Labor Economics:
Intro to Empirics

Motaz Al-Chanati

Summer 2019

Outline

1. Understanding Data

2. Hypothesis Testing

3. Regressions

Understanding Data

Data

- Empirical studies are all about using data
  - Think about data as a table (e.g. an Excel spreadsheet)
  - Easier to know what is going on if you can visualize the table (dataset) the researchers are using
- Table has two aspects:
  - Each row is an **observation**, e.g. a person, a county, a person at a particular year
  - Each column is a **variable**: a characteristic or measurement of the observation, e.g. the person's date of birth, the county's population, the person's income in that year

Dataset

	Var 1	Var 2	Var 3	Var 4
Obs 1				
Obs 2				
Obs 3				
Obs 4				

Dataset

	Age	Education	Income	State
Jane, 2009	31	College	50,000	CA
Jane, 2010	32	College	53,000	CA
Ali, 2009	19	HS	4,000	NY
Ali, 2010	20	HS	4,000	PA

- What is the unit of observation?
  - Person-year level
  - This is an example of **panel data** (can track people/units over time)

Notation

- Denote variables by a letter, e.g. X
- Denote observation by subscripts, e.g. X_i indicates the value of variable X for observation i
  - If unit of observation has multiple aspects use multiple subscripts
  - Example: take a person-year level data. X_{it} is the value of variable X for the observation where the person is i and the year is t
- Simply put: the variable tells you which column to look in, the subscript tells you which row. The intersection is the cell (value) that you want

Types of Variables

- **Discrete:** Variables that take on whole number values (e.g. counts)
  - *Dummy variables:* Special case, they can only be 0 or 1 (binary)

- **Continuous:** Variables that take on a range of values
  - For measurement purposes, continuous variables are often recorded in a discrete way (e.g. age, income)

- **Categorical:** Variables that take on a set of possible values (e.g. marital status, county)
  - Often there is a catch-all "other" category (e.g. race)

  - Can interpret some categorical variables as dummies, e.g. $X_i \in \{\text{Female, Not Female}\}$ then let 1 = Female and 0 = Not Female

Types of Variables

	Age	Education	Income	State	Male	# Kids
Jane, 2009	31	College	50,000	CA	0	1
Jane, 2010	32	College	53,000	CA	0	1
Ali, 2009	19	HS	4,000	NY	1	0
Ali, 2010	20	HS	4,000	PA	1	1

	Cont.	Categ.	Cont.	Cat.	Dummy	Discrete

Describing Data

- Before getting into fancy techniques, we first want to understand the data
  - Especially important since we often don't see the data ourselves (this is a whole other problem!)
  - Example tables were 4×4. Actual datasets could have millions of observations! How to process that complexity?
- Analogy: imagine someone is telling you about their friend for the first time. What information would they first tell you?
  - Likely candidates: name, background, job, where they to school, hometown, hobbies, personality
  - Data is your new friend! You are introduced to it through **descriptive** or **summary statistics**

Summary Statistics

- Summary statistics take a variable (X) or a set of variables (e.g. X and Y) and gives us a number to characterize them
  - Let i be the row index. Suppose there are N observations (i.e. $i = 1, \ldots, N$)
- Lots of possible statistics:
  - Statistics about one variable (X)
  - Statistics about one variable (X), *conditional* on another variable (Y)
  - Statistics about the relationship between two variables (X and Y)

One-Variable Statistics

- **Mean (Average):** The mean (often denoted as $\bar{X}$) is:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

  The mean tells us the "central tendency" of a variable

- **Variance:** The variance (often denoted as σ_X^2) tells us how much the data varies around the mean (i.e how disperse are the values)
  - The variance is never negative: $\sigma_X^2 > 0$
  - The higher the variance, the higher the dispersion
  - Similar measure: standard deviation is the square root of variance, $\sigma_X = \sqrt{\sigma_X^2}$

One-Variable Statistics

- **Percentiles:** Order all the numbers in the column from smallest to largest. Pick how far down you want to go down the list (as a percentage)
  - 0% = smallest number, 100% = largest number
  - 50% = middle number, a.k.a the **median**
- Often, we group percentiles into **quantiles**
  - Quartiles: 4 groups (0-25%, 25-50%, 50-75%, 75-100%)
  - Quintiles: 5 groups (0-20%, 20-40%, 40-60%, 60-80%, 80-100%)
  - Deciles: 10 groups (0-10%,..., 90-100%)

Conditional Statistics

- Sometimes we are interested in a statistic of *X* for a subset of the data
  - e.g. “what is the mean income, conditional on sex being female?”
  - Picture this as taking only the rows that meet your condition, and then calculating the statistic on the new (smaller) dataset

Income	Sex
50,000	M
12,500	F
98,500	F
63,000	M
39,100	M
82,200	F

Mean: 57,550

Income	Sex
12,500	F
98,500	F
82,200	F

Mean: 64,400

Two-Variable Statistics

- If we have two variables, X and Y, we often want to learn about their relationship: how much do X and Y move together?
- **Covariance:** σ_{XY} or $Cov(X, Y)$ tells us this
  - $\sigma_{XY} > 0$ (< 0) means positive (negative) relationship: higher X associated with higher (lower) Y
  - $\sigma_{XY} = 0$ means there is no relationship between X and Y
  - Covariance depends on units, e.g. if X was in dollars, but then we change it into cents, the covariance with Y would change (even though the real association hasn't changed)

Two-Variable Statistics

- **Correlation:** a units-free measure of the association between two variables, denoted as ρ_{XY} or $corr(X, Y)$

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

  - Ranges from -1 (perfectly negative relationship) to 1 (perfectly positive relationship)

Plots

- Statistics are nice because they provide a lot information with just one number
  - But they are not perfect
  - Mean can be swayed by extreme values - may not be so “central”
  - Correlation is just an average relationship - could have a lot of variation
- Best way to represent a lot of data: **graphs!**

Distributions

- **Distribution:** How often does each value in the variable come up?
- To show on a graph:
  - On the x-axis: the possible values
  - On the y-axis: a count of how many times in appears in the column (or represent it as a fraction of total observations)
- Different graphs for different types of variables:
  - Discrete/Categorical: bar chart (each value is separate)
  - Continuous: histogram (group values) or density plot (line)

Example: Math Scores Data

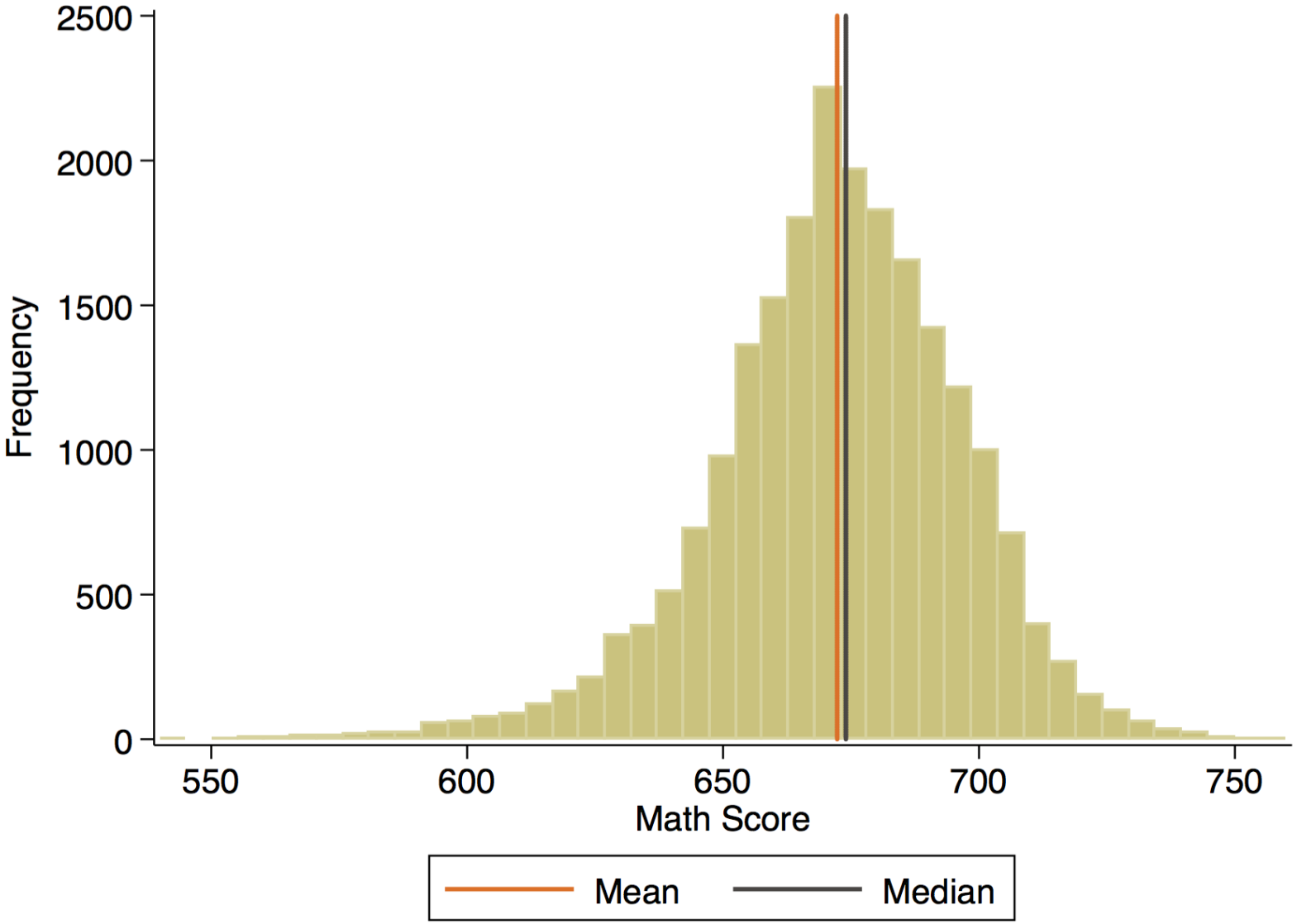
- From NYC Open Data: mean math scores by school-grade-year

Figure 1: Sample Data

	school	grade	year	numbertested	sat
1	01M015	3	2006	39	667
2	01M015	3	2007	31	672
3	01M015	3	2008	37	668
4	01M015	3	2009	33	668
5	01M015	3	2010	26	677
6	01M015	3	2011	28	671
7	01M015	4	2006	49	629
8	01M015	4	2007	40	659
9	01M015	4	2008	41	655
10	01M015	4	2009	39	655
11	01M015	4	2010	29	663
12	01M015	4	2011	28	668
13	01M015	5	2006	31	630
14	01M015	5	2007	50	637
15	01M015	5	2008	36	660
16	01M015	5	2009	39	661
17	01M015	5	2010	35	662
18	01M015	5	2011	25	667
19	01M015	6	2006	39	639

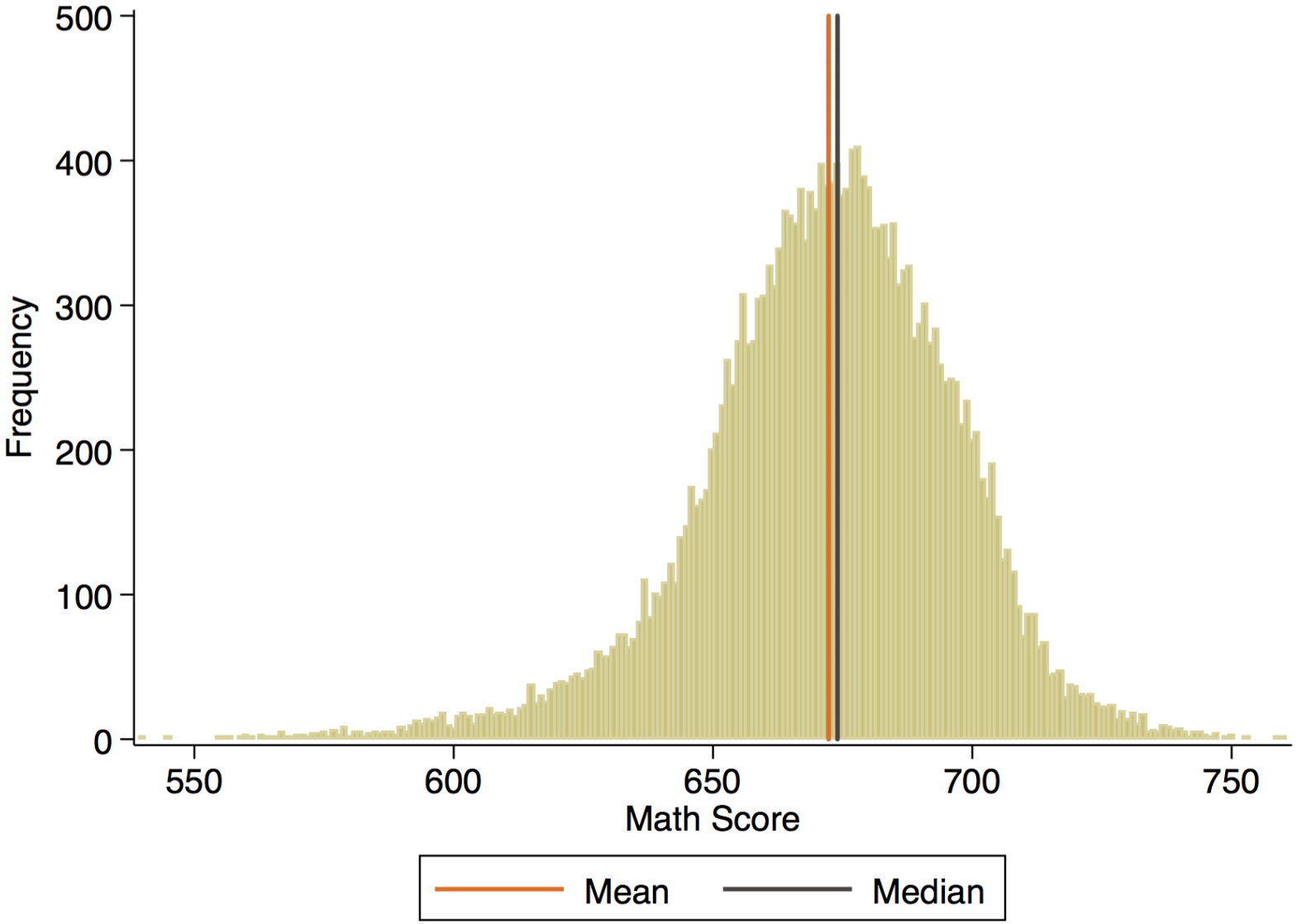
Math Score Distribution

Figure 2: Score Distribution



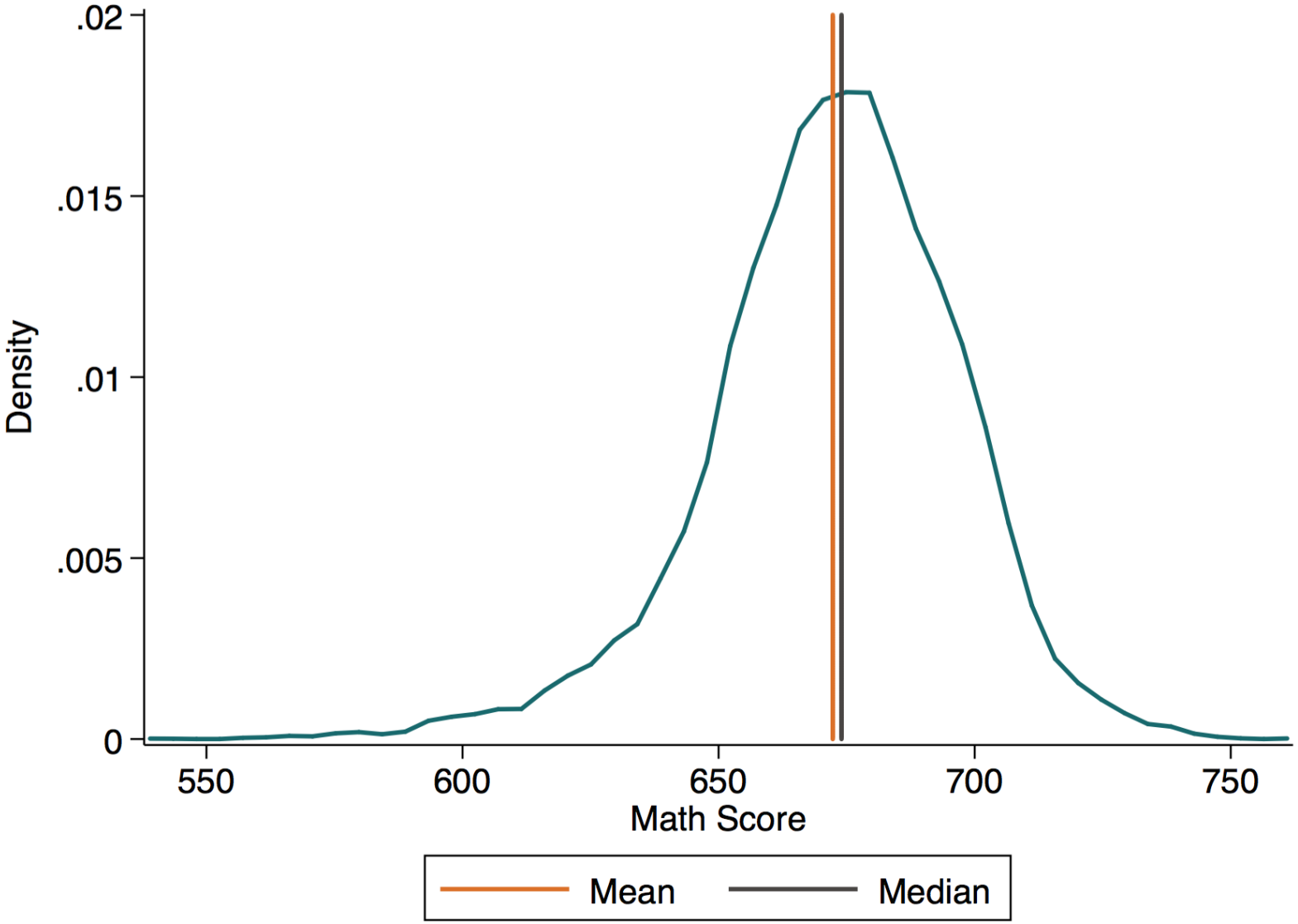
Math Score Distribution

Figure 3: Score Distribution (Discrete)



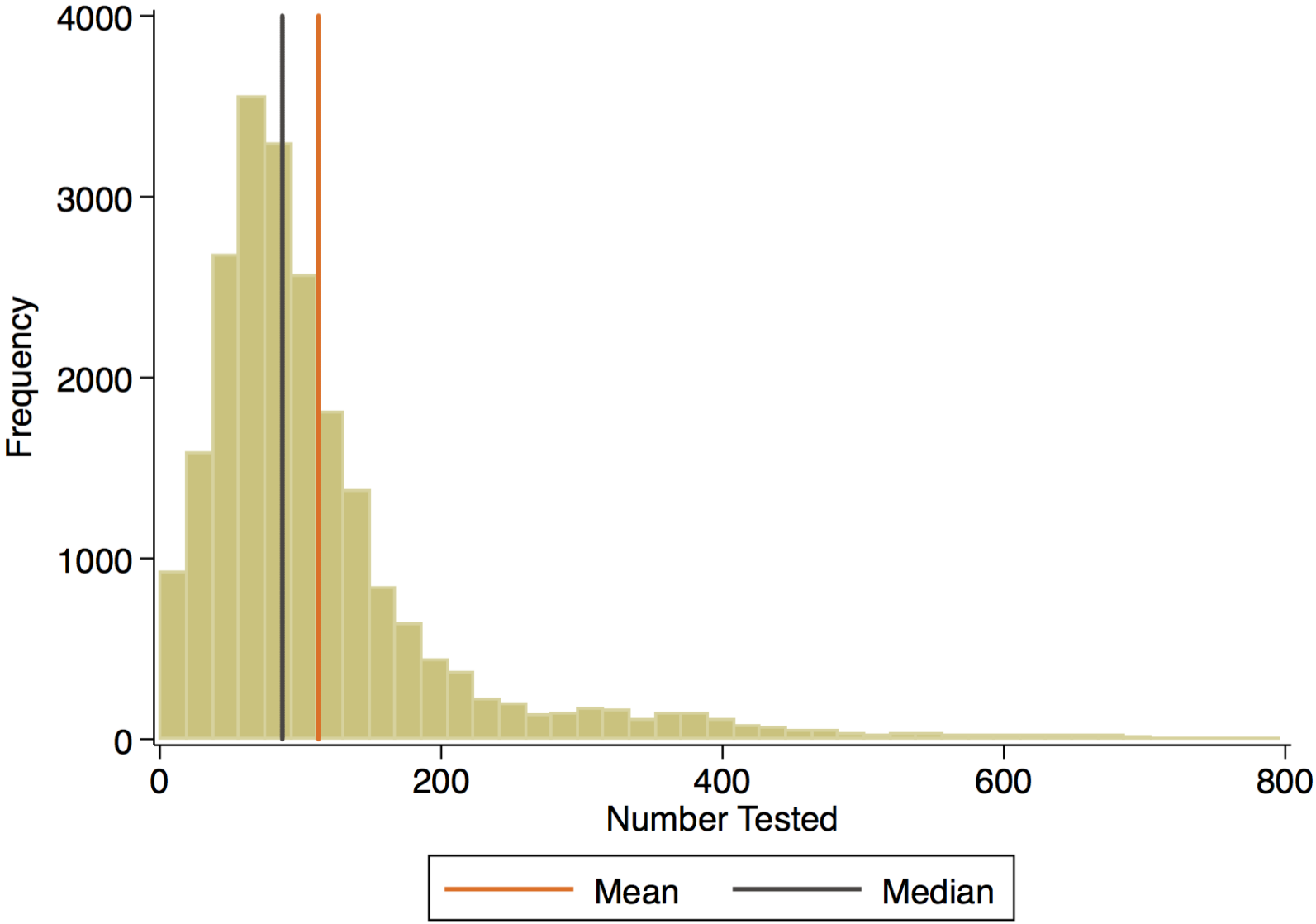
Math Score Distribution

Figure 4: Score Distribution (Density)



Number Tested Distribution

Figure 5: Number Tested Distribution

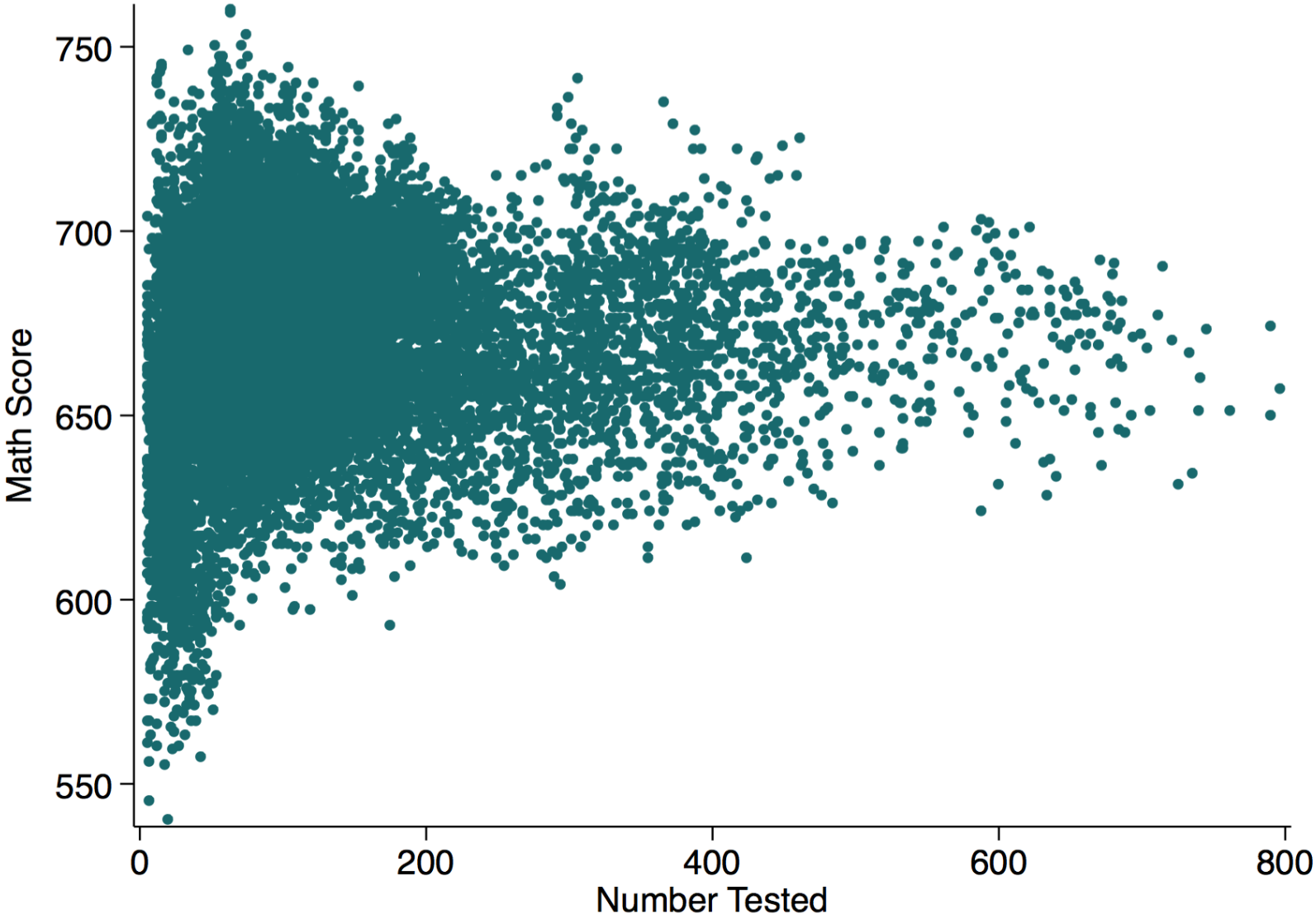


Plotting Two Variables

- To show the relationship between two variables, use a scatter plot
  - x-axis: independent/explanatory variable, y-axis: dependent/outcome variable
  - Correlation captured by the line of best fit
- Scatter plot with lots of data is not very informative
  - Common strategy: binned scatter plots
  - Group *x* variable into bins/ranges (just like histogram)
  - Find mean of *y* variable within each bin (i.e. conditional mean)

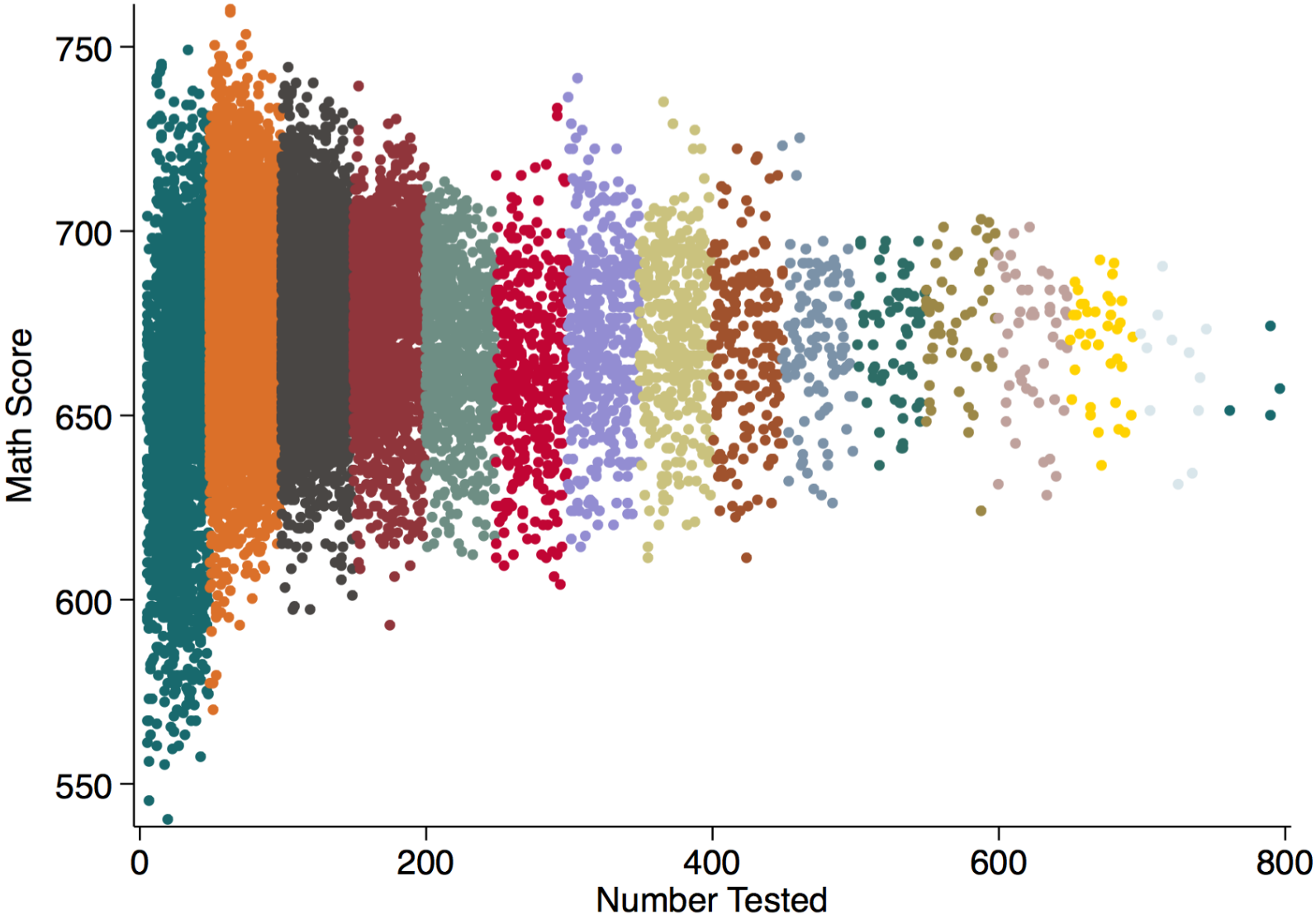
Scatter Plot

Figure 6: Score vs Number Tested



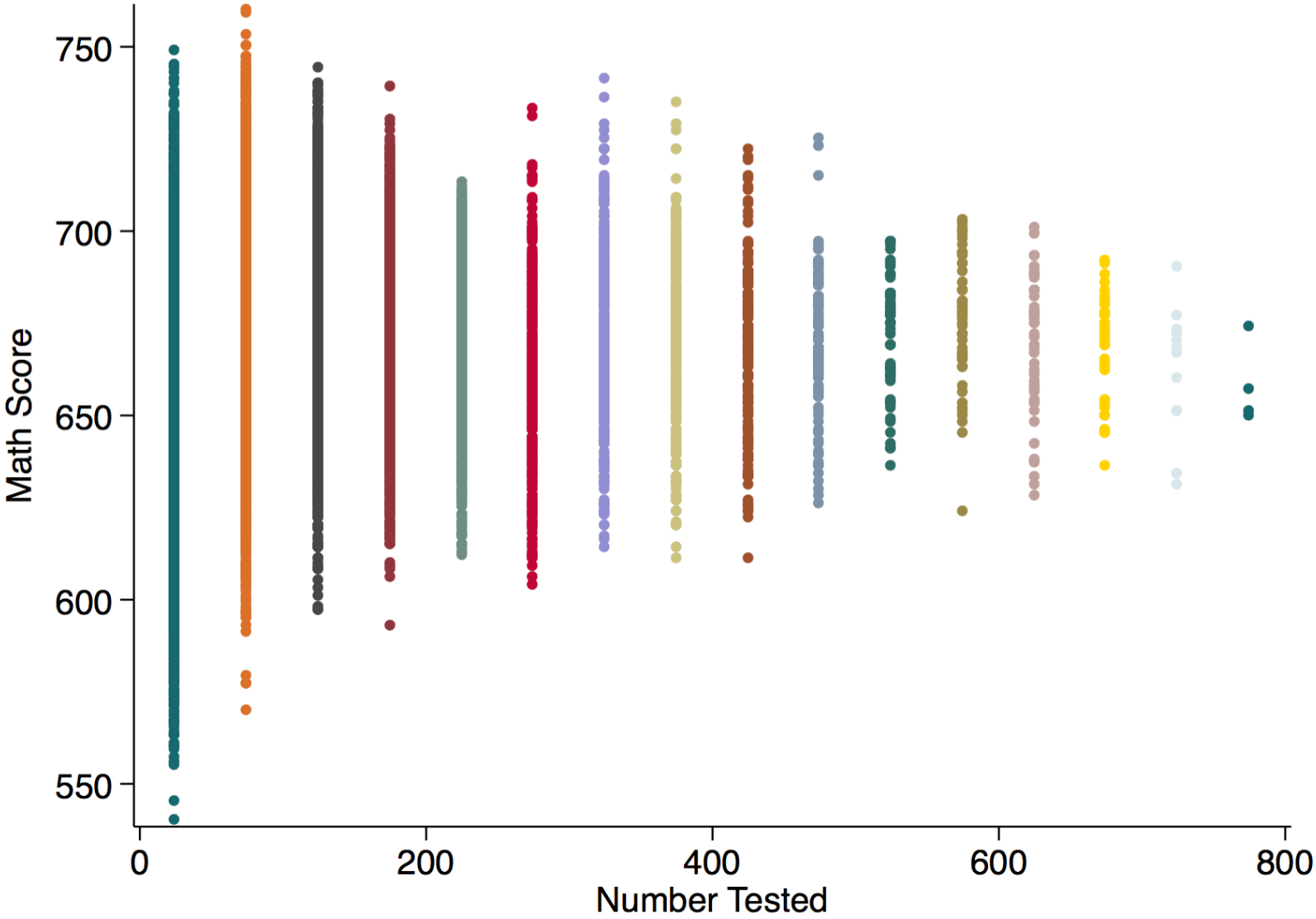
Scatter Plot

Figure 7: Score vs Number Tested (Grouped)



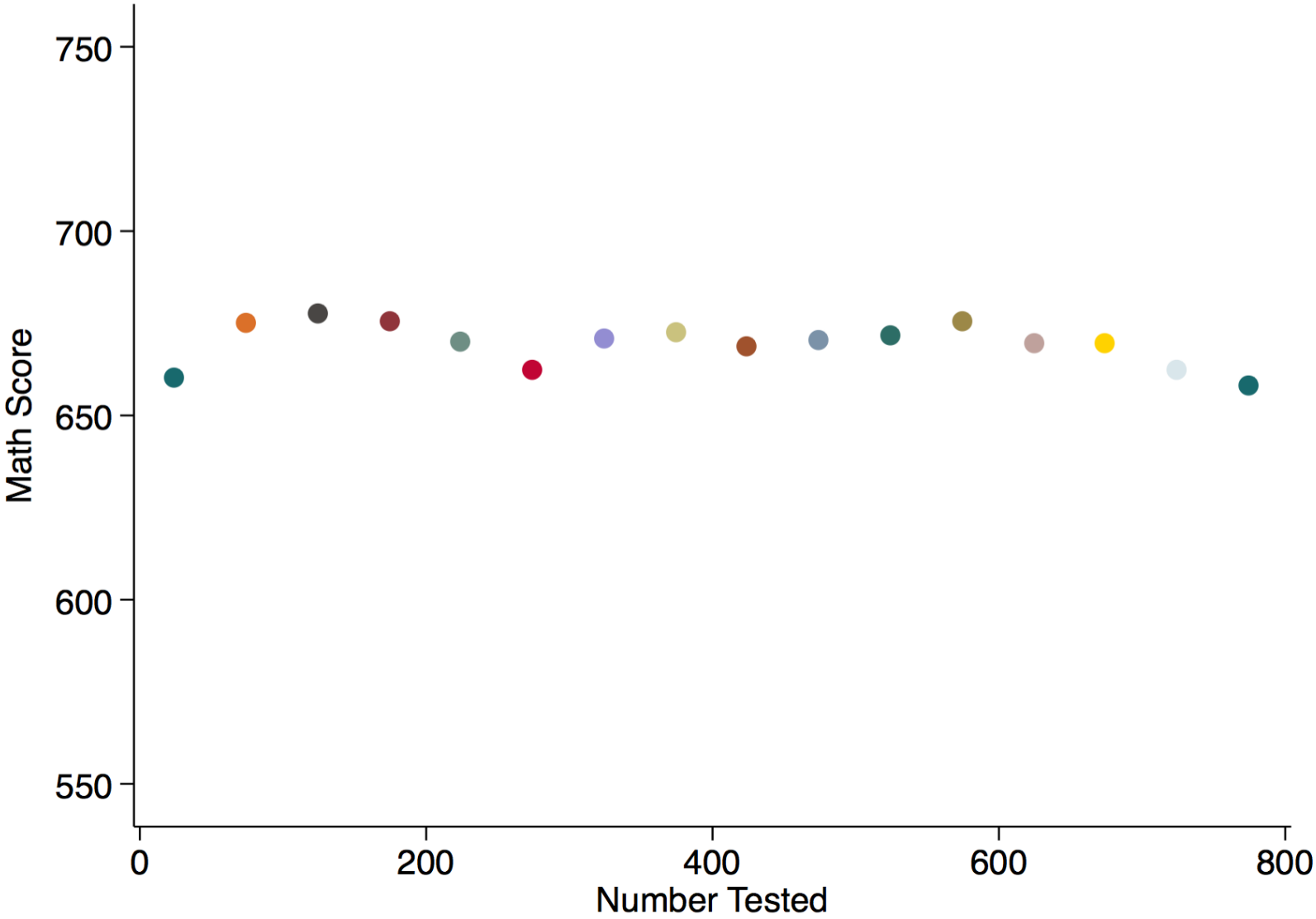
Scatter Plot

Figure 8: Score vs Number Tested (Grouped)



Scatter Plot

Figure 9: Score vs Number Tested (Binned)



Hypothesis Testing

Population vs Sample

- If we have a dataset, it is pretty easy to calculate the mean $\bar{X}$ for a variable
  - But think of our dataset as just a sample from the "true" population dataset
  - So $\bar{X}$ is only the mean for the sample. It might not be the same as the "true" mean (call this μ)
- Key issue: we do not observe the true mean μ
  - Can we use our data to say something about μ?

Hypothesis Testing

- Hypothesis testing works like this:
  - Choose some value for μ (we call this the *null hypothesis*)
  - Ask yourself: what's the chance that I observed the *sample* mean $\bar{X}$, given that the *true* mean was μ?
- This is like a game of chance. You never know if you are 100% correct
  - How confident in your answer do you have to be to bet on it?
  - Get comfortable with thinking in a probabilistic way

Example: Polling

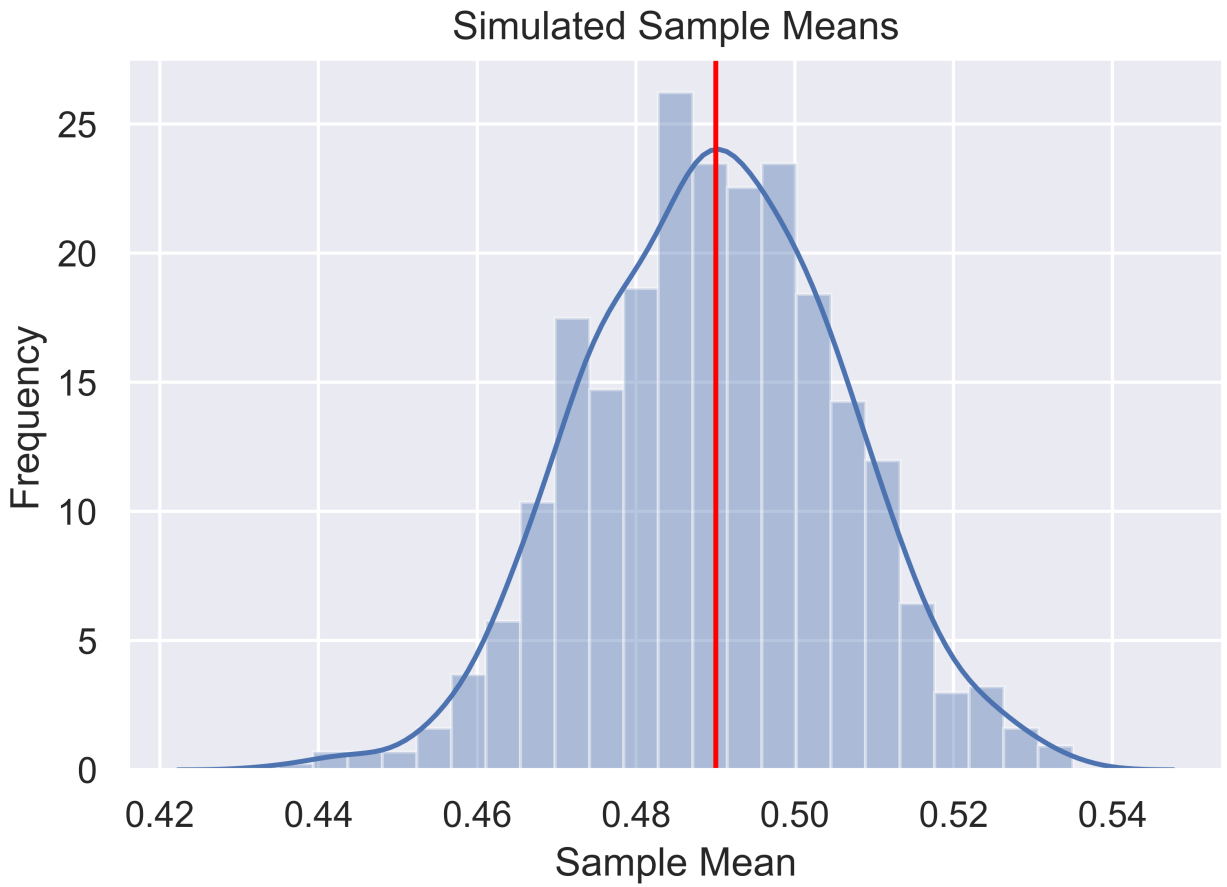
- One example where we often see hypothesis testing is political polling
- We want to know what percentage of people will vote for Trump (assume everyone's voting choice is locked in already)
  - There is some true number (μ), we just can't observe it until the election. But we want the answer now
  - Polling companies will take a sample of people and try to predict what μ will be based on their sample mean ($\bar{X}$)
  - Suppose we *think* that $\mu = 49\%$. But the poll finds that $\bar{X} = 51\%$
  - Did they just happen to pick slightly more pro-Trump people? Or was our guess of μ wrong to begin with?

Example: Polling

- If true value was 49%, then we should expect some random sampling variation
  - $\bar{X} = 51\%$ could be a plausible number for a sample mean
- But suppose we found that $\bar{X} = 30\%$!
  - That is so far away from our initial belief!
  - Assuming no weird sampling by the company, we can rule out sampling variation
  - Leaves us with one possibility: it's very unlikely that μ was indeed 49%

Example: Polling Simulation

- Simulated data: 100,000 observations, with true mean of 49%.
  - Sampled 1000 obs (1%), 1000 times. Took sample mean each time. Distribution of those means:



Testing Intuition

- The key intuition for testing works as follows: if we observe a sample mean that's far from the null, then there are two possible reasons:
  1. The null is true, and we just got a very unlikely random draw
  2. The null was not true

- Hypothesis testing looks at option 1. We *assume* that the null is true and show that the draw is very unlikely
  - If the draw is unlikely enough, we think that option 2 is a better option ("reject the null")

- Note: we can never know whether null is actually true. The hypothesis test has to assume it is true
  - This is why you will see "fail to reject the null" as opposed to "accept the null"

Jury Analogy

- Hypothesis testing is like being on a jury
- The null is innocence (“innocent until proven guilty”)
- The jury chooses between guilty or not guilty
  - Guilty = reject the null of innocence
  - Not guilty = fail to reject the null of innocence
- Choose guilty if beyond a reasonable doubt = evidence is unlikely if they were in fact innocent
- Notice that juries do not choose between guilty and innocent

Testing Procedure

- "Unlikely" is a subjective term. We want an objective way of determining when something is unlikely
- We have two equivalent approaches of doing this:
  1. ***p*-values**: is the data consistent with the null μ?
  2. **Confidence intervals**: what are all the μ's that are consistent with the data?
- In economics, we almost always have the null as $\mu = 0$

p-values

- *p*-values are probabilities that tell the chance of observing the sample mean $\bar{X}$ (or something more extreme), assuming that the null hypothesis is true (i.e. that the true mean is zero)
  - Low *p*-values mean that the null is unlikely to be true, so we can "*reject the null*"
- What is a "low" value? No hard and fast rule, but common cutoffs are 0.1, 0.05, and 0.01
  - If $p < 0.05$, the estimate is **statistically significant** at the **5% level**
  - If $p < 0.01$, the estimate is **statistically significant** at the **1% level**

Confidence Intervals

- Construct a 95% confidence interval (CI) as (approximately):

$$[\bar{X} - 2 \cdot SE_X, \bar{X} + 2 \cdot SE_X]$$

- Where SE_X is the **standard error**, $SE_X = \frac{\sigma_X}{\sqrt{N}}$
- Interpretation: the interval contains the true μ 95% of the time (when we keep re-sampling the population data)
  - Equivalently, we can reject values *outside* the interval at the 5% level
  - To get lower sig. levels, the CI needs to be wider (e.g. for 1%, multiply SE_X by 2.6)
- See this in polling all the time: *margin of error* is just a CI

Improving Precision

- Ideally, we want low *p*-values/small CIs. Two main ways this happens:
  1. Large sample size (bigger data is more generalizable)
  2. Low standard deviation (if data has little variation, makes the window of possible values smaller)

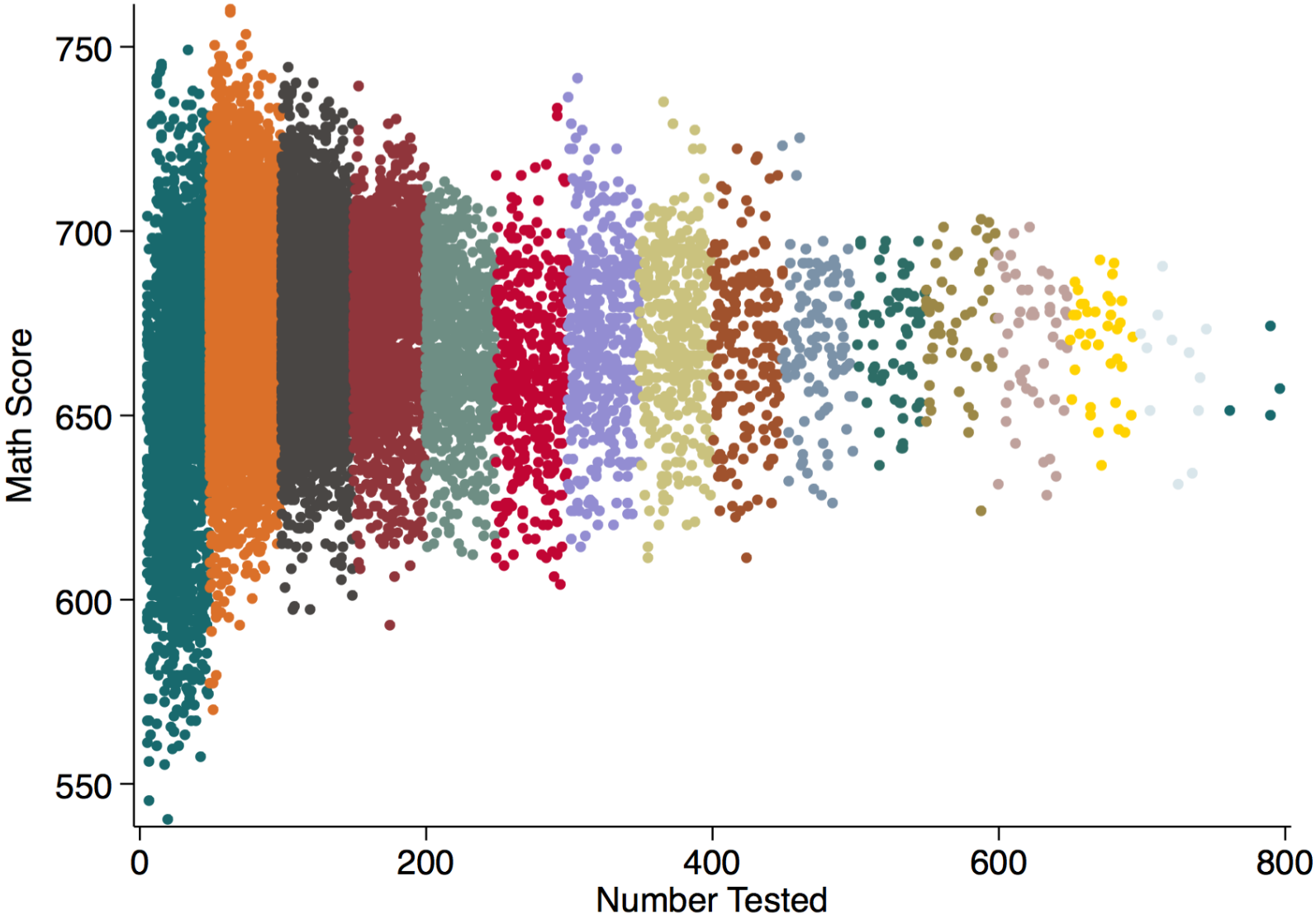
- Can't really change standard deviation, so best strategy for researchers is to collect more data

Example

- Suppose we have some data and the variable X has a mean of $\bar{X} = 2.56$
- We want to test whether the population mean is $\mu = 0$ (null)
- The analysis shows the following:
  - p-value: 0.02
  - SE = 1.10
- What can we conclude?
  - Reject $\mu = 0$ at the 5% level. Note the the 95% CI is $[0.40, 4.72]$
  - *Fail* to reject $\mu = 0$ at the 1% level. Note the the 99% CI is $[-2.78, 5.40]$

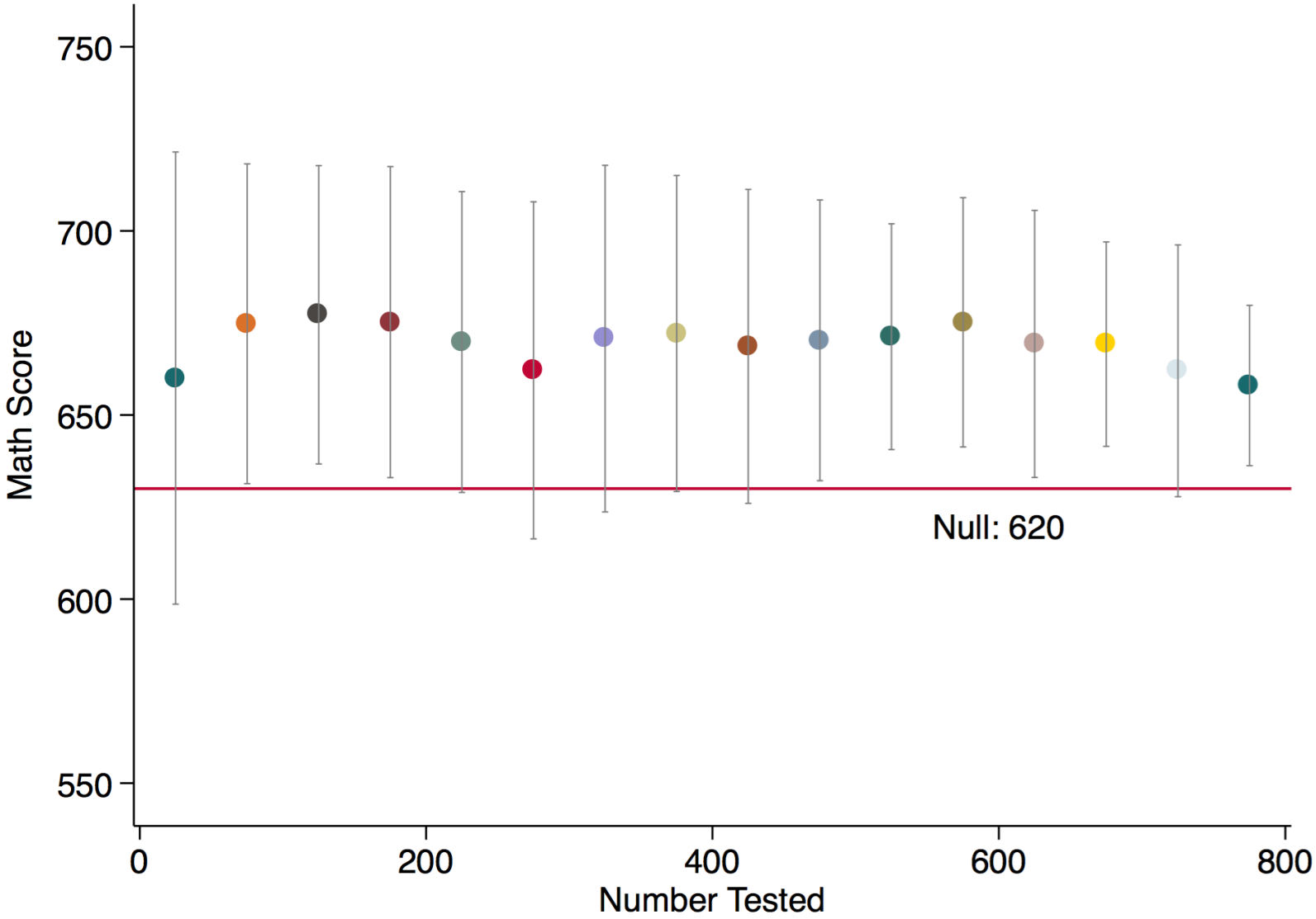
Plotting CIs

Figure 10: Score vs Number Tested



Plotting CIs

Figure 11: Score vs Number Tested (Binned)



Comparing Means

- Testing the mean of one variable isn't usually very interesting
- Often, we have two groups and we want to see if there is a difference in their means
- We do hypothesis testing on the difference in means ($\Delta\mu$)
  - Our null hypothesis is that there no difference ($\Delta\mu = 0$)
  - Same rules as before in terms of p-values/CIs

Comparing Means Example

	Grade 3	Grade 4	Difference
Math Score	683.2	678.8	4.40
SE	0.278	0.335	0.434
p-value			0.000
95% CI			[3.55, 5.25]
Number Tested	94.38	93.96	0.42
SE	0.750	0.745	1.06
p-value			0.689
95% CI			[−1.65, 2.50]

Regressions

Regression

- So far the only way to relate two variables is through covariance and correlations
- **Regressions** are a way to show the relationship between variables in a more complicated way
  - Suppose we have three variables: Y, X, and Z and we think X and Z predict the value of Y
  - We represent this relationship as an equation:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$

  - We take an educated guess at this relationship, but often a linear equation works well

Regression Equation

- Let's break down this equation: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$
- Y_i is the *dependent* variable or *outcome*
  - This is in our dataset
- X_i and Z_i are *independent* or *explanatory* variables
  - This is in our dataset
- β_0, β_1, β_2 are constant numbers called *parameters* or *coefficients*
  - This is what we want to figure out
- ε_i is called the *error term*
  - We cannot observe this!

Regression Example

	Y	*X*	*Z*
1	12.7	8	7
2	10.4	3	7
3	9.0	9	2
4	3.6	2	1
5	10.7	2	8
6	10.7	9	4

Regression Example

	Y	X	Z	$2 + 0.5X + Z$	ε_i
1	12.7	8	7	13	-0.3
2	10.4	3	7	10.5	0.1
3	9.0	9	2	8.5	0.5
4	3.6	2	1	4	-0.4
5	10.7	2	8	11	-0.3
6	10.7	9	4	10.5	0.2

In this case, $\beta_0 = 2$, $\beta_1 = 0.5$, $\beta_3 = 1$

Coefficients

- We don't just figure out the parameters by inspecting the data - this is what the regression does for us
  - Does this by finding the β's which best fit the data
- The coefficients help us interpret the relationship between the variables
  - Above example: $Y_i = 2 + 0.5X_i + Z_i + \varepsilon_i$
  - An increase in X by 1 is *associated* with an increase in Y by 0.5 (holding everything else constant)
  - An increase in Z by 1 is *associated* with an increase in Y by 1 (holding everything else constant)
- Interpretation of coefficient depends on the variable type

Example Regression

- Throughout this, let's work with a particular regression as an example:

$$Y_i = \beta_0 + \beta_1 Size_i + \varepsilon_i$$

  where Y_i is person i's test score and $Size_i$ is the number of people in person i's class

- We have a person-level dataset (each observation is a person) and in the columns we have test scores and class size
  - We are trying to predict someone's test score based on class size
  - Interpretation: an increase in class size by 1 person is associated with a change in test scores by β_1

Dummy Variables

- We can include dummy variables in the regression
  - Interpretation is always *relative* to what the group is not
- Example: let's add a *Hispanic* dummy to our regression

$$Y_i = \beta_0 + \beta_1 Size_i + \beta_2 Hispanic_i + \varepsilon_i$$

  where $Hispanic_i$ equals 1 if person i is Hispanic, and 0 if not

- Interpretation: the difference in test scores between Hispanic and non-Hispanic is β_2
  - e.g. if $\beta_2 > 0$, then Hispanics have higher test scores than non-Hispanics, on average

Categorical Variables

- To include a categorical variable, we use lots of dummy variables
  - If there are K categories, put in $K-1$ dummies
  - Interpretation is relative to the *omitted* dummy
- Example: suppose we had a race variable, with the following categories: $\{White, Black, Asian, Other\}$
- To include this in the regression, we need 3 dummies:

$$Y_i = \beta_0 + \beta_1 Size_i + \delta_1 Black_i + \delta_2 Asian_i + \delta_3 Other_i + \varepsilon_i$$

  where $Black_i$ equals 1 if person i is Black, and 0 if not (and similarly for the other two)
- Interpretation: the difference in test scores between black and white students is δ_1

Categorical Variables

- Note that omitting *White* was arbitrary
  - We could have omitted another group (just changes the interpretation)

- Example:

$$Y_i = \beta_0 + \beta_1 Size_i + \delta_1 Black_i + \delta_2 Asian_i + \delta_3 White_i + \varepsilon_i$$

- Interpretation: the difference in test scores between black and other race students is δ_1
  - Author's will pick baseline group that give the most meaningful interpretation

  - In this case, we may be most interested in the minority-white score gap

Interactions

- Dummies are very useful for **interaction variables**
- Interaction variables are created by multiplying two variables

Score	Size	Sex	Female	Race	Black	Female × Black
680	12	M	0	B	1	0
661	18	F	1	W	0	0
700	15	M	0	B	1	0
593	31	F	1	A	0	0
676	19	M	0	W	0	0
644	20	F	1	B	1	1

Interactions

- Now we can run a regression like so:

$$Y_i = \beta_0 + \beta_1 Size_i + \beta_2 Female_i + \beta_3 Black_i + \beta_4 Female_i \times Black_i + \varepsilon_i$$

- Interpretation: there are four possible groups (Female/Male and Black/Not Black). Let's ignore size for now
  - Male/Non-Black: β_0 (baseline group)
  - Male/Black: $\beta_0 + \beta_3$
  - Female/Non-Black: $\beta_0 + \beta_2$
  - Female/Black: $\beta_0 + \beta_2 + \beta_3 + \beta_4$
- Each group can now have a different average score (conditional on size)

Interactions

- Can also create interactions between dummies and continuous variables

Score	Size	Black	Size × Black
680	12	1	12
661	18	0	0
700	15	1	15
593	31	0	0
676	19	0	0
644	20	1	20

Interactions

- Now we can run this regression:

$$Y_i = \beta_0 + \beta_1 Size_i + \beta_2 Black_i + \beta_3 Size_i \times Black_i + \varepsilon_i$$

- Interpretation: the predictiveness of class size for scores is different for black students
  - Non-Black:

$$\beta_0 + \beta_1 Size_i$$

  - Black:

$$\begin{aligned} &\beta_0 + \beta_1 Size_i + \beta_2 + \beta_3 Size_i \\ &= \underbrace{(\beta_0 + \beta_2)}_{\text{Change in intercept}} + \underbrace{(\beta_1 + \beta_3)}_{\text{Change in slope}} Size_i \end{aligned}$$

Fixed Effects

- Suppose that we know that people in School *A* generally score higher than people in School *B*
  - To include school, just treat it as a categorical variable, i.e. put dummies

- For a person *i* who goes to school *s*, our regression model is now:

$$Y_{is} = \beta_0 + \beta_1 Size_i + \phi_s + \varepsilon_{is}$$

  - Note that we have a *s* subscript now
  - ϕ_s are called **fixed effects**, they tell us the average level of scores in each school (common prediction for all students who go to school *s*)

Fixed Effects

- Notation is not technically correct (but often done like this) - makes it seem like there is one variable
  - In fact, we actually need dummies for all (but one) of the school. Suppose there are 1, ..., K schools

$$\phi_s = \phi_1 School_1 + \cdots + \phi_K School_K = \sum_{k=1}^{K} \phi_k School_k$$

- Fixed effects are very important and useful (we will talk more about them later)
  - But we usually do not care about the actual values of the ϕ_k's
- Used when you have multiple observations that belong to the same group
  - Examples: state, year
  - Group cannot have one observation or be the entire sample

Choosing a Specification

- So far we've seen a few different things we could include in our regression
- How do we choose what to include?
  - Intuition/past papers - you will soon pick up on common patterns
  - Data limitations
  - Trial and error
- If we include a variable, how do we know it was the right decision?

Choosing a Specification

- With a regression, we are trying to predict the outcome variable Y using the explanatory variables
- If a variable has a coefficient equal to zero, then it doesn't help predict Y
  - This means that we shouldn't include it in the model
  - It's pretty rare for a coefficient to exactly equal to zero, so how "close" to zero should it be?

Testing

- The bar we set is whether the coefficient estimate is **statistically different from zero**
  - This is where hypothesis testing comes in!
  - Let the null hypothesis be that it is zero
- Common standard (in economics): different from zero if significant at 5 or 10% level
  - Of course, the more significant, the better!

Regression Steps

- For a researcher, running a regression often looks like this:
  1. Choose a specification, i.e. write down an equation
  2. Use statistical software to run the regressions
  3. Software estimates the *coefficients* and gives you *p-values/SEs*
  4. Decide if you're happy with it. If not, go back to step (1)

Reading Regression Output

- As consumers of research, our job is not to run the regressions but to be able to interpret a regression table
- General conventions (but every table is different):
  - Table column: output variable (Y)
  - Table rows: explanatory variables (X)
  - A cell often indicates three things:
    - The coefficient estimate (β)
    - Test information below the coefficient and in brackets (e.g. standard errors, p-values)
    - Statistical significance stars next to the coefficient estimate ($*$=10%, $**$=5%, $***$=1% level)

Interpreting Stars

*

**

***

Reading Regression Output

- You can often figure out the model specification from the table
  - Each column is a different specification
  - A blank cell indicates that the X variable was not part of that column's model
  - Some variables are in the estimate but usually not reported on table, e.g. constant parameter (β_0), fixed effects
- At the bottom, extra notes about the specification and other stats about the regression (always good to read the notes)

Output Example

Table 1: Example Regression Output

	(1) Price	(2) Price	(3) Price	(4) Mileage (mpg)
Weight (lbs.)	1.747***	3.465***	4.614***	−0.00647***
	(0.641)	(0.631)	(0.725)	(0.0007)
Mileage (mpg)	−49.51	21.85	263.2**	
	(86.16)	(74.22)	(110.8)	
Foreign		3673.1***	11240.3***	−1.655
		(684.0)	(2751.7)	(1.082)
Foreign X Mileage (mpg)			−307.2***	
			(108.5)	
Headroom (in.)				−0.219
				(0.542)
Constant	1946.1	−5853.7*	−14449.6***	41.99***
	(3597.0)	(3377.0)	(4425.7)	(2.312)
Observations	74	74	74	74

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Regression Interpretation

- So far, regressions used for “predicting” (basically more complex versions of correlations)
- Predictions just look for patterns in the data, it doesn’t tell us anything about direction or what causes the pattern
  - If you are sweating and in workout clothes, I predict that you probably went to the gym
  - But you are sweating and in workout clothes *because* you went to the gym
  - Other way makes no sense: you went to the gym *because* you were sweating and in workout clothes…?
- This example is silly, but in many cases, can be hard to tease out what is happening. We’ll dive into this in the next class!