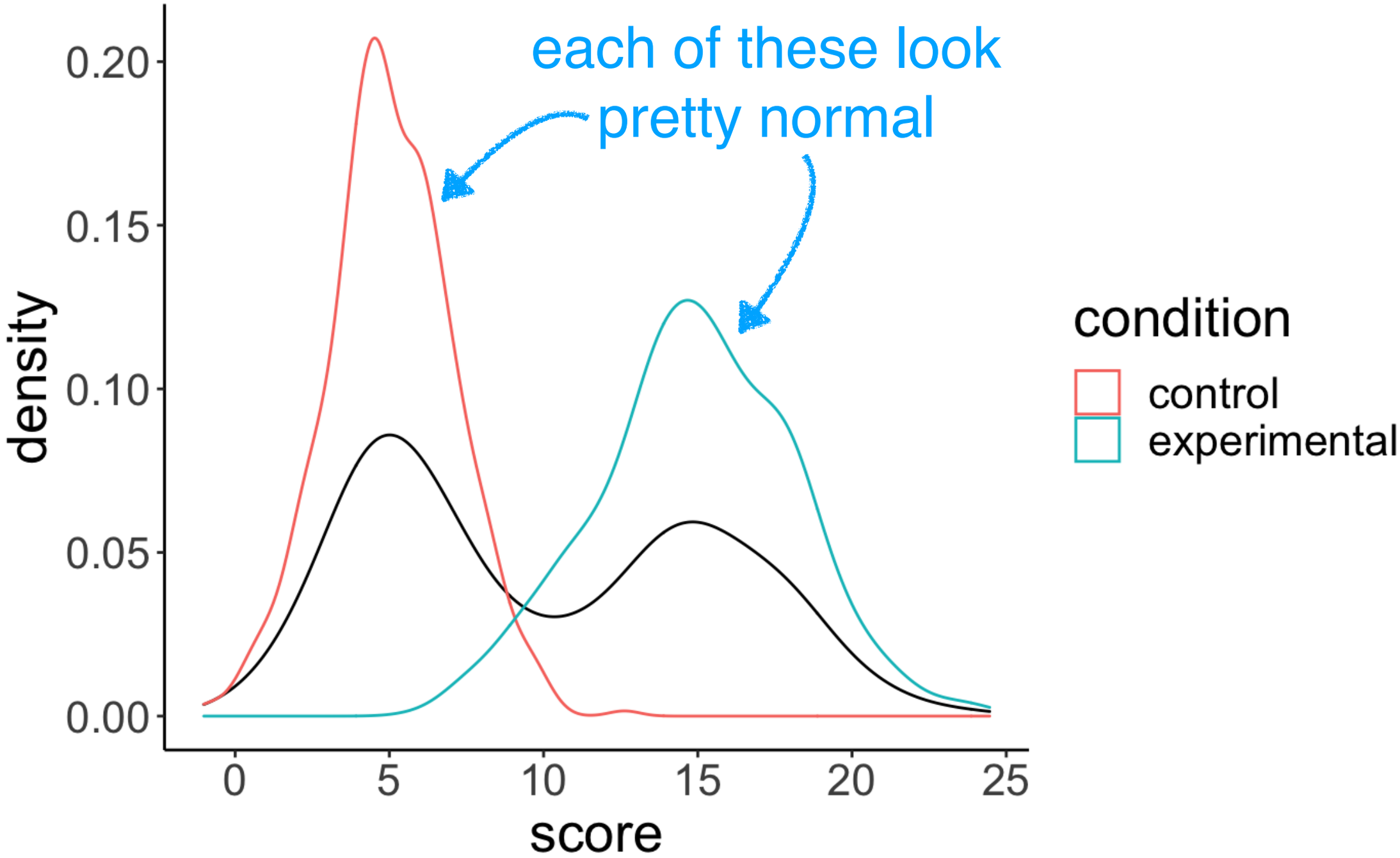
don't need to be normally distributed!!

very common misconception!!!

Distribution of **data**

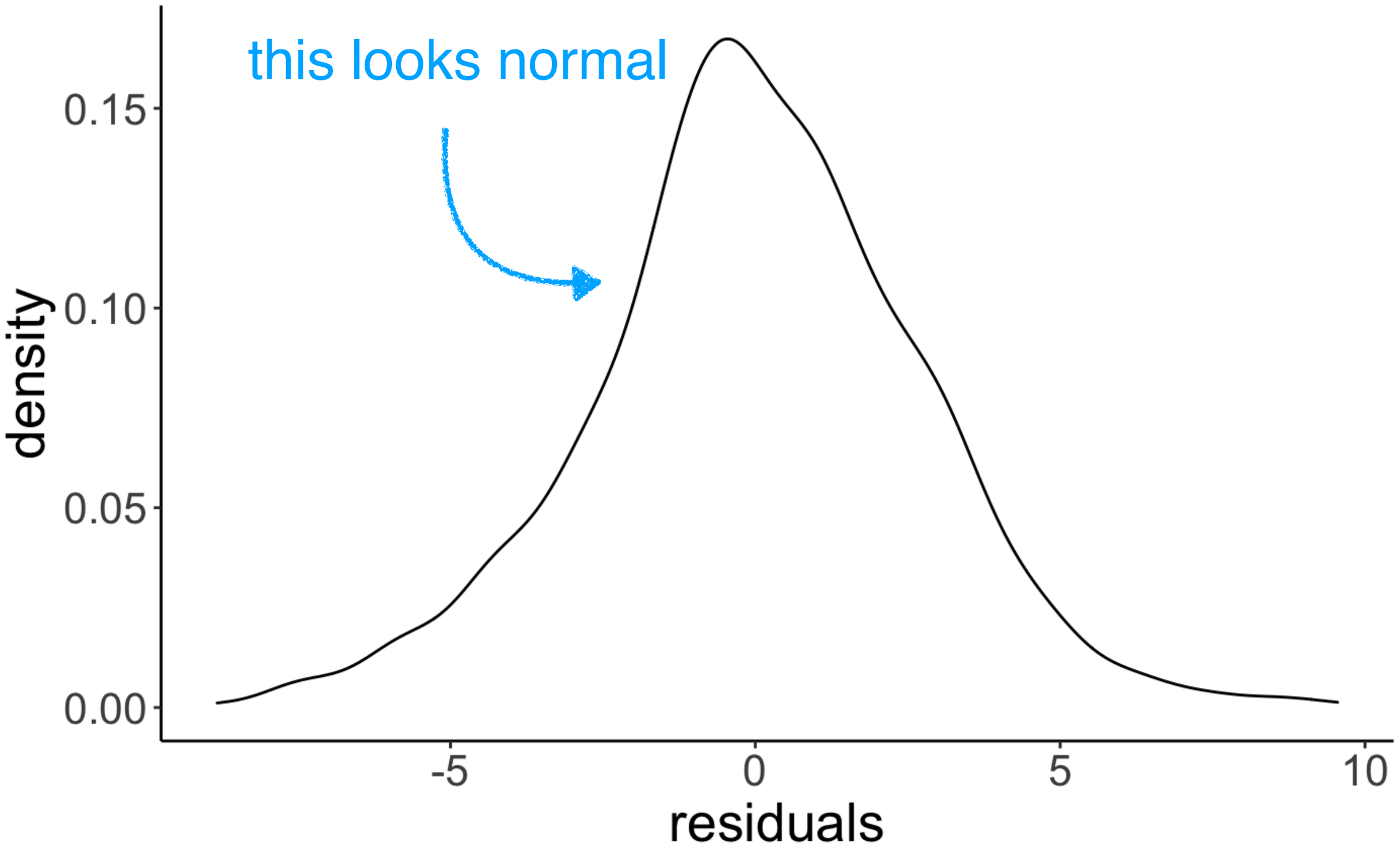


Distribution of **data**



Distribution of **residuals**

`Error` `= Data - Model`

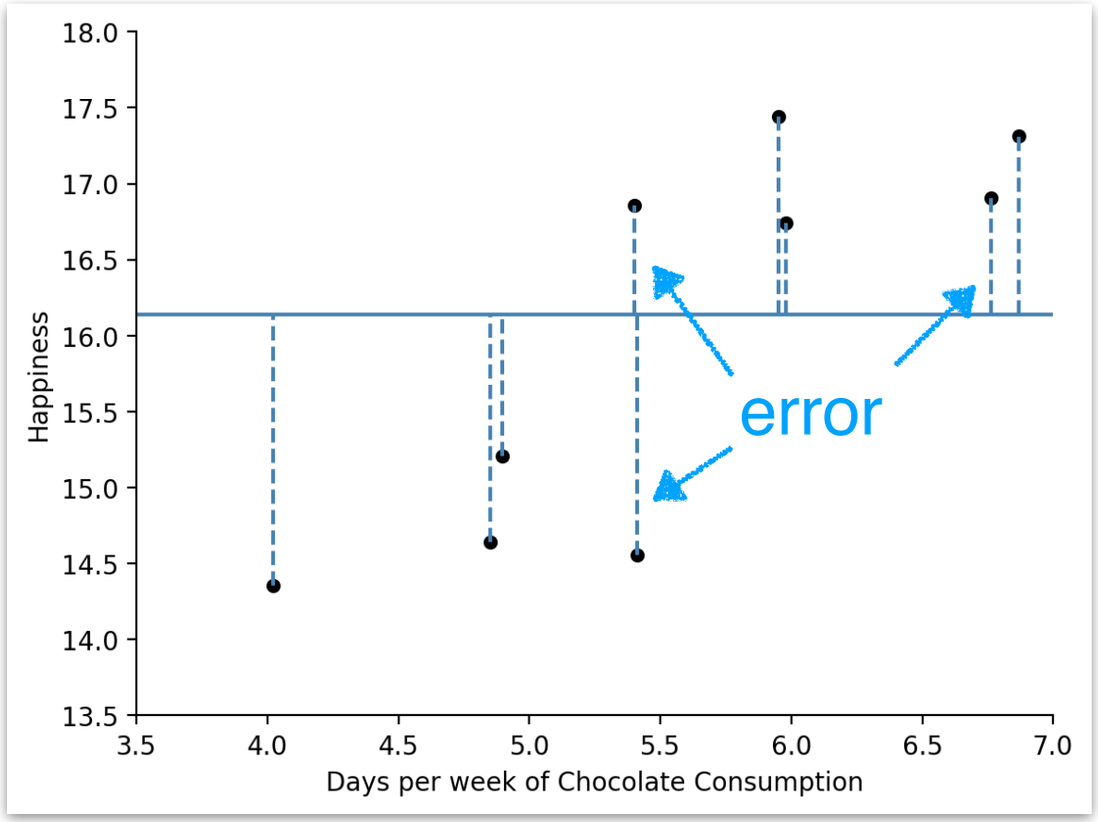


Error = Data - Model

concretely: fit models to minimize the

sum-of-squared-errors

square these and add them up



why squared error?

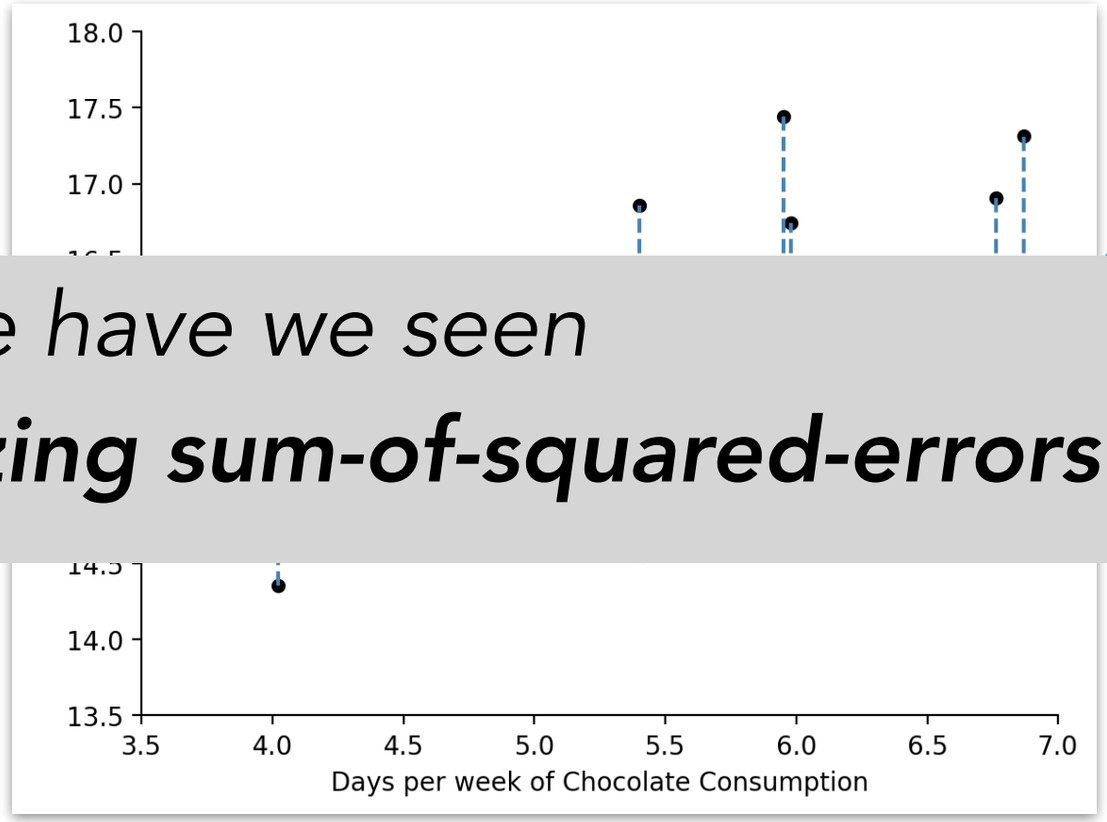
- positive and negative prediction errors don't cancel out
- larger errors are weighted more

Error = Data - Model

concretely: fit models to minimize the

sum-of-squared-errors

square these and add them up



…where have we seen

minimizing sum-of-squared-errors *before….*

why squared error?

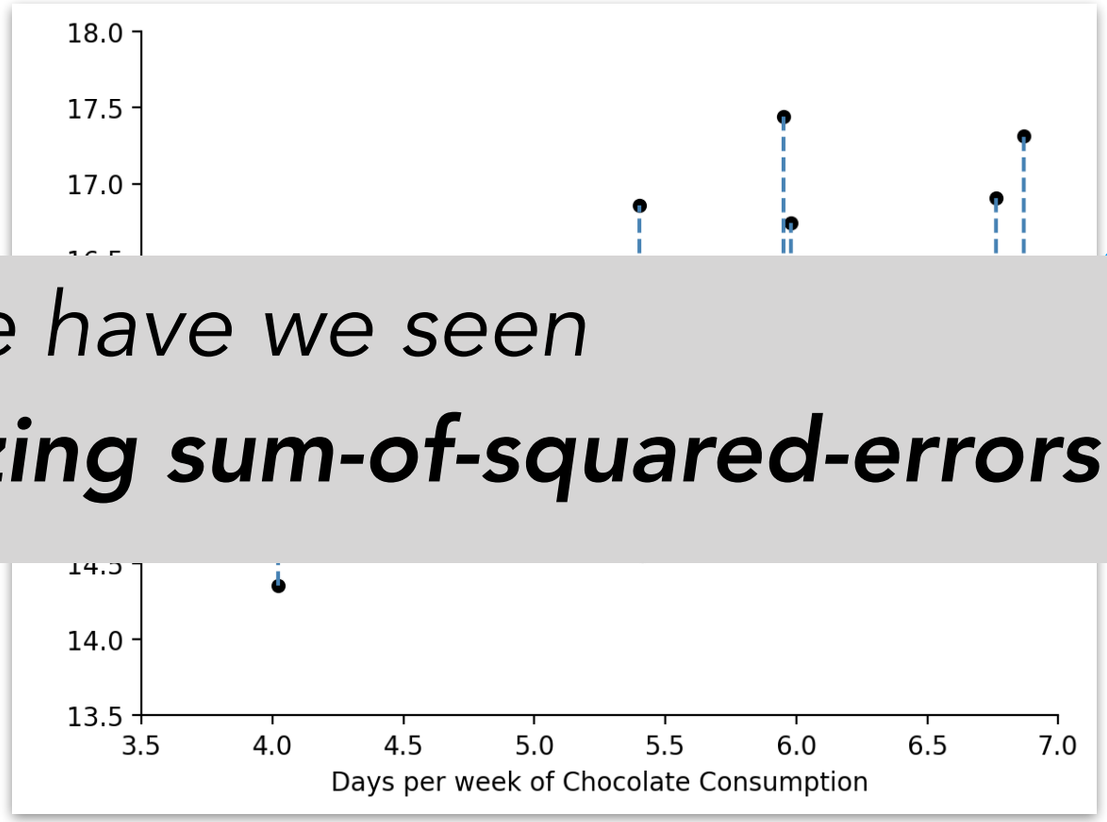
- positive and negative prediction errors don't cancel out
- larger errors are weighted more

Error = Data - Model

concretely: fit models to minimize the

sum-of-squared-errors

square these and add them up



…where have we seen

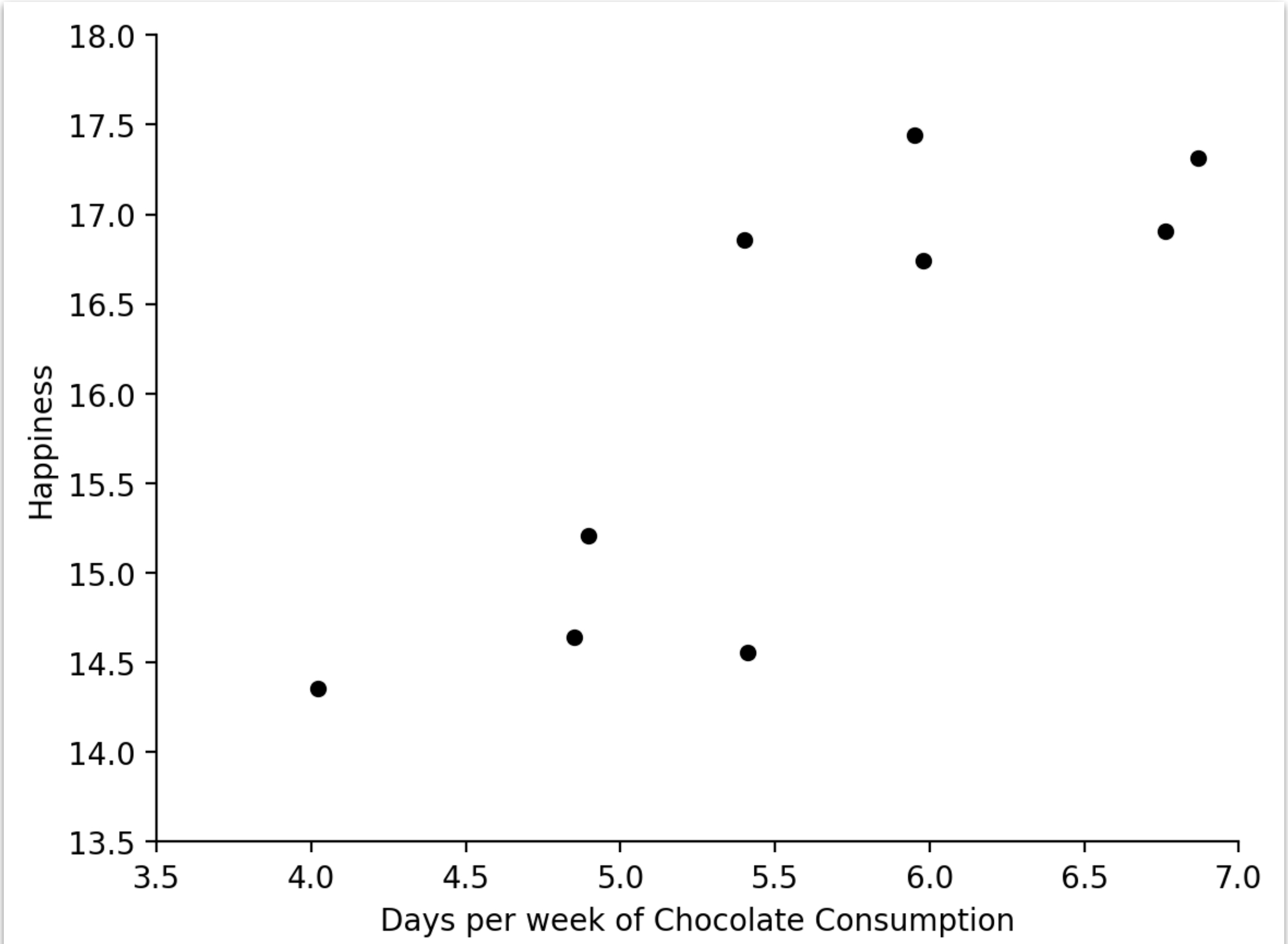
minimizing sum-of-squared-errors *before….*

why squared error?

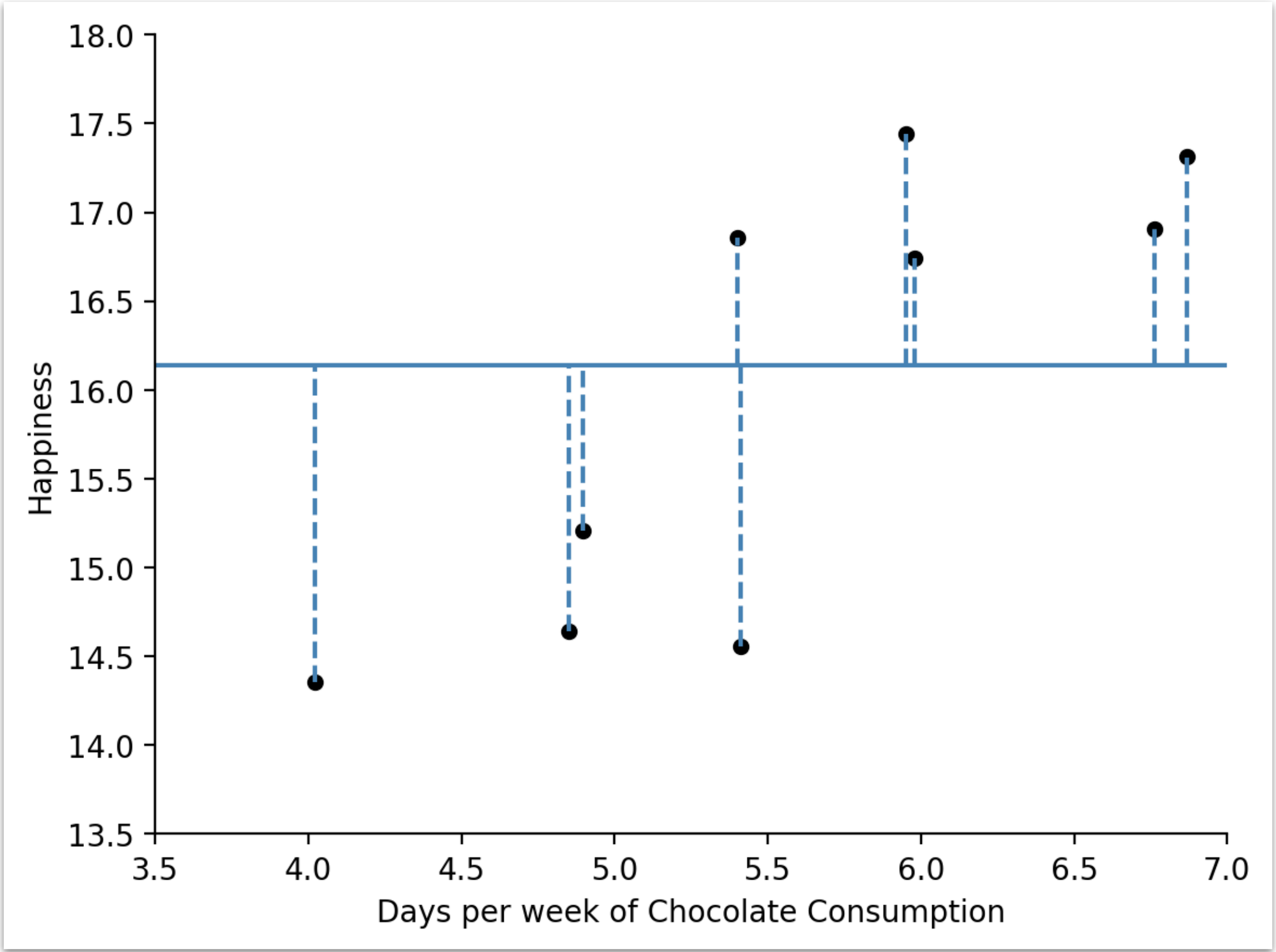
- positive and negative prediction errors don't cancel out
- larger errors are weighted more

The *mean* as a model

Is there a relationship between chocolate consumption and happiness?

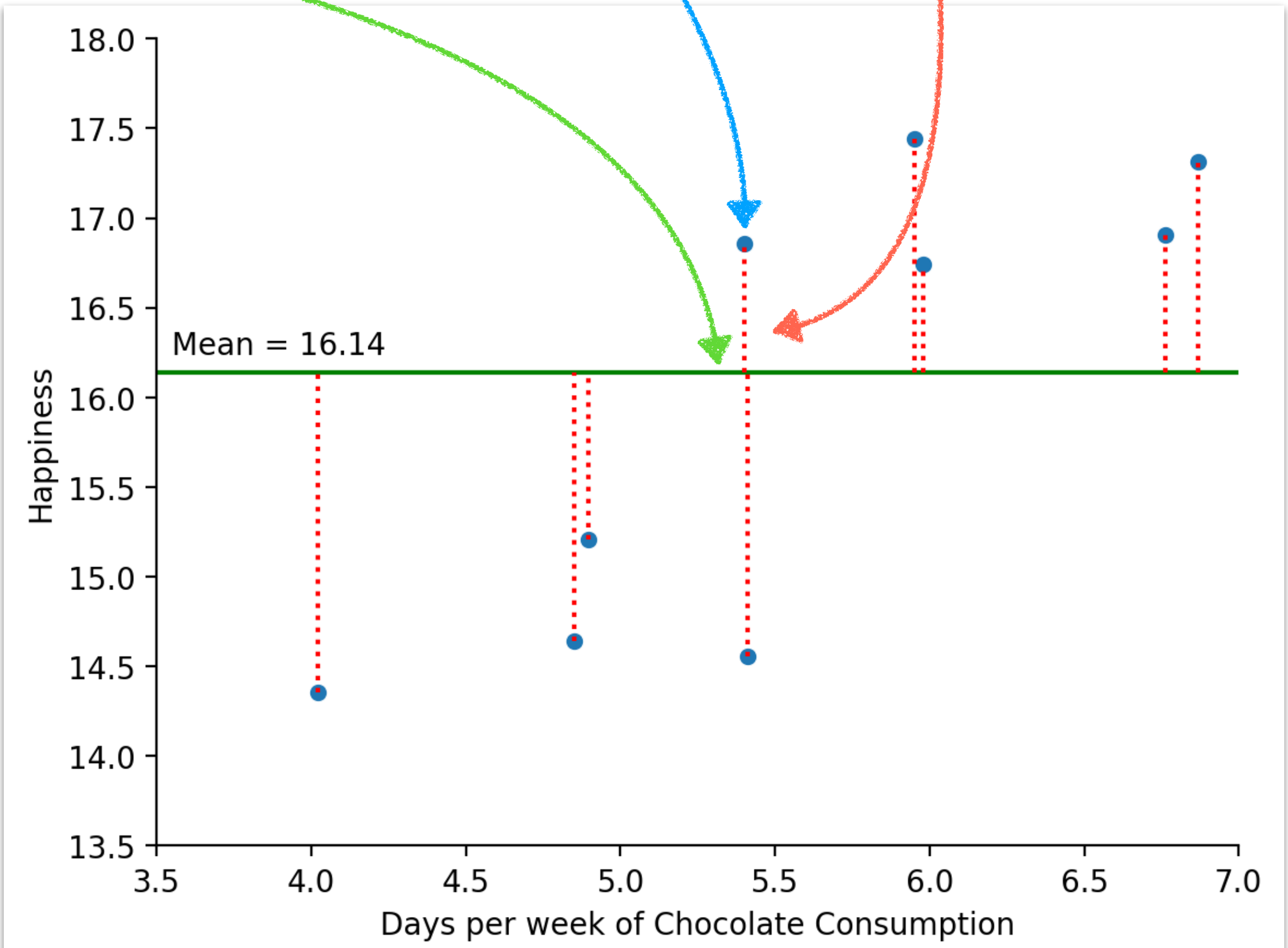


Is there a relationship between chocolate consumption and happiness?



The ***mean*** as a **model** of happiness:

$happiness_{prediction} = mean(happiness) + error$



The ***mean*** as a **model** of happiness:

$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

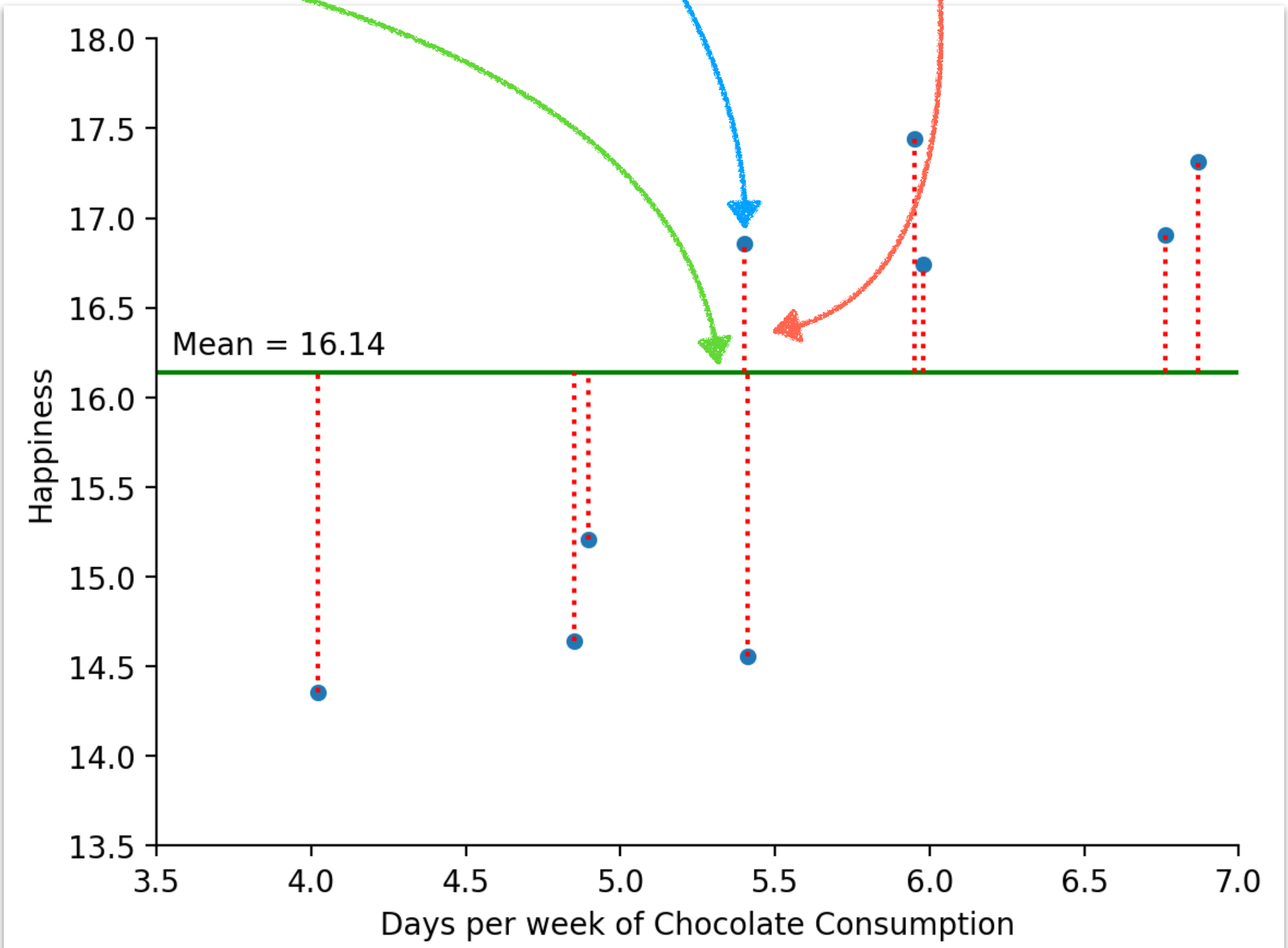
$$\texttt{happiness}_{\texttt{prediction}} = \texttt{mean(happiness)} + \texttt{error}$$

$$Y_i = \beta_0 + \epsilon_i$$

estimated from data

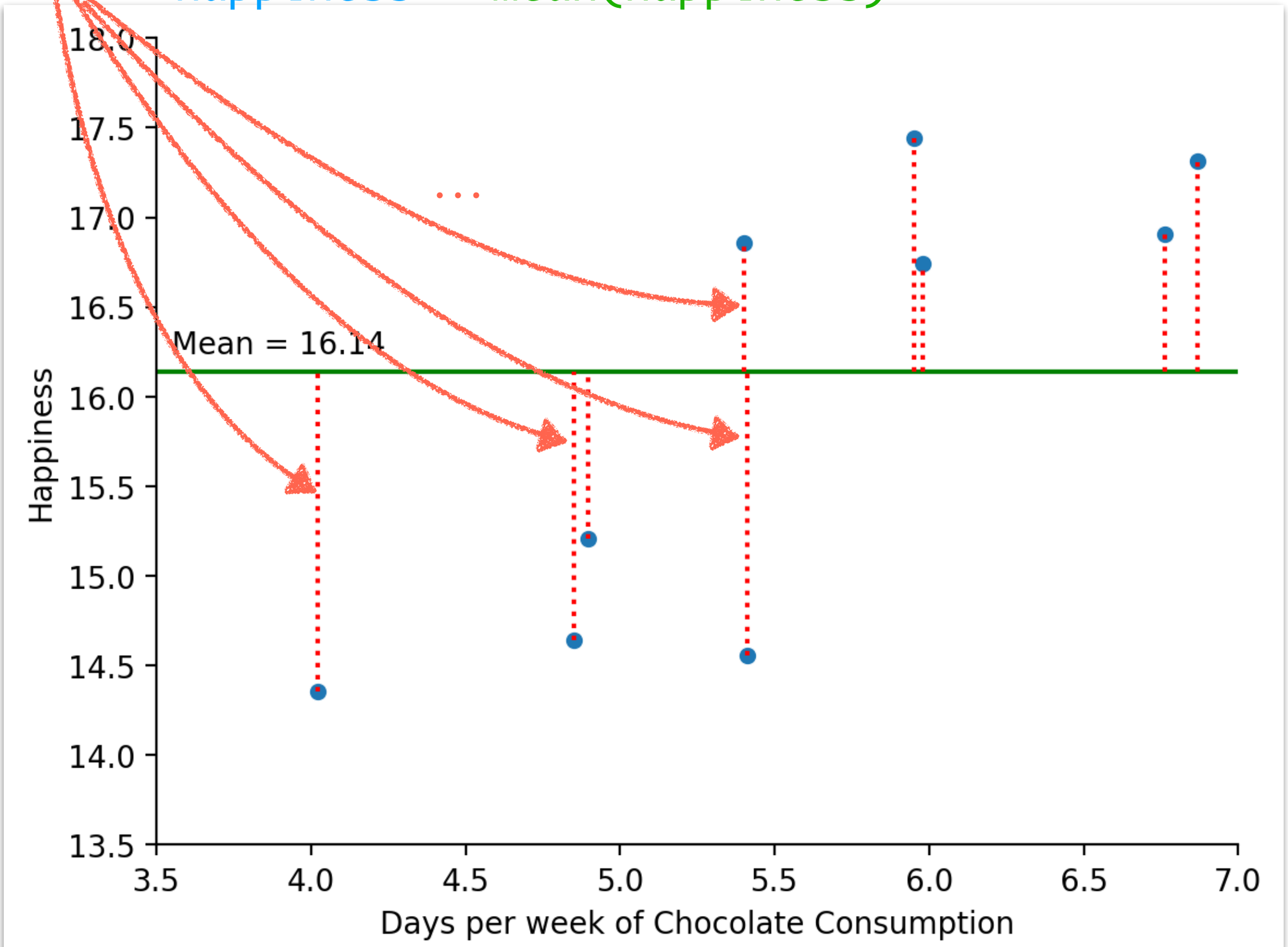
The ***mean*** as a **model** of happiness:

$happiness_{prediction} = mean(happiness) + error$



The ***variance*** as the average ***error*** of the model

$$\text{error} = \text{happiness} - \text{happiness}_{\text{prediction}}$$
$$= \text{happiness} - \text{mean}(\text{happiness})$$



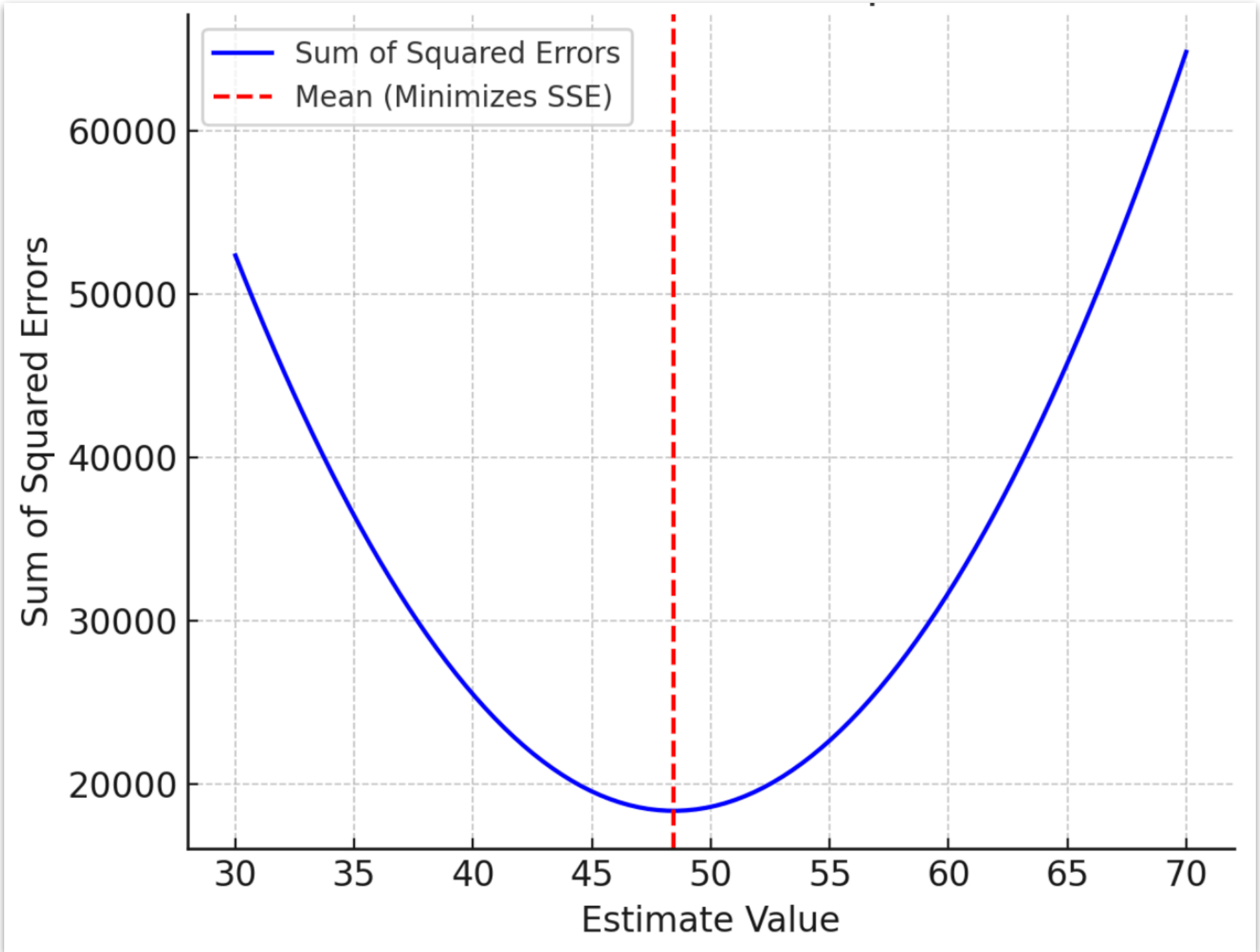
The ***variance*** as the average ***error*** of the model

model = sample mean $\bar{x}$

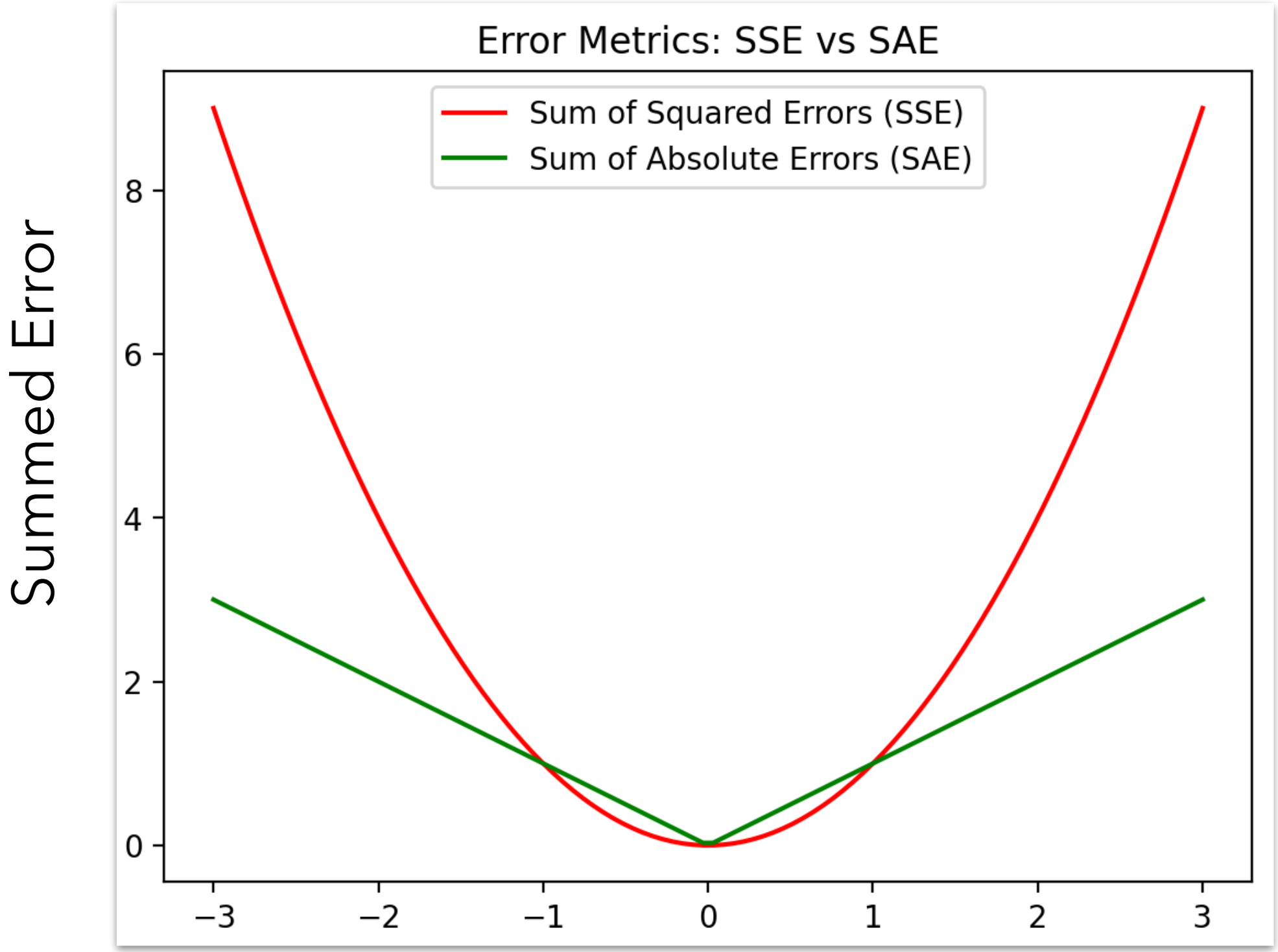
$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N} = \frac{\text{SSE}}{N} = MSE$$

error =
average distance^2 of each data point from model

The mean is the **best 1 parameter model** when best = **minimize** sum-of-squared **error**

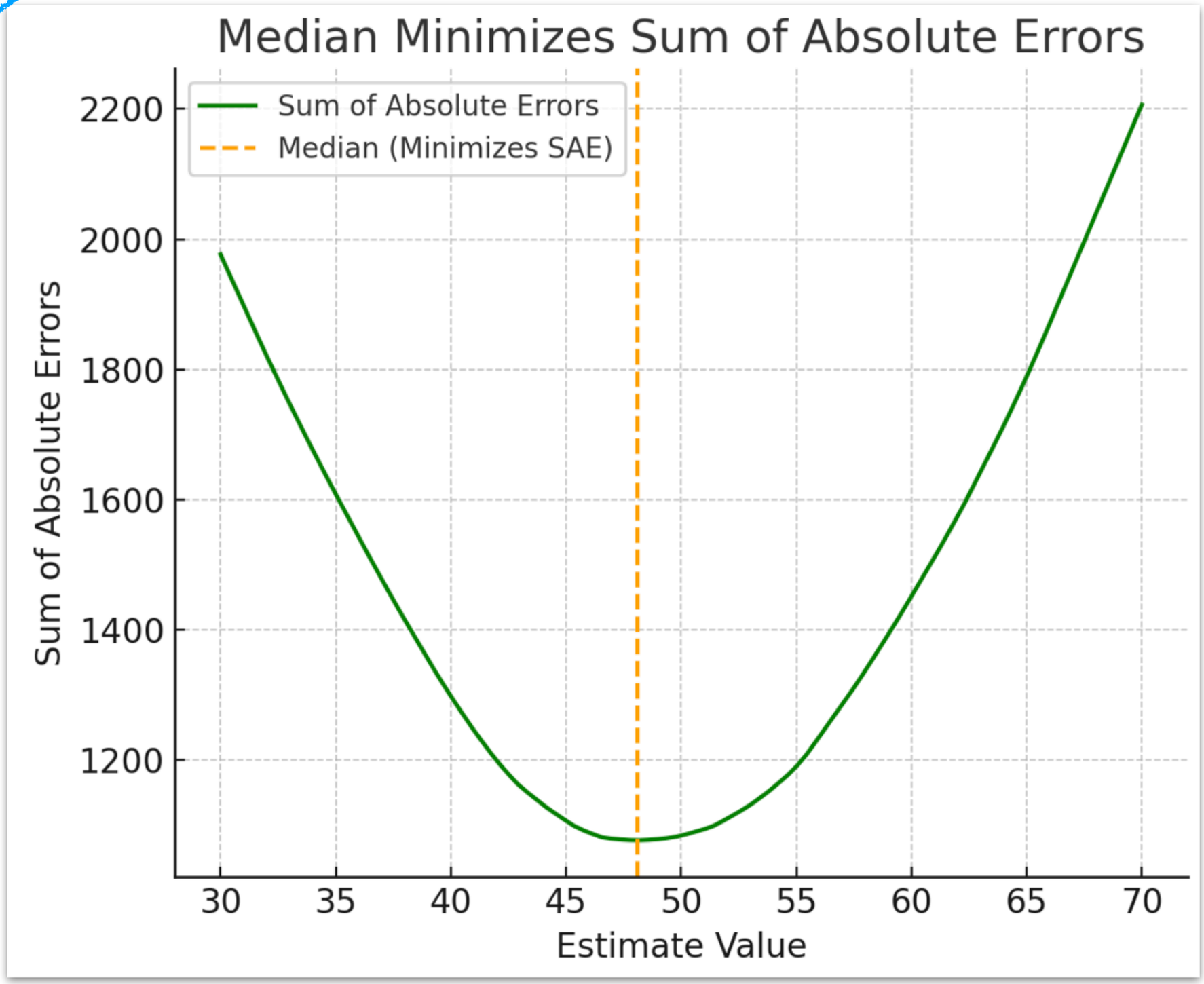


We could calculate **error** differently…



When best = minimize sum-of-**absolute error**
the *median* is the **best 1 parameter model**

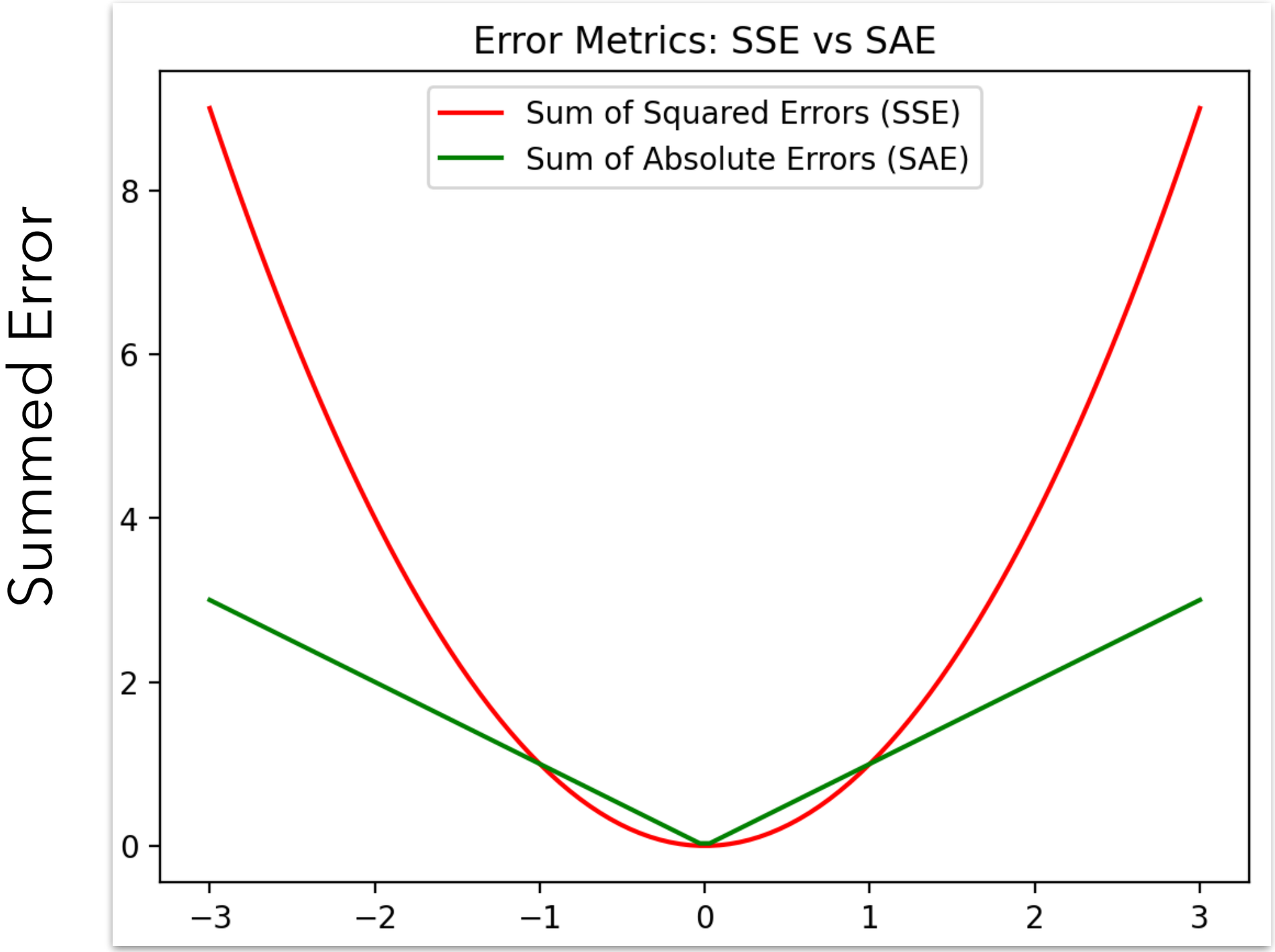
we saw this in week 2: link



Sum-of-Squared-Error (SSE), grow quadratically with error - **sensitive to outliers**

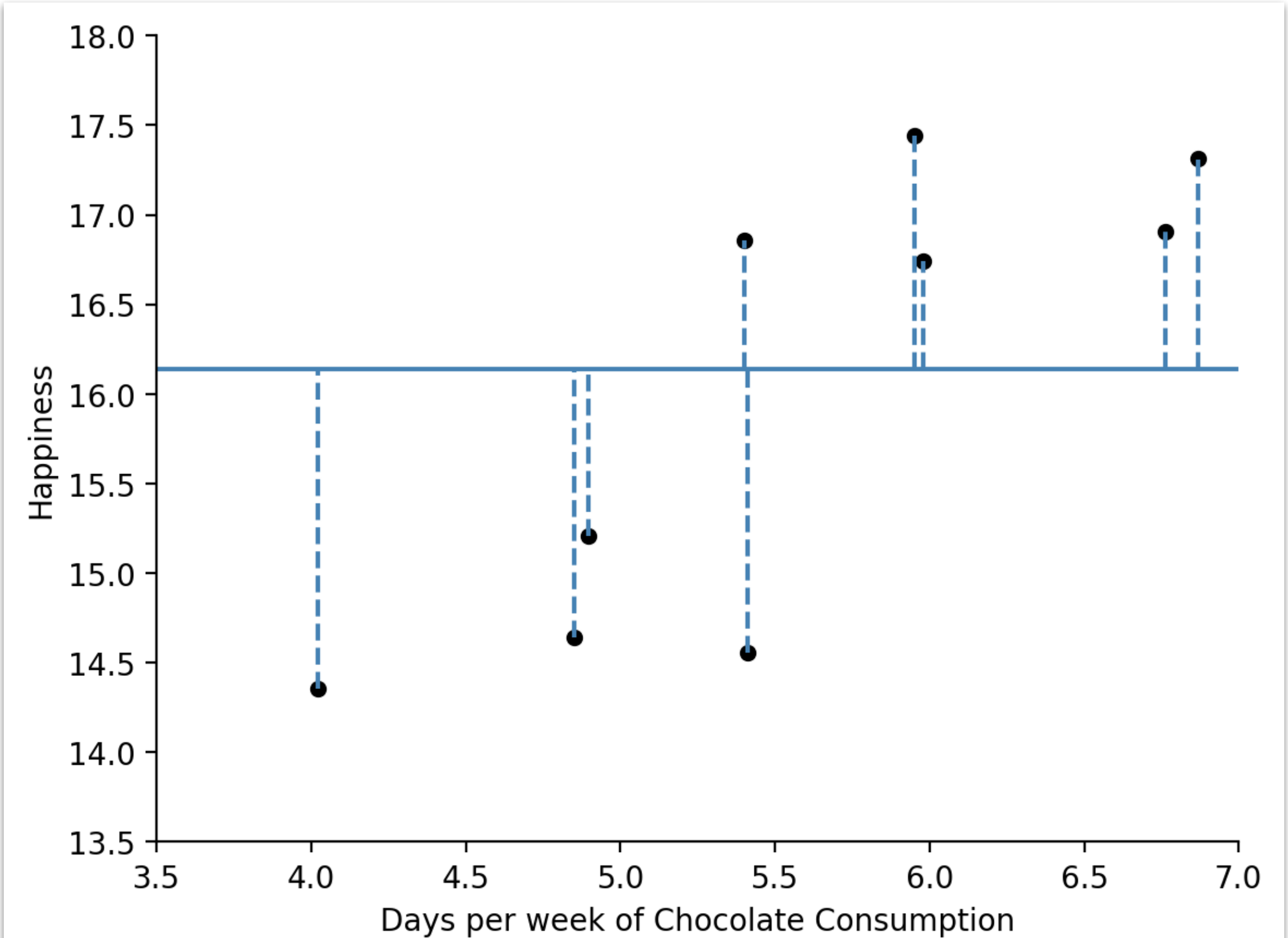
Sum-of-Absolute-Error (SAE) - grows linearly with error - **more robust to outliers**

New perspective: ***Outliers*** *are* ***Predictions*** *with very high* ***Error***



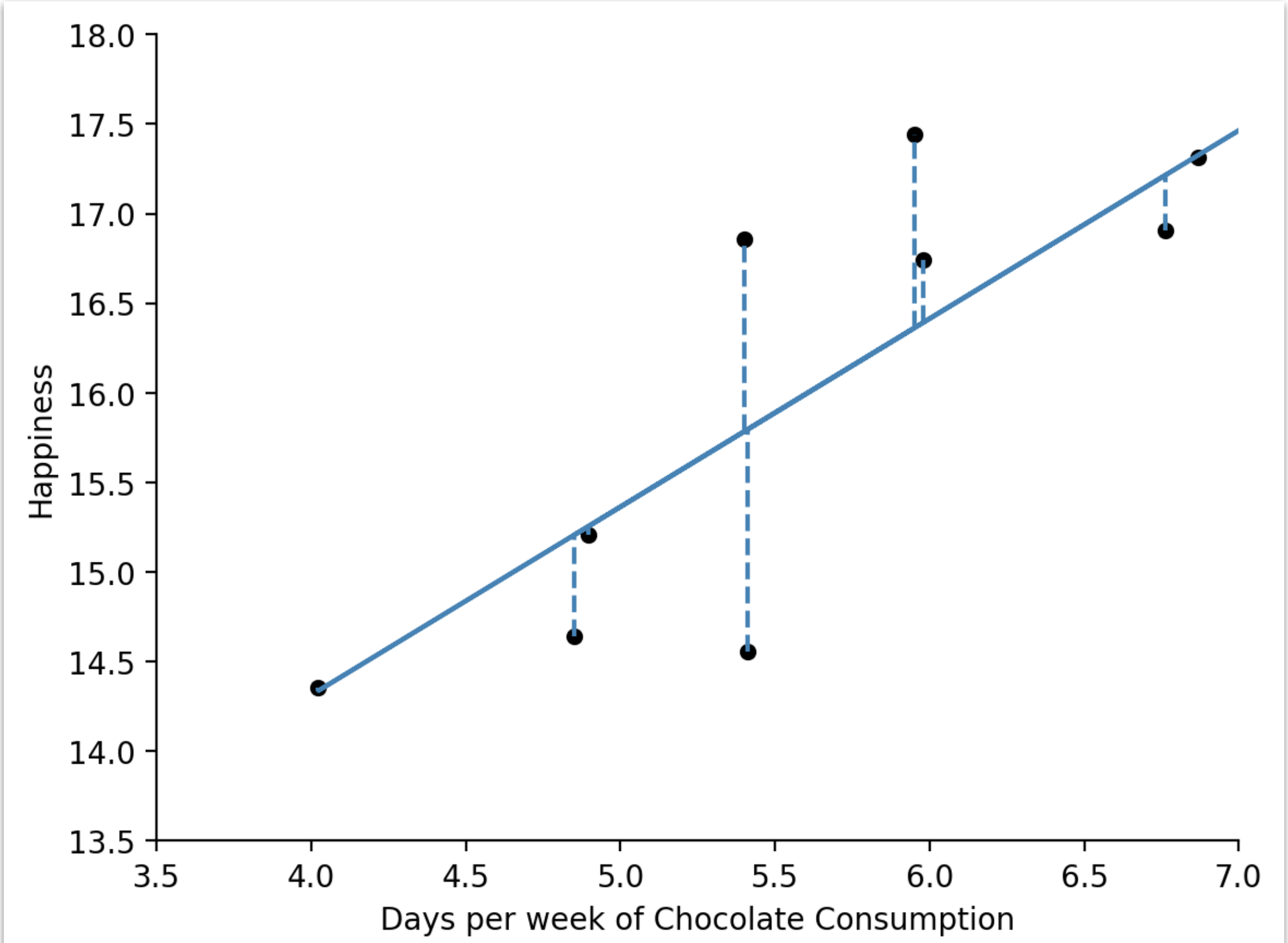
Can we do *better*?

The ***mean*** as a **model** of happiness



The ***linear*** model of happiness

(using information about chocolate consumption)



Adding a second parameter: **slope**

$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

$\texttt{happiness}_{\texttt{prediction}} = \texttt{mean}_{\texttt{happiness}} + \texttt{error}$

$$Y_i = \beta_0 + \epsilon_i$$

$\texttt{happiness}_{\texttt{prediction}} = \texttt{mean}_{\texttt{happiness}} + \texttt{slope}_{\texttt{chocolate}} + \texttt{error}$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

the model is a **linear** combination of predictors

Adding a second parameter: **slope**

$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

$\texttt{happiness}_{\texttt{prediction}} = \texttt{mean}_{\texttt{happiness}} + \texttt{error}$

$$Y_i = \beta_0 + \epsilon_i$$

worth it?

$\texttt{happiness}_{\texttt{prediction}} = \texttt{mean}_{\texttt{happiness}} + \boxed{\texttt{slope}_{\texttt{chocolate}}} + \texttt{error}$

$$Y_i = \beta_0 + \boxed{\beta_1 X_i} + \epsilon_i$$

$$Data = Model + Error$$

what makes for
a good model?

- we build models with parameters, and fit those parameters to minimize error
- adding **additional parameters** to the model will **always** improve the model fit and **reduce error**
- fundamental trade-off between **simplicity** and **accuracy** = **worth it?**

Hypothesis testing as model *comparison*

(The **worth it?** question)

The **worth it?** question

set before looking at the data

0 parameter(s) → model$_1$: $Y_i = 75 + \text{ERROR}$

worth it?

fit to the data

1 parameter(s) → model$_2$: $Y_i = \beta_0 + \text{ERROR}$

worth it?

additional predictor

2 parameter(s) → model$_3$: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

fit to the data

Compact model

model_C: $Y_i = \beta_0 + \text{ERROR}$

Augmented model

model_A: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

$$\text{ERROR}(C) \geq \text{ERROR}(A)$$

Proportional reduction in error (PRE)

$$\text{PRE} = \frac{\text{ERROR}(C) - \text{ERROR}(A)}{\text{ERROR}(C)}$$

Compact model

model_C: $Y_i = \beta_0 + \text{ERROR}$

$$\text{ERROR}(C) = 50$$

Augmented model

model_A: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

$$\text{ERROR}(A) = 30$$

Proportional reduction in error (PRE)

$$\text{PRE} = 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)}$$

$$= 1 - \frac{30}{50} = .40$$

Increasing the complexity of the model by 1 parameter reduced the error by 40%. worth it?

The **worth it?** question

PRE is the estimate of an unknown true reduction of error η^2

We need a **sampling distribution** of PRE:

- a distribution of what PRE would look like if Model C (our H_0) were *true*
- and then **compare the observed value** of PRE to that **distribution**

We *could* just simulate such a sampling distribution ...

But PRE is closely related to the ***F* statistic**!

Sampling distribution of PRE = F Statistic

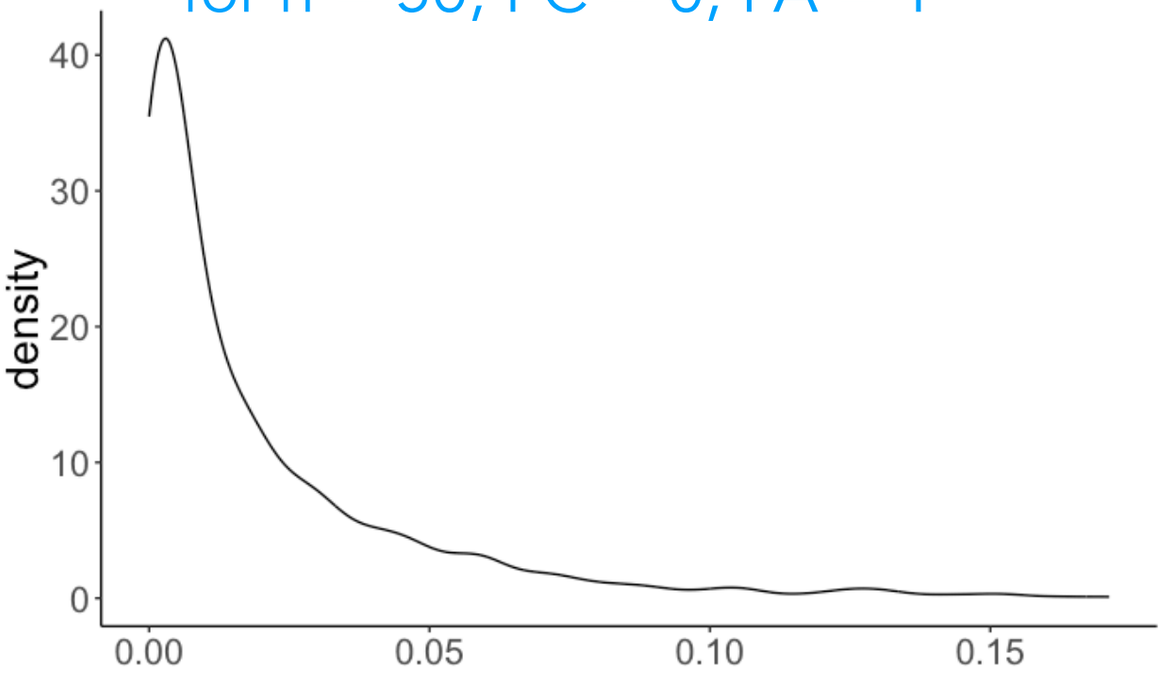
deterministic mapping

not in most textbooks

but in readings for this week!

sampling distribution of PRE

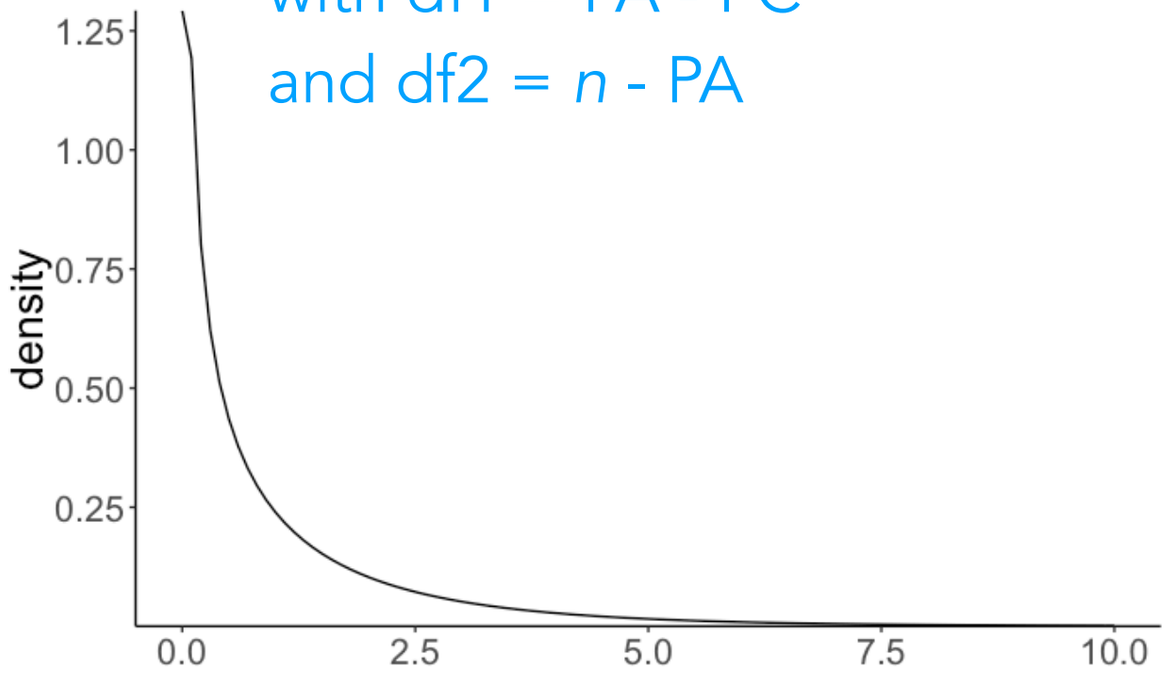
for n = 50, PC = 0, PA = 1



***F*(df1, df2) distribution**

with df1 = PA - PC

and df2 = *n* - PA



F-statistic: **ratio** of variances/errors

The **worth it?** question

Compact model

model_C: $Y_i = \beta_0 + \text{ERROR}$

Augmented model

model_A: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

Proportional reduction in error (PRE)

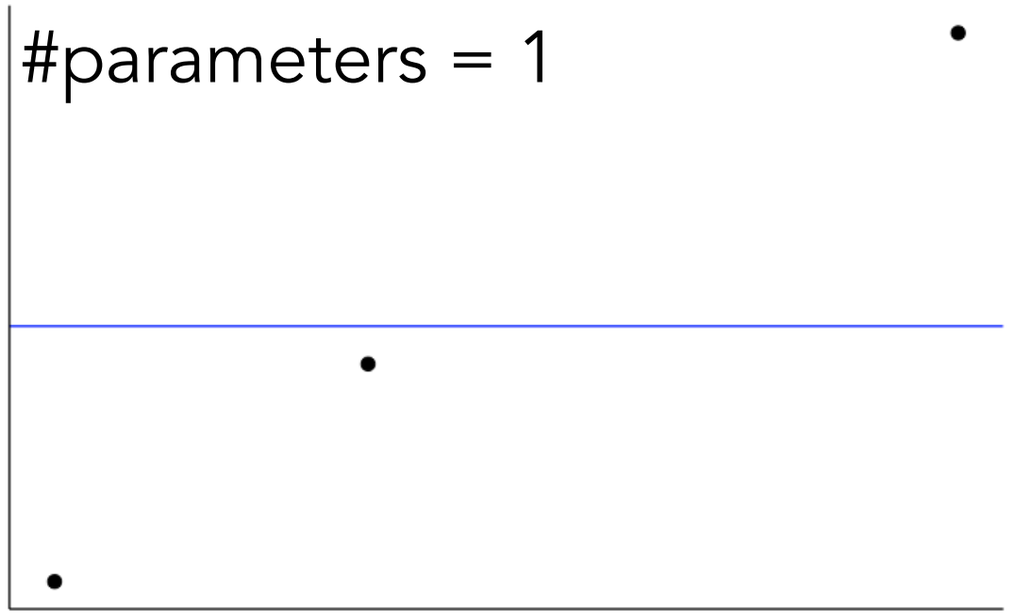
$$\text{PRE} = 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)}$$

$$= 1 - \frac{30}{50} = .40$$

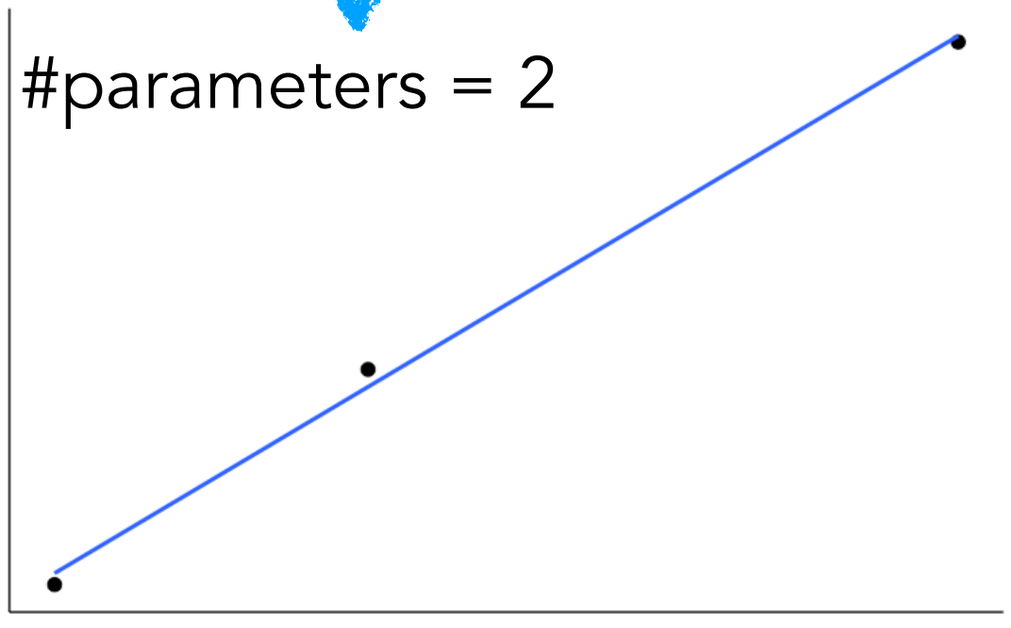
- to answer the worth it? question we need inferential statistics (define ERROR, sampling distributions, etc.)
- more likely to be worth it if:
  1. **PRE** is high
  2. the number of additional parameters in A compared to C is low (**PRE per additional parameter**)
  3. the number of parameters we *could have, but didn't use* to create model_A is high; it costs fewer *degrees-of-freedom*

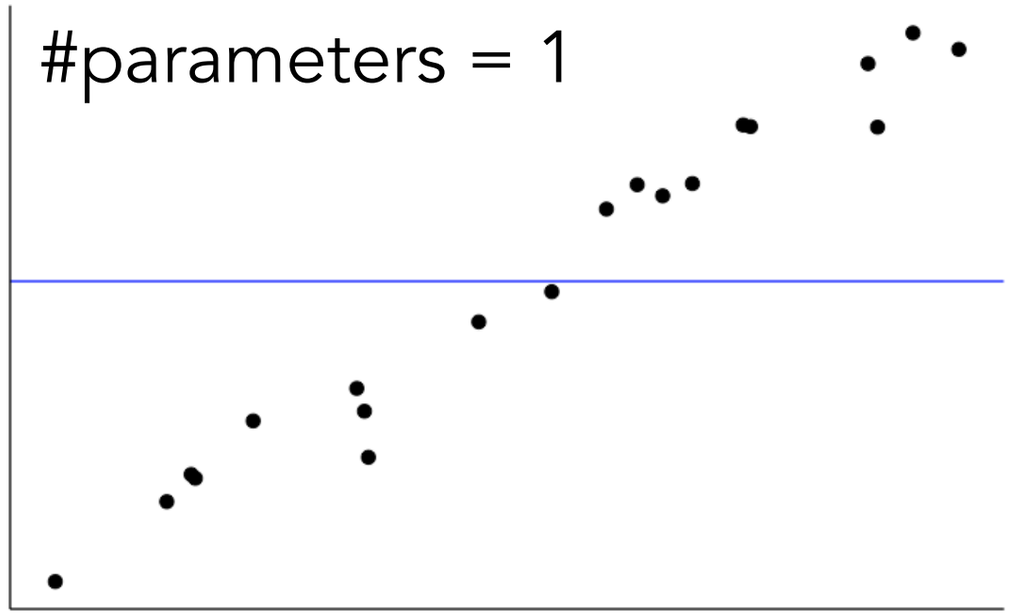
more impressive if the number of parameters (p) far fewer than number of observations (n)

PRE per parameter for different *n*

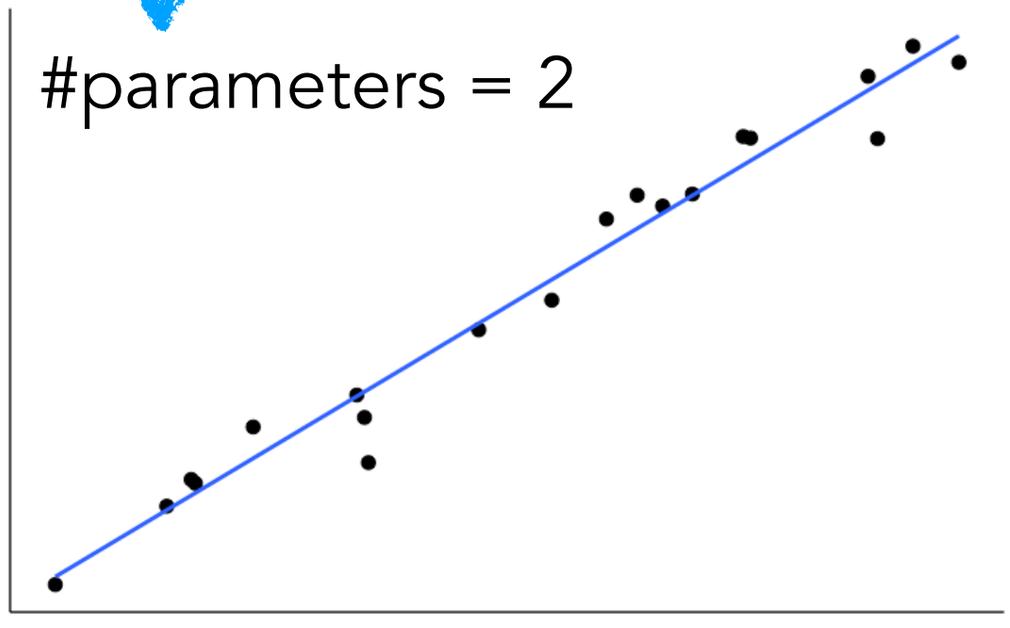


only 2 parameters
predict 3 data points
neato!





only 2 parameters
predict *lots* of data points
impressive!



Hypothesis testing as model comparison

1. Start with research question
2. Formulate **hypothesis as a comparison** between compact and augmented model
3. Fit parameters in each model
4. Calculate the proportional reduction of error (PRE) to **compare models**
5. Decide whether PRE is worth it

Hypothesis testing as model comparison

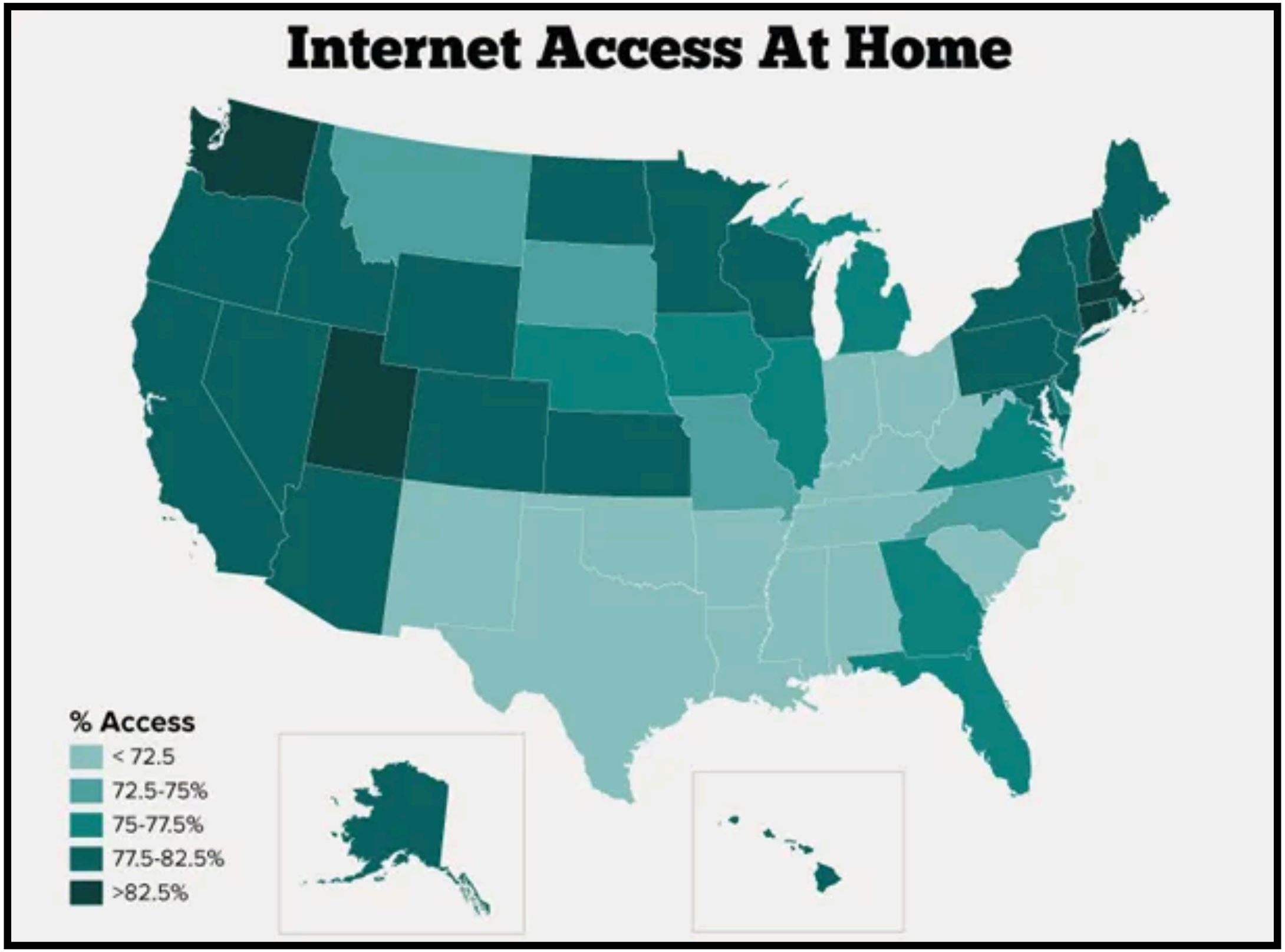
Frequentist statistics lingo:

- $model_C$ = compact model = H_0 (null hypothesis)
- $model_A$ = augmented model = H_1 (alternative hypothesis)

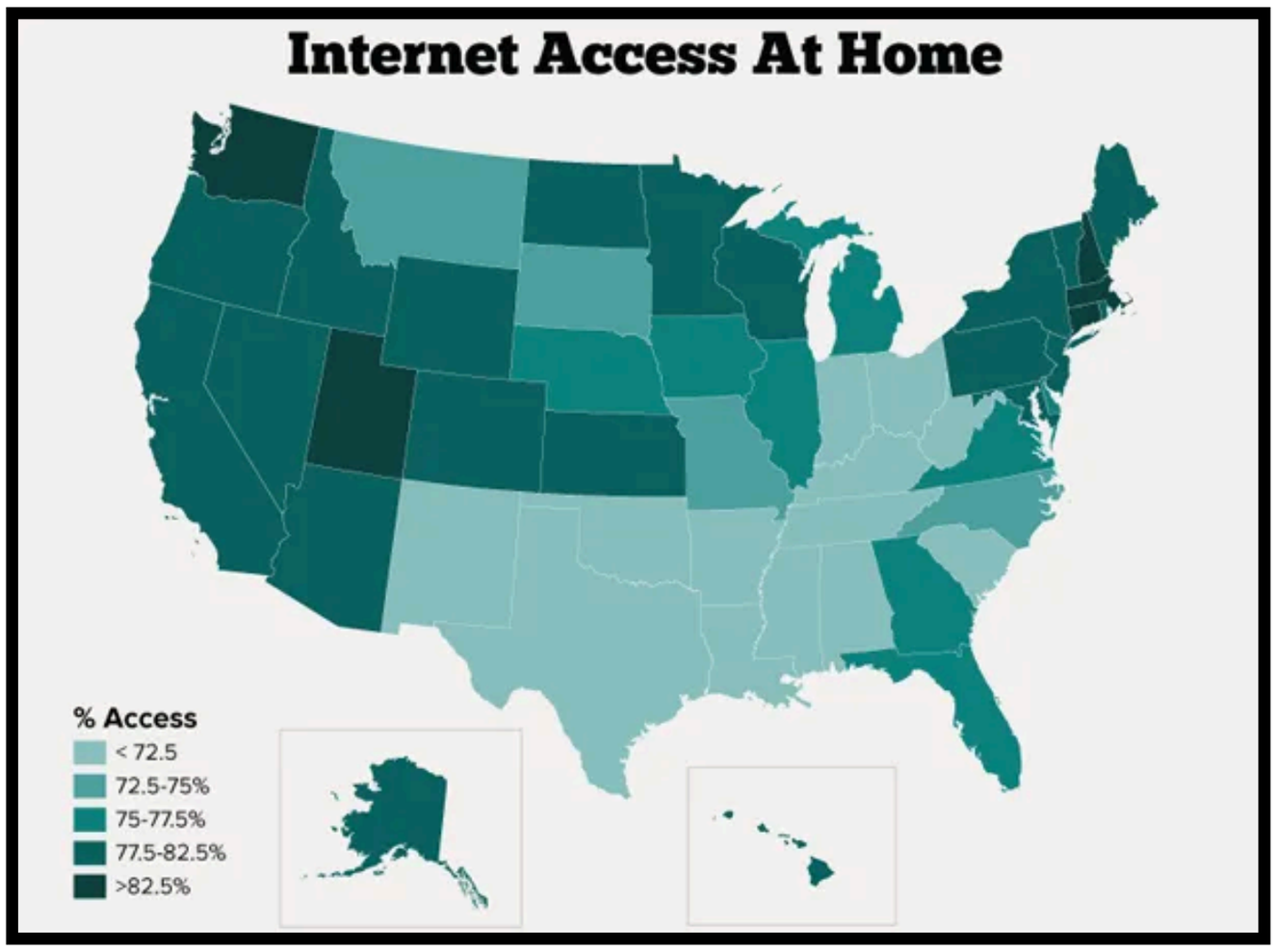
Hypothesis test:

- H_0: **all** the parameters that are included in $model_A$ but not in $model_C$ are 0
- H_1: **not all** the parameters that are included in $model_A$ but not in $model_C$ are 0

Example



http://thedailyviz.com/2012/07/19/mapping-internet-access-in-u-s-homes/



State	Internet	College	Auto	Density
str	f64	f64	f64	f64
"AK"	79.0	28.0	1.17	1.2
"AL"	63.5	23.5	1.34	94.4
"AR"	60.9	20.6	1.7	56.0
"AZ"	73.9	27.4	1.27	56.3
"CA"	77.9	31.0	0.84	239.1
"CO"	79.4	37.8	0.96	48.5
"CT"	77.5	37.2	1.02	738.1
"DE"	74.5	29.8	1.13	460.8
"FL"	74.3	27.2	1.25	350.6
"GA"	72.2	28.3	1.12	168.4

1. Research question

Is the average percentage of internet users per state significantly different from 75%?

Model$_C$: $Y_i = \beta_0 + \epsilon_i$

0 parameters

$Y_i = 75 + e_i$

Model$_A$: $Y_i = \beta_0 + \epsilon_i$

1 parameter

$Y_i = b_0 + e_i$

$= \bar{Y} + e_i$

State	Internet	College	Auto	Density
str	f64	f64	f64	f64
"AK"	79.0	28.0	1.17	1.2
"AL"	63.5	23.5	1.34	94.4
"AR"	60.9	20.6	1.7	56.0
"AZ"	73.9	27.4	1.27	56.3
"CA"	77.9	31.0	0.84	239.1
"CO"	79.4	37.8	0.96	48.5
"CT"	77.5	37.2	1.02	738.1
"DE"	74.5	29.8	1.13	460.8
"FL"	74.3	27.2	1.25	350.6
"GA"	72.2	28.3	1.12	168.4

2. Hypothesis as comparison

Population distribution?

$$Y_i = 75 + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(\mu = 0, \sigma = 5)$$

Model C

$$Y_i = 75 + e_i$$

0 parameters

vs

Model A

$$Y_i = \overline{Y} + e_i$$

1 parameter

3 & 4. Fit parameters and calculate PRE

$$\textbf{C: } Y_i = 75 + e_i \qquad \textbf{A: } Y_i = \overline{Y} + e_i$$

Y_i

State	Internet	College	Auto	Density	model_c	model_a	model_c_se	model_a_se
str	f64	f64	f64	f64	i32	f64	f64	f64
"AK"	79.0	28.0	1.17	1.2	75	72.806	16.0	38.365636
"AL"	63.5	23.5	1.34	94.4	75	72.806	132.25	86.601636
"AR"	60.9	20.6	1.7	56.0	75	72.806	198.81	141.752836
"AZ"	73.9	27.4	1.27	56.3	75	72.806	1.21	1.196836
"CA"	77.9	31.0	0.84	239.1	75	72.806	8.41	25.948836
"CO"	79.4	37.8	0.96	48.5	75	72.806	19.36	43.480836
"CT"	77.5	37.2	1.02	738.1	75	72.806	6.25	22.033636
"DE"	74.5	29.8	1.13	460.8	75	72.806	0.25	2.869636
"FL"	74.3	27.2	1.25	350.6	75	72.806	0.49	2.232036
"GA"	72.2	28.3	1.12	168.4	75	72.806	7.84	0.367236

Model A has 15% less error than Model C.

$$\text{PRE} = 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)}$$

$$= 1 - \frac{1355}{1595} \approx .15$$

SSE_C	SSE_A
1595	1355

5. Decide whether it's worth it

- To compute the *F* statistic, we need:
  - PRE
  - number of parameters in Model C (PC) and Model A (PA)
  - number of observations *n*

more likely to be worth it if:

1. **PRE** is high
2. the number of additional parameters in A compared to C is low (**PRE per additional parameter**)
3. the number of parameters we *could have, but didn't use* to create $model_A$ is high; it costs fewer *degrees-of-freedom*

difference in parameters between models A and C

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

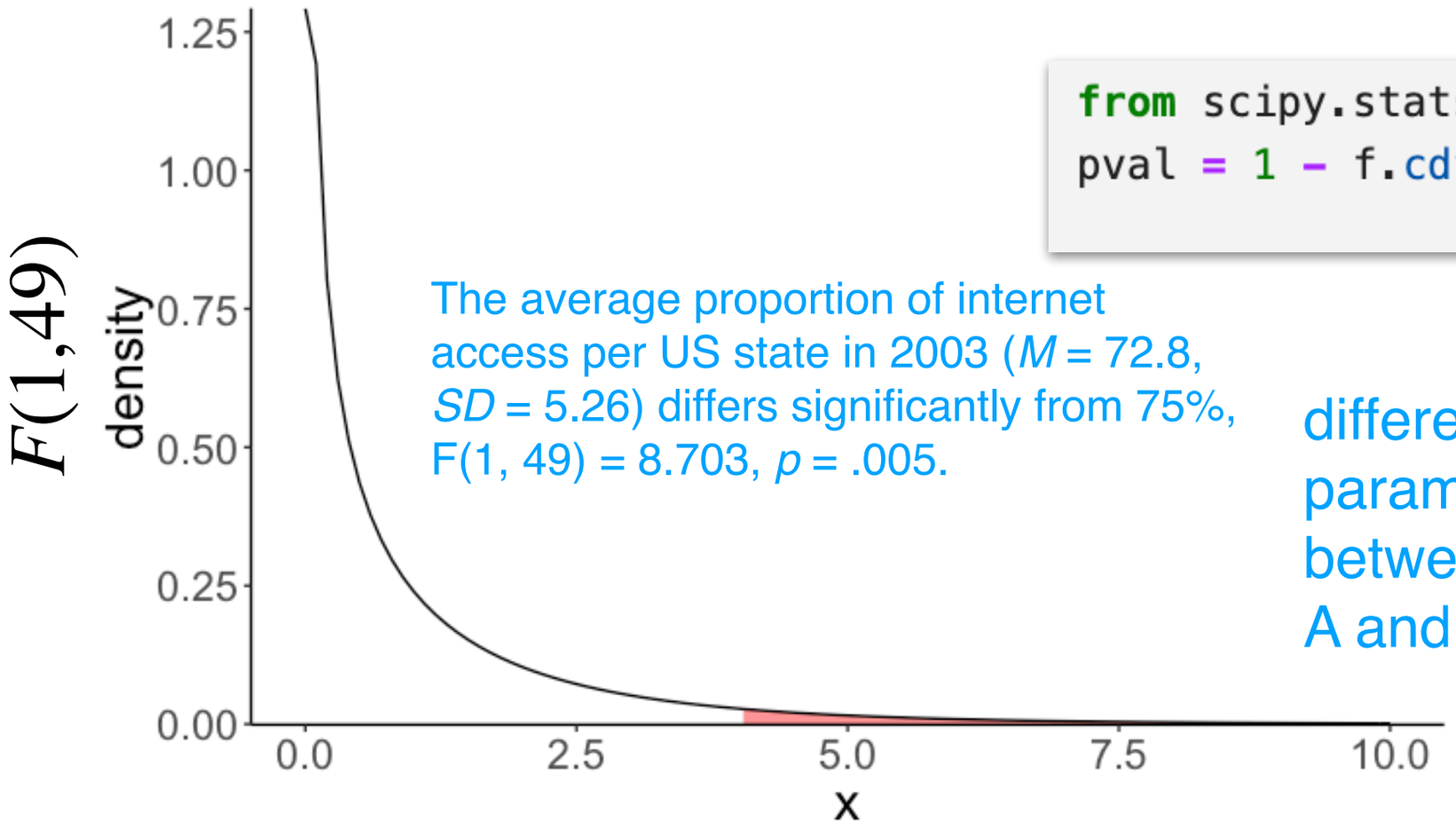
number of observations vs. parameters in Model A

5. Decide whether it's worth it

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

$$= \frac{.15/(1 - 0)}{(1 - .15)/(50 - 1)} = 8.703$$

$p = .00486$



The average proportion of internet access per US state in 2003 ($M = 72.8$, $SD = 5.26$) differs significantly from 75%, F(1, 49) = 8.703, $p = .005$.

```
from scipy.stats import f
pval = 1 - f.cdf(8.703, dfn=1, dfd=49)
```

difference in parameters between models A and C

model A observations - parameters

This is just a model comparison version of **1-sample t-test**!

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

$$= \frac{.15/(1-0)}{(1-.15)/(50-1)} = 8.703$$

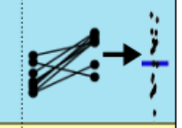
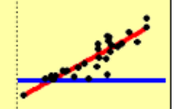
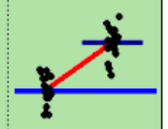
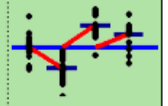
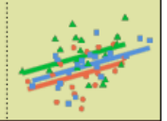
```
from scipy.stats import f
pval = 1 - f.cdf(8.703, dfn=1, dfd=49)
```

$p = .00486$

$$t = \sqrt{(F)}$$

```
from scipy.stats import ttest_1samp

t, p = ttest_1samp(df['Internet'], popmean=75)
```

Most statistical tests are just a **linear model** in disguise…

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise $\mathbf{y_2}$-$\mathbf{y_1}$ differences. - (Same, but it predicts the *signed rank* of $\mathbf{y_2}$-$\mathbf{y_1}$.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=…)[B][A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x_1 + x_2 +…)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +…+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +…+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +…+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 + … + G_N + S_2 + S_3+ … + S_K + G_2*S_2+G_3*S_3+…+G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: $\mathbf{G_{2\,to\,N}}$ is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for $\mathbf{S_{2\,to\,K}}$ for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + … + G_N + S_2 + S_3+ … + S_K + G_2*S_2+G_3*S_3+…+G_N*S_K, family=…)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments:* `glm(model, family=poisson())` *As linear-model, the Chi-square test is $\log(y_{ij}) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +…+ G_N, family=…)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for $y = 1 \cdot b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.
[B] Same model, but with one variance per group: `gls(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.

Jonas Kristoffer Lindeløv
https://lindeloev.net

Summarizing so far

- **Statistical models** allow us specify a theory for how **data** were **generated**
- Good models **balance** *complexity* - number of parameters - with *accuracy* - minimizing error of predictions
- The **mean** is among the simplest models (1-parameter) models we can estimate from data - it minimizes the **sum-of-squared-errors**
- The **variance** summarizes the average predictive error of the **mean**
- We can formulate hypotheses as **model comparison** - proportional reduction in error relative to addition of more parameters

Remember?

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Two Cultures: to explain or predict?

Explanation ⟷ **Prediction**

Classical inference / Out-of-sample generalization

Theory		
null-hypothesis testing & multiple comparisons	bias-variance decomposition	Vapnik-Chervonenkis dimensions & curse of dimensionality
degrees of freedom df		hypothesis space $\mathcal{H}$
asymptotic consistency		finite-sample theorems
Invalidations of inferential process		
double dipping/circular analysis	post-selection inference	data snooping/peeking
Outcome metrics		
(in-sample) p values sensitivity/specificity effect size/power	explained variance metrics AUC/ROC curve confidence intervals	out-of-sample prediction accuracy/precision/recall/F1 scores learning curves certainty estimates via bootstrap
Representative methods		
Student's *t*-test *F*-test ANOVA Binomial test X^2-test linear regression	general(ized) linear model	support vector machines LASSO/ridge regression/elastic net logistic regression nearest neighbors random forests kernel methods („deep") neural networks

Two Cultures: One Model

Explanation ⟷ **Prediction**

Classical inference / Out-of-sample generalization

	Theory		
	null-hypothesis testing & multiple comparisons	bias-variance decomposition	Vapnik-Chervonenkis dimensions & curse of dimensionality
	degrees of freedom df		hypothesis space $\mathcal{H}$
	asymptotic consistency		finite-sample theorems
	Invalidations of inferential process		
	double dipping/circular analysis	post-selection inference	data snooping/peeking
	Outcome metrics		
	(in-sample) p values sensitivity/specificity effect size/power	explained variance metrics AUC/ROC curve confidence intervals	out-of-sample prediction accuracy/precision/recall/F1 scores learning curves certainty estimates via bootstrap
	Representative methods		
	Student's *t*-test *F*-test ANOVA Binomial test X^2-test linear regression	**G**eneral(*ized*) **L**inear **M**odel **GLM**	support vector machines LASSO/ridge regression/elastic net logistic regression nearest neighbors random forests kernel methods („deep") neural networks

Two Cultures: One Model

$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

Two Cultures: One Model

$$\hat{Data} = Model + Error$$

$$\hat{Y} = a + bX + \varepsilon$$

Prediction

Explanation

***Explanations** depend on how we evaluate our **predictions**!*

Next time

- Keep working on HW2 - reach out for questions!
- Please do readings for this week
- Tomorrow
  - Finish Up/Review
  - Estimating models - *manually*
  - OLS: Linear algebra to the rescue
  - Intuitions about Vectors & Matrices
  - The General Linear Model