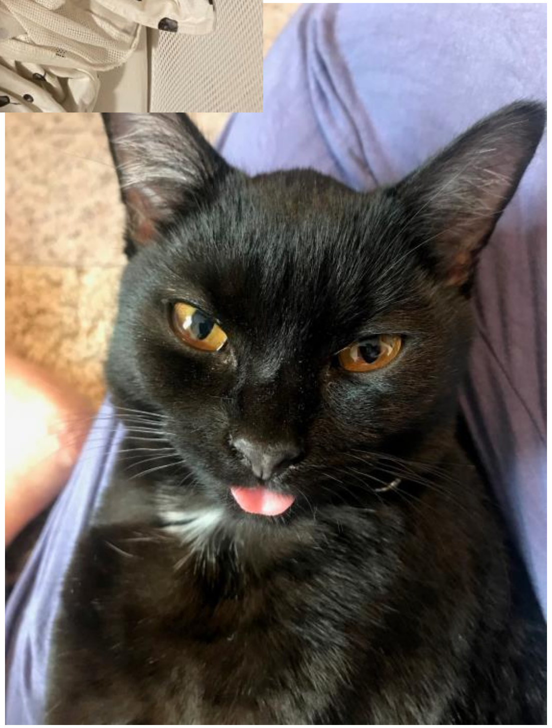
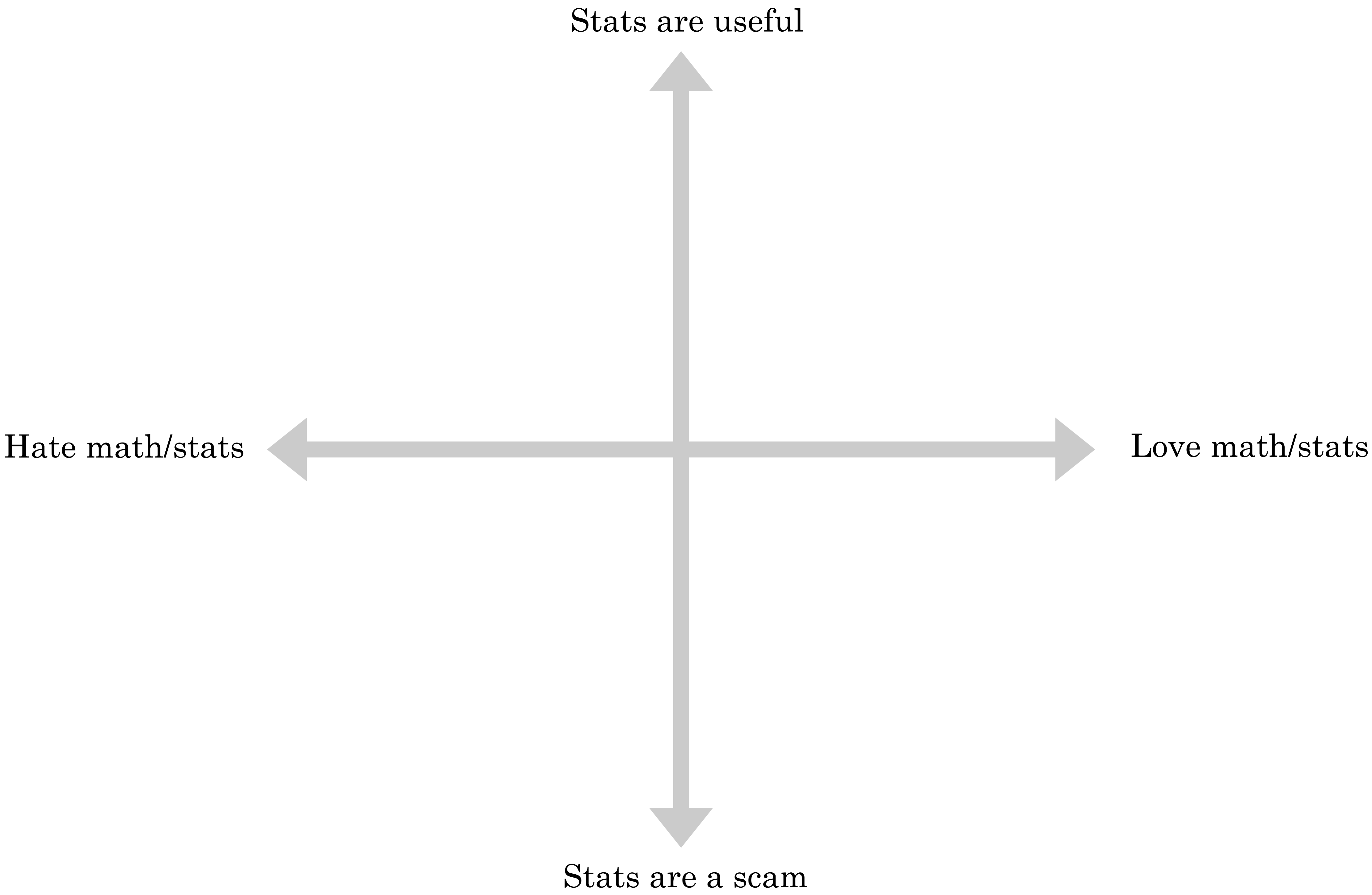
Day 1: Understanding Math and Notation


S'MOOSH YOUR




Stats are useful
Hate math/stats
Love math/stats
Stats are a scam

Goals

- Help you understand why we're doing this

Goals

- Help you understand why we're doing this
- Increase your mathematical literacy

Not goals

- Teaching you how to do calculations

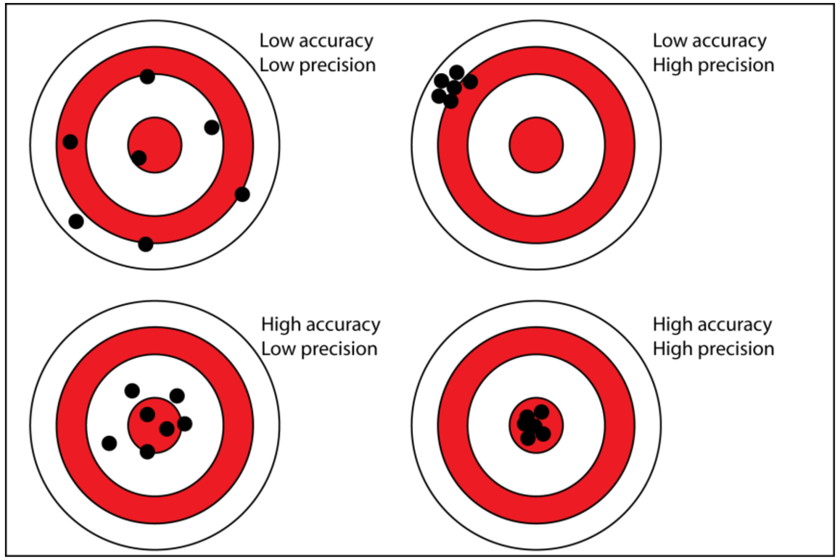
Why do you think we use math in social science?

Why do you think we use math in social science?

- Mathematics is a way of expressing relationships or theories between concepts.

Why do you think we use math in social science?

- Mathematics is a way of expressing relationships or theories between concepts.
- Mathematics is precise.
  - Note: Precise does not imply correct or good. Equating these is where a lot of quantitative research goes wrong



Big idea: Mathematics express relationships and operators define the nature of that relationship

Methods isn’t about being good at calculations. It’s about being able to precisely define your theory and the relationships between variables

Example: The Downsian Model of Voter Participation

- Anthony Downs (1957) proposed an economic model of voter participation

Utility = Probability you vote determines the outcome × perceived difference in benefit from preferred candidate − the cost of voting

Example: The Downsian Model of Voter Participation

- Anthony Downs (1957) proposed an economic model of voter participation

Utility = Probability you vote determines the outcome × perceived difference in benefit from preferred candidate − the cost of voting

Let's make this more concise:

$$U = P \times B - C$$

$U = P \times B - C$

- What are the “terms” in our equation? What does each term represent?

$U = P \times B - C$

- What are the “terms” in our equation? What does each term represent?
- Why is C being subtracted?

$U = P \times B - C$

- What are the “terms” in our equation? What does each term represent?
- Why is C being subtracted?
- Why are P and B being multiplied?

$U = P \times B - C$

- What are the "terms" in our equation? What does each term represent?
- Why is C being subtracted?
- Why are P and B being multiplied?

Note that this equation can be written more or less explicitly to emphasize different things:

- $U = P(B_1 - B_2) - C$
- $U = D - C$

How to approach mathematical models

When interpreting equations think about what each individual term means substantively, and what the operators between the terms tell you about their relationships.

How to approach mathematical models

When interpreting equations think about what each individual term means substantively, and what the operators between the terms tell you about their relationships.

What does the Downsian model imply about voting?

$U = P \times B - C$

How to approach mathematical models

When interpreting equations think about what each individual term means substantively, and what the operators between the terms tell you about their relationships.

What does the Downsian model imply about voting?

$U = P \times B - C$

That nobody should vote! Some people call this the paradox of voting. I call it a bad model.

Notation: Variables

Common Undefined Variables

Variable	Notes
x	Default value for independent variable or main variable of interest
y	Default value for dependent variable
z	When x and y isn't enough
α	Alpha, area under a curve
γ	Gamma, often some scaling factor or rate of change
θ	Theta, used in likelihood functions
λ	Lambda, often represent eigenvalues

This is just a small sample of what you will see. Different subdisciplines in statistics have different norms and commonly used variables. In general, keep it simple

$U = P \times B - C$ Revisited

- $y = px - c$

$U = P \times B - C$ Revisited

- $y = px - c$
- $y = px - z$

$U = P \times B - C$ Revisited

- $y = px \; - c$
- $y = px \; - z$
- $y = px_1 \; - x_2$

U = P × B − C Revisited

- $y = px - c$
- $y = px - z$
- $y = px_1 - x_2$
- $\upsilon = \pi x - \gamma$

Again, my advice is to keep it simple. Don’t make your readers do more work than necessary. Choose variables that remind them of the underlying concept or follow field conventions.

Notation: Vectors, Matrices, and indexing

Vector/Matrix Notation and Indexing

Vector: An ordered, one dimensional, list of numbers. Often used interchangeably with ‘array’ in computing contexts. Can be thought of as a matrix with one column.

Matrix: An ordered two dimensional set of numbers. Can be thought of as a collection of vectors.

Notation can often tell us something about the nature of the data. Capital bold letters indicate a matrix or a vector of random (unobserved) variables. Lower case letters usually indicate vectors or individual values.

What do these two different forms of our equation communicate? For what types of data would you use each?

$$\mathbf{U = PB - C}$$

$$u = pb - c$$

Indexing is used to indicate a specific element of a vector or matrix

Index	Notes
i	Primary index by convention, usually used for rows
j	Secondary index
k	Usually used for columns, dimensions, clusters, etc.
h	Less commonly used secondary index

Indexing is used to indicate a specific element of a vector or matrix

Index	Notes
i	Primary index by convention, usually used for rows
j	Secondary index
k	Usually used for columns, dimensions, clusters, etc.
h	Less commonly used secondary index

Say we collected data from a single election, how would we index it?

$$u = pb - c$$

Indexing is used to indicate a specific element of a vector or matrix

Index	Notes
i	Primary index by convention, usually used for rows
j	Secondary index
k	Usually used for columns, dimensions, clusters, etc.
h	Less commonly used secondary index

Say we collected data from a single election, how would we index it?

$$u_i = p_i b_i - c_i$$

What does our data set look like for this equation?

Another example

We want to test the downsian model so we collect a sample of 1,000 users across three elections. We assume that for each election there are different candidates, and the probability that your vote decides the election changes. The cost of voting is assumed to be constant across elections, but not individuals.

What does our equation and data look like?

Another example

We want to test the downsian model so we collect a sample of 1,000 users across three elections. We assume that for each election there are different candidates, and the probability that your vote decides the election changes. The cost of voting is assumed to be constant across elections, but not individuals.

What does our equation and data look like?

$$u_{ik} = p_{ik} b_{ik} - c_i$$

- 3 vectors with 1,000 rows for variables u,p, and b
- 1 vector with 1,000 rows for variable c

Notation: Models

Common Defined Variables

Variable	Notes
μ	Mu, the population mean
$\overline{X}$	X-bar, the sample mean
β	Beta, the estimated parameters of a linear model. Intercept and coefficients
$\hat{x}, \hat{y}, \hat{u}, \hat{s}$	Hat, the estimated value of a given variable
u, e	The error term in a model. The distance between a predicted value and an observed value.
σ	Sigma, Population standard deviation. σ^2 is the population variance.
s	Sample standard deviation. s^2 is the sample variance
N	Population
n	Sample population

Theorizing Downs

We want to test the downsian theory of turnout by predicting the percent of elections individuals turn out for.

We hypothesize a model with the following variables:

y: the percent of elections they turned out for in their life

x: The utility gained by if their preferred party wins, weighted by the probability that their vote decides an election.

c: The average cost of voting across elections

Expressing Downs

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + u_i$$

This model is expressing a theory. How do I know this?

Expressing Downs

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \ u_i$$

This model is expressing a theory. How do I know this?

- Capital variables indicate random (unobserved) variables. Use lowercase variables if this is vector you have actual values for.
- Absence of hats on parameters

Expressing Downs

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \ u_i$$

This model is expressing a theory. How do I know this?

- Capital variables indicate random (unobserved) variables. Use lowercase variables if this is vector you have actual values for.
- Absence of hats on parameters

Why is cost now being added rather than subtracted?

Expressing Downs

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \ u_i$$

This model is expressing a theory. How do I know this?

- Capital variables indicate random (unobserved) variables. Use lowercase variables if this is vector you have actual values for.
- Absence of hats on parameters

Why is cost now being added rather than subtracted?

- Subtraction is just the addition of negative numbers. In our model we are trying to estimate the effect of cost. We don’t assume its negative or positive before estimating the value so we just use addition by convention. If higher cost has a negative effect, β_2 will be a negative number and it will be the same as subtracting the cost.

Expressing Downs

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + u_i$$

What is u_i ?

Expressing Downs

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \; u_i$$

What is u_i ?

- u_i is random variation that our model does not account for

Expressing Downs

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \ u_i$$

What is u_i ?

- u_i is random variation that our model does not account for

What is the role of β_0 in the equation? In other words, what is its relationship with Y?

Expressing Downs

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \ u_i$$

What is u_i ?

- u_i is random variation that our model does not account for

What is the role of β_0 in the equation? In other words, what is its relationship with Y?

- It is the value Y takes when all other values are zero. Think of it as the starting point.

Expressing Downs

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \ u_i$$

What is u_i ?

- u_i is random variation that our model does not account for

What is the role of β_0 in the equation? In other words, what is its relationship with Y?

- It is the value Y takes when all other values are zero. Think of it as the starting point.

What is the role of β_1 and β_2?

Expressing Downs

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \ u_i$$

What is u_i ?

- u_i is random variation that our model does not account for

What is the role of β_0 in the equation? In other words, what is its relationship with Y?

- It is the value Y takes when all other values are zero. Think of it as the starting point.

What is the role of β_1 and β_2?

- They determine how much of an effect X and C have on Y.

What's the substantive difference between these equations?

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + u_i$$

$$y_i = \widehat{\beta_0} + \widehat{\beta_1} x_i + \widehat{\beta_2} c_i + \widehat{u_i}$$

What does this third equation tell us?

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \ u_i$$

$$y_i = \widehat{\beta_0} \ + \widehat{\beta_1} x_i + \widehat{\beta_2} c_i + \widehat{u}_i$$

$$\widehat{y}_i = \widehat{\beta_0} \ + \widehat{\beta_1} x_i + \widehat{\beta_2} c_i$$

What does this third equation tell us?

$$y_i = \widehat{\beta_0} + \widehat{\beta_1} x_i + \widehat{\beta_2} c_i + u_i$$

$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i + \widehat{\beta_2} c_i$$

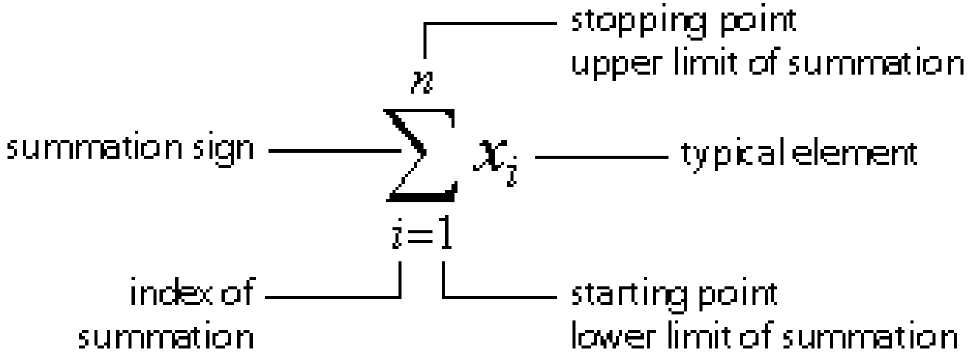
$$y_i = \hat{y}_i + \hat{u}_i$$

Or

$$\hat{u}_i = \hat{y}_i - y_i$$

Some operators you'll need…

Operator	
$\sum_{i=1}^{n} x$	Summation. The sum of numbers in a vector,
$\prod_{i=1}^{n} x$	Product operator. The product of operators in a vector



Population Mean	Sample Mean
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$ N = number of items in the population	$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$ n = number of items in the sample