DSAI Summer School

CHOOSING THE RIGHT STATISTICAL TEST

JUNE 28, 2017

Policy questions

- What is the effect of reducing class size by one student (or by 5 students) per class on academic achievements?
- What is the effect of reducing class length by 10 minutes on test scores?
- What is the effect of increasing educational system funding by £100 million on university ranking (other measures of the university performance)?

We must use data and rely on statistical techniques to find this out (is there any way to answer this *without* data?)

Data

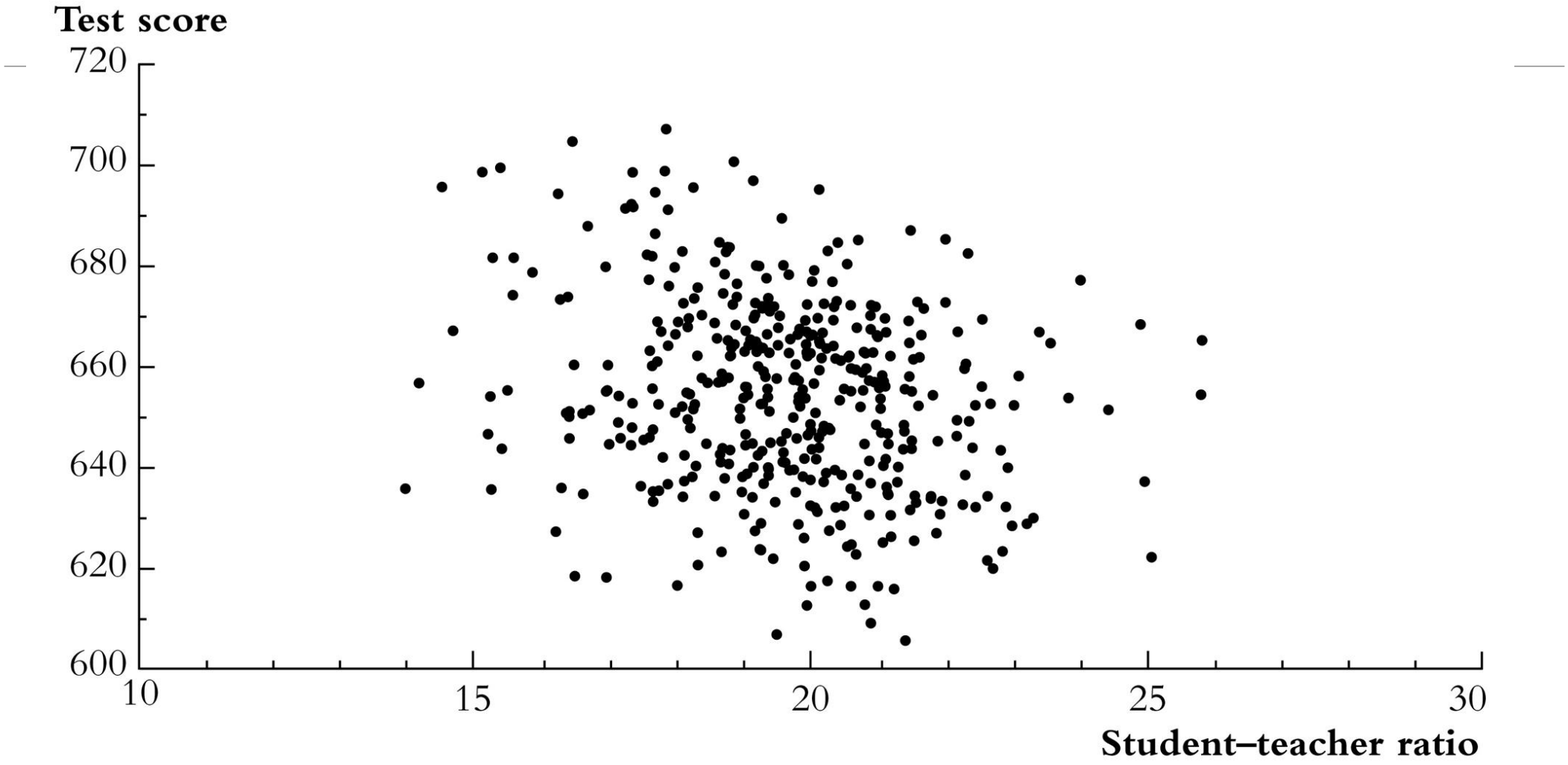
This table doesn't tell us anything about the relationship between test scores and the *STR*.

TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	665.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

Do districts with smaller classes have higher test scores?

Scatterplot of test score v. student-teacher ratio



Plan

A. Statistical inference

B. Estimation
   - Compare average test scores in districts with low STRs to those with high STRs

C. Testing
   - Test the “null” hypothesis that the mean test scores in the two types of districts are the same, against the “alternative” hypothesis that they differ

D. Confidence intervals
   - Estimate an interval for the difference in the mean test scores, high v. low STR districts

E. Linear regression

F. RCTs

Statistical inference

Probability framework

Population

The group or collection of all possible entities of interest (school districts)

We will think of populations as infinitely large (∞ is an approximation to "very big")

Random variable Y

Numerical summary of a random outcome (district average test score, district STR)

Population distribution of Y

The probabilities of different values of Y that occur in the population, for ex. $\Pr[Y = 650]$

or: The probabilities of sets of these values, $\Pr[640 < Y < 660]$ (when Y is continuous).

Moments of a population distribution

mean = expected value (expectation) of Y= $E(Y) = \mu_Y$ = long-run average value of Y over repeated realizations of Y

variance = $E(Y - \mu_Y)^2 = \sigma^2_Y$ = measure of the squared spread of the distribution

standard deviation = σ_Y

skewness $= \frac{E[(Y-\mu_Y)^3]}{\sigma^3_Y}$ = measure of asymmetry of a distribution

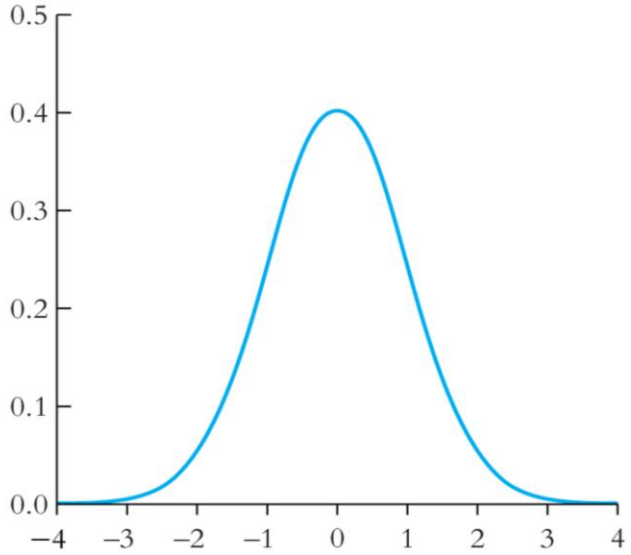
skewness = 0: distribution is symmetric

skewness > (<) 0: distribution has long right (left) tail

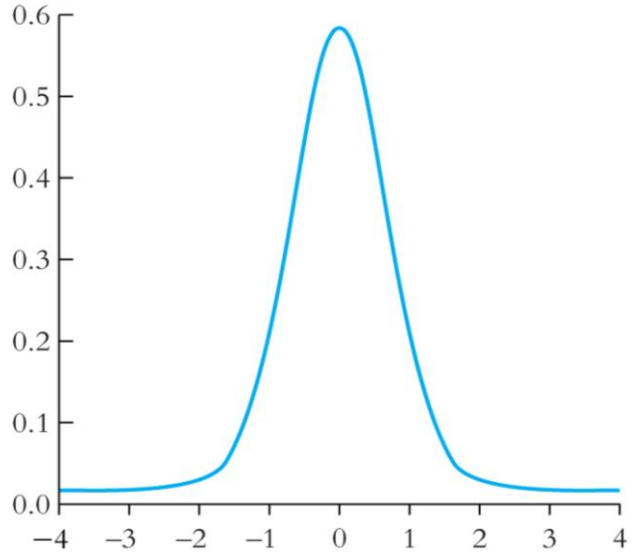
kurtosis $= \frac{E[(Y-\mu_Y)^4]}{\sigma^4_Y}$ = measure of mass in tails = measure of probability of large values

kurtosis = 3: normal distribution

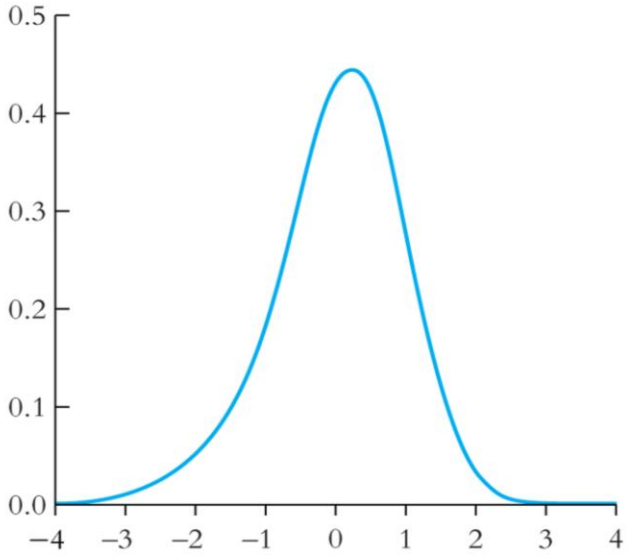
skewness > 3: heavy tails (“***leptokurtotic***”)



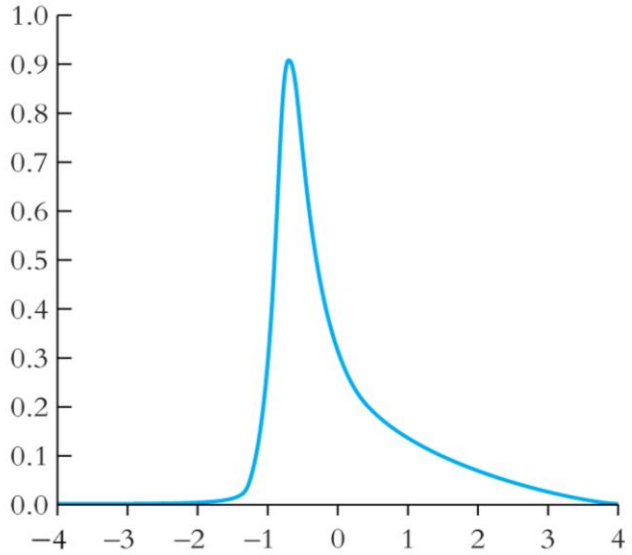
(a) Skewness = 0, kurtosis = 3



(b) Skewness = 0, kurtosis = 20



(c) Skewness = −0.1, kurtosis = 5



(d) Skewness = 0.6, kurtosis = 5

Correlation

The correlation measures the linear association between *X* and *Z*. The ***correlation coefficient*** is defined in terms of the covariance:

$$\text{corr}(X,Z) = \frac{cov(X,Z)}{\sqrt{var(X)var(Y)}} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z} = r_{XZ}$$

$$-1 \leq \text{corr}(X,Z) \leq 1$$

- corr(*X*,*Z*) = 1 mean perfect positive linear association
- corr(*X*,*Z*) = –1 means perfect negative linear association
- corr(*X*,*Z*) = 0 means no linear association

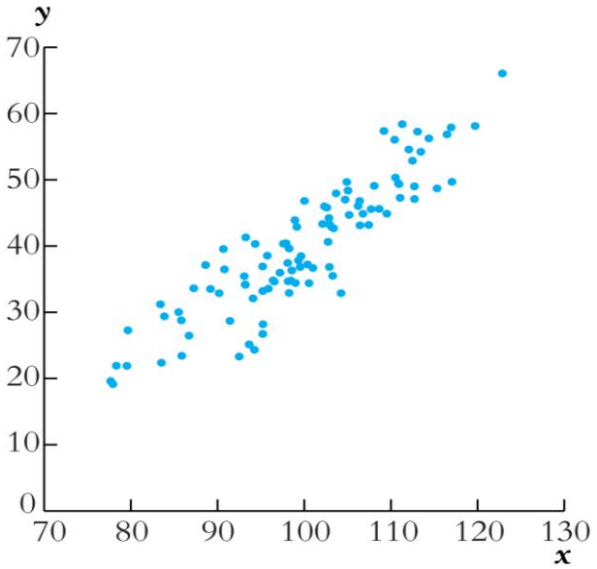
Covariance

Random variables X and Z have a ***joint distribution.*** The ***covariance*** between X and Z is

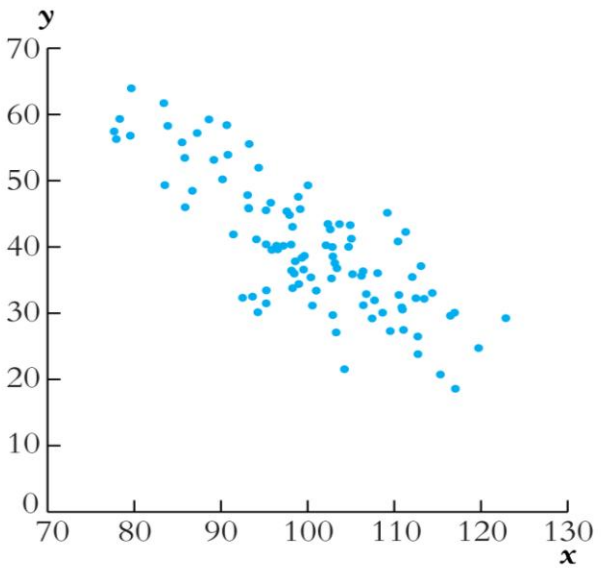
$$\text{cov}(X,Z) = E[(X - \mu_X)(Z - \mu_Z)] = \sigma_{XZ}$$

The covariance is the mean value of the product of the deviations of two variates from their respective means.

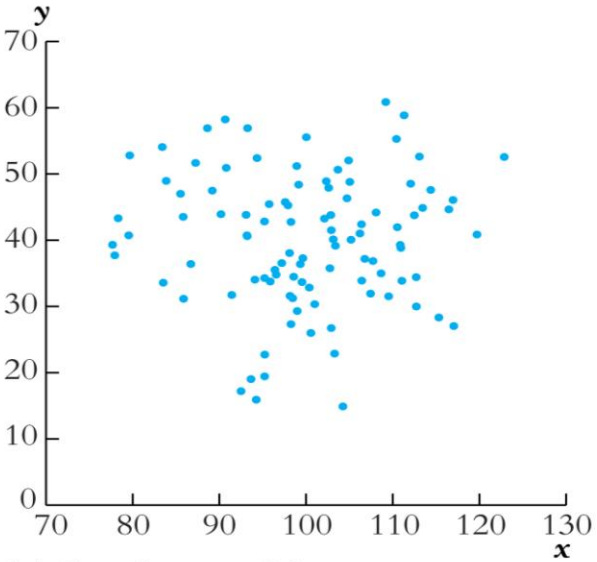
- $\text{cov}(X,Z) > 0$ means a positive relation between X and Z



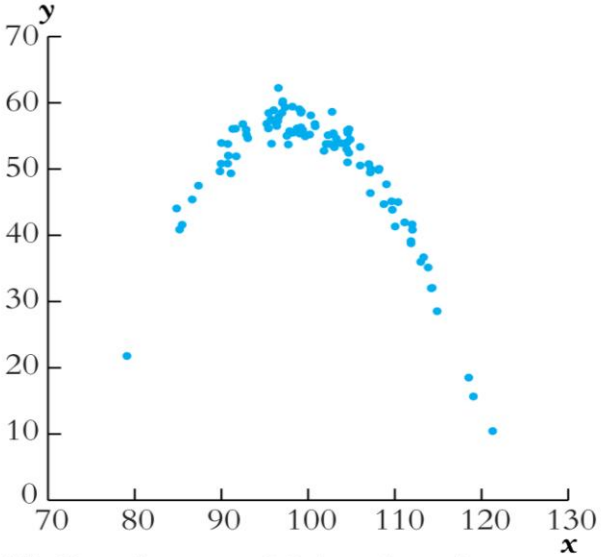
(a) Correlation = +0.9



(b) Correlation = –0.8



(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)

Distribution of a sample of data drawn randomly from a population: $Y_1,\ldots, Y_n$

We will assume simple random sampling

- Choose and individual (district, entity) at random from the population

Randomness and data

- Prior to sample selection, the value of Y is random because the individual selected is random
- Once the individual is selected and the value of Y is observed, then Y is just a number – not random
- The data set is (Y_1, Y_2,…, Y_n), where Y_i = value of Y for the i^{th} individual (district, entity) sampled

Distribution of a sample of data drawn randomly from a population: $Y_1,\ldots, Y_n$

Given that the individuals #1 and #2 are selected at random, the value of Y_1 has no information content for Y_2. Thus:

- ✓ Y_1 and Y_2 are independently distributed
- ✓ Y_1 and Y_2 come from the same distribution, that is, Y_1, Y_2 are identically distributed
- ✓ That is, under simple random sampling, Y_1 and Y_2 are independently and identically distributed (i.i.d.).
- ✓ More generally, under simple random sampling, $\{Y_i\}$, i = 1,…, n, are i.i.d.

Estimation

$\bar{Y}$ is the natural estimator of the mean. $\bar{Y}$ is a random variable, and its properties are determined by the sampling distribution of $\bar{Y}$

- The individuals in the sample are drawn at random.
- Thus the values of (Y_1,…, Y_n) are random
- Thus functions of (Y_1,…, Y_n), such as $\bar{Y}$, are random: had a different sample been drawn, they would have taken on a different value
- The distribution of $\bar{Y}$ over different possible samples of size *n* is called the sampling distribution .
- The mean and variance of $\bar{Y}$ are the mean and variance of its sampling distribution, $E(\bar{Y})$ and $\text{var}(\bar{Y})$.
- The concept of the sampling distribution underpins all of econometrics.

Things we want to know about the sampling distribution:

$$E(\bar{Y}) = \mu_Y \qquad \text{var}(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

Coming back to STR example

Initial data analysis: compare districts with “small” (STR < 20) and “large” (STR ≥ 20) class sizes:

Class Size	Average score $\overline{Y}$	Standard deviation	n
Small	657.4	19.4	238
Large	650.0	17.9	182

1. Estimation of Δ = difference between group means
2. Test the hypothesis that $\Delta = 0$
3. Construct a confidence interval for Δ

Estimation

$$\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = \frac{1}{n_{\text{small}}} \sum_{i=1}^{n_{\text{small}}} Y_i - \frac{1}{n_{\text{large}}} \sum_{i=1}^{n_{\text{large}}} Y_i$$

$$= 657.4 - 650.0$$

$$= 7.4$$

Is this a large difference in a real-world sense?

- Standard deviation across districts = 19.1
- Difference between 60P$^{\text{thP}}$ and 75P$^{\text{thP}}$ percentiles of test score distribution is 667.6 – 659.4 = 8.2
- This is a big enough difference to be important for school reform discussions, for parents, or for a school committee?

Hypothesis Testing

Hypothesis Testing

The hypothesis test is used to evaluate the results from a research study in which

1. A sample is selected from the population
2. The treatment is related to the sample
3. After treatment, the individuals in the sample are measured

If the individuals in the sample are noticeably different from the individuals in the original population, we have evidence that the treatment has an effect

However, it is also possible that the difference between the sample and the population is simply sampling error

Hypothesis Testing

The purpose of the hypothesis test is to decide between two explanations:

1. The difference between the sample and the population can be explained by sampling error (there is no evidence of a treatment effect)
2. The difference between the sample and the population is too large to be explained by sampling error (there is evidence of a treatment effect)

$$\text{Test statistic} = \frac{\text{Sample statistic} - \text{Population parameter under } H_0}{\text{Standard error of the sample statistic}}$$

The hypothesis testing problem (for the mean): make a provisional decision, based on the evidence at hand, whether a null hypothesis is true, or instead that some alternative hypothesis is true. That is, test

$$H_0: E(Y) = \propto_{Y,0} \text{ vs. } H_1: E(Y) > \propto_{Y,0} \text{ (1-sided, >)}$$
$$H_0: E(Y) = \propto_{Y,0} \text{ vs. } H_1: E(Y) < \propto_{Y,0} \text{ (1-sided, <)}$$
$$H_0: E(Y) = \propto_{Y,0} \text{ vs. } H_1: E(Y) \neq \propto_{Y,0} \text{ (2-sided)}$$

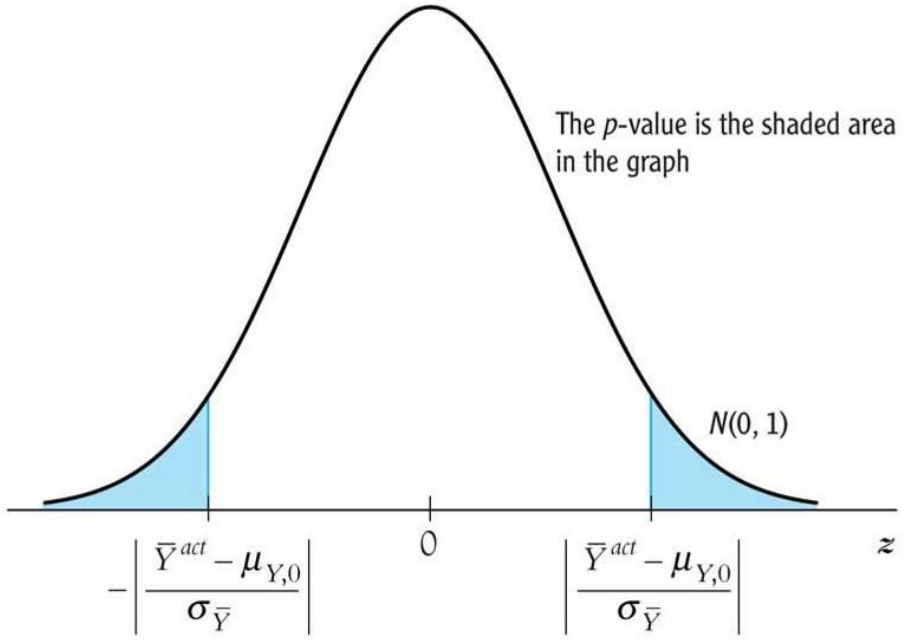
- **p-value** - probability of drawing a statistic at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true.
- The **significance level** of a test is a pre-specified probability of incorrectly rejecting the null, when the null is true.

Calculating the p-value based on :

$$\Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$$

Where $\bar{Y}^{act}$ is the value of $\bar{Y}$ actually observed (nonrandom). To compute the p-value, you need the to know the sampling distribution of $\bar{Y}$

Calculating the p-value with σ_Y known:



For large n, p-value = the probability that a N(0,1) random variable falls outside $|(\bar{Y}^{act} - \mu_{Y,0})/s_{\bar{Y}}|$ In practice, σ_Y is unknown – it must be estimated

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$ = "sample variance of *Y*"

p value= $\Pr_{H_0}[|t| > |t^{act}|]$ $\gtrsim$ probability under normal tails outside $|t^{act}|$

where $t = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}$ (the usual t-statistic)

What is the link between the *p*-value and the significance level?

The significance level is prespecified. For example, if the prespecified significance level is 5%

- you reject the null hypothesis if $|t| \geq 1.96$
- equivalently, you reject if $p \leq 0.05$.
- the *p*-value is sometimes called the ***marginal significance level***.
- often, it is better to communicate the *p*-value than simply whether a test rejects or not – the *p*-value contains more information than the "yes/no" statement about whether the test rejects.

Difference-in-means test: compute the *t*-statistic,

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$

where *SE* is the "standard error" for *s* "small" and "large" STR districts.

Class Size	Average score	Standard deviation	n
Small	657.4	19.4	238
Large	650.0	17.9	182

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657.4 - 650.0}{\sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}}} = \frac{7.4}{1.83} = 4.05$$

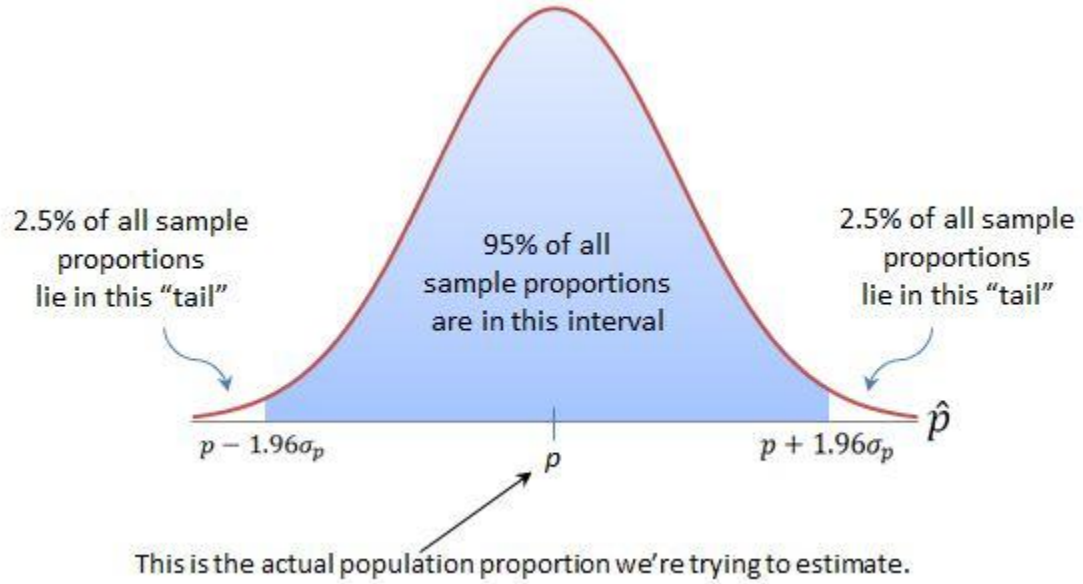
|t| > 1.96, so reject (at the 5% significance level) the null hypothesis that the two means are the same.

Confidence Intervals

A 95% ***confidence interval*** for μ_Y is an interval that contains the true value of μ_Y in 95% of **repeated samples**.

Digression: What is random here? The values of $Y_1,\ldots,Y_n$ and thus any functions of them – including the confidence interval. The confidence interval it will differ from one sample to the next. The population parameter, μ_Y, is not random, we just don't know it.

A 95% confidence interval can always be constructed as the set of values of μ_Y not rejected by a hypothesis test with a 5% significance level.



A 95% confidence interval for the difference between the means is,

$$(\bar{Y}_s - \bar{Y}_l) \pm 1.96 \times SE(\bar{Y}_s - \bar{Y}_l) = 7.4 \pm 1.96 \times 1.83 = (3.8, 11.0)$$

Two equivalent statements:

1. The 95% confidence interval for Δ doesn't include 0;
2. The hypothesis that $\Delta = 0$ is rejected at the 5% level.

Linear Regression

Let’s re-think the question again

Policy question: What is the effect of reducing class size by one student per class?

What is the right output (performance) measure?

- parent satisfaction
- student personal development
- future adult welfare
- future adult earnings
- performance on standardized tests

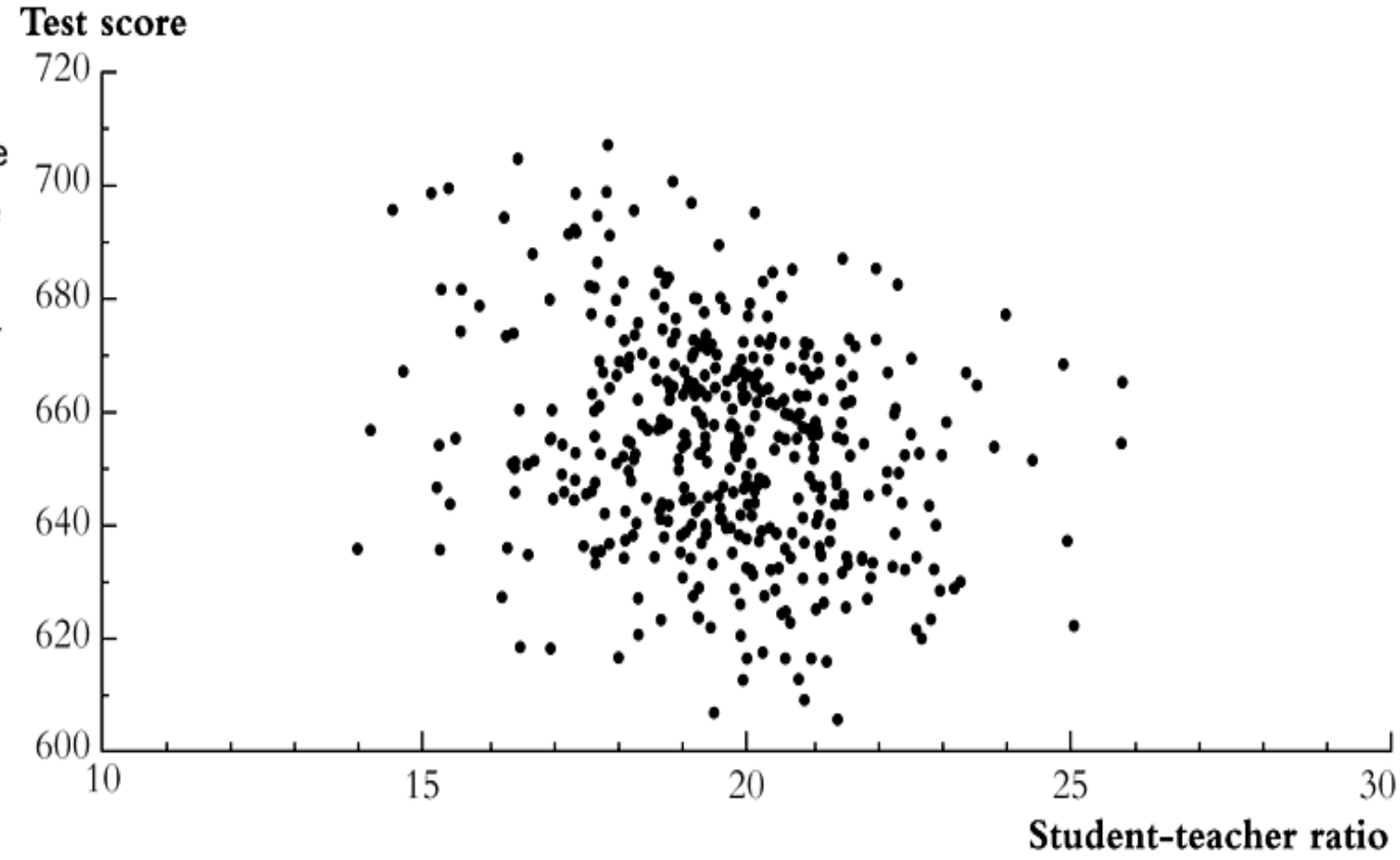
We want to know $\frac{\Delta \text{Test score}}{\Delta STR}$

This is the slope of the line relating test score and *STR*

This suggests that we want to draw a line through the *Test Score v. STR* scatterplot

FIGURE 4.2 Scatterplot of Test Score vs. Student-Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student-teacher ratio and test scores: the sample correlation is –0.23.



The population regression line:

$$\text{Test Score} = \beta_0 + \beta_1 \text{STR}$$

β_1 = slope of population regression line = $\frac{\Delta \text{Test score}}{\Delta STR}$ = change in test score for a unit change in STR

- We would like to know the population value of β_1.
- We don't know β_1, so must estimate it using data.

- We will focus on the least squares ("ordinary least squares" or "OLS") estimator of the unknown parameters β_0 and β_1, which solves,

$$\min_{b_0,b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

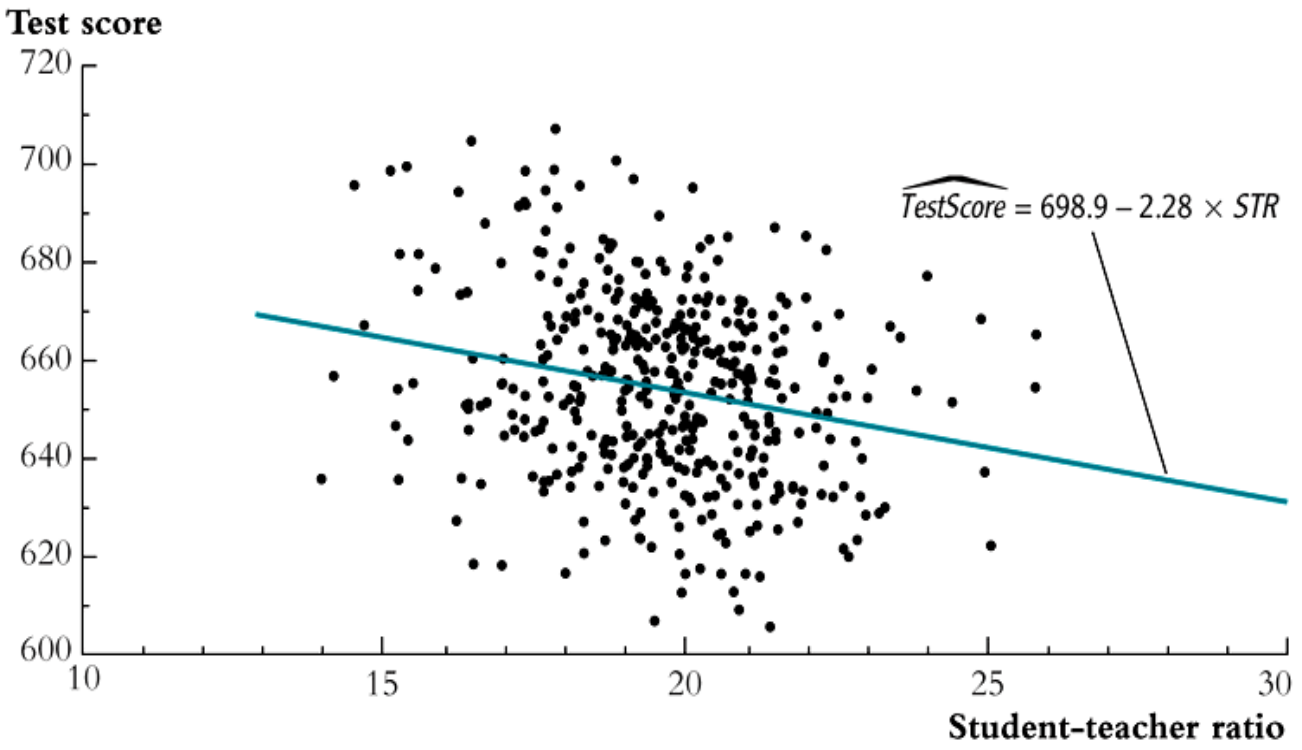
- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.
- This minimization problem can be solved using calculus.
- The result is the OLS estimators of β_0 and β_1.

Why use OLS, rather than some other estimator?

- OLS is a generalization of the sample average: if the "line" is just an intercept (no *X*), then the OLS estimator is just the sample average of $Y_1,\ldots Y_n$ ($\bar{Y}$)
- Like $\bar{Y}$, the OLS estimator has some desirable properties: under certain assumptions, it is unbiased (that is, E($\hat{\beta}_1$)= β_1), and it has a tighter sampling distribution than some other candidate estimators of β_1 (more on this later)
- Importantly, this is what everyone uses – the common "language" of linear regression.

FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. If class sizes fall by 1 student, the estimated regression predicts that test scores will increase by 2.28 points.



Estimated slope = $\hat{\beta}_1 = -2.28$

Estimated intercept = $\hat{\beta}_0 = 698.9$

Estimated regression line: $= 698.9 - 2.28 \times STR$

Interpretation of the estimated slope and intercept

$$\text{Test Score} = 698.9 - 2.28 \times STR$$

- Districts with one more student per teacher on average have test scores that are 2.28 points lower. That is,
- $$\frac{\Delta \text{Test score}}{\Delta STR} = -2.28$$

- The intercept (taken literally) means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9.
- This interpretation of the intercept makes no sense – it extrapolates the line outside the range of the data – in this application, the intercept is not itself economically meaningful.
- We can also sub in numbers. One of the districts in the data set is Antelope, CA, for which *STR* = 19.33 and *Test Score* = 657.8
  - predicted value: $\hat{Y}_{Antelope} = 698.9 - 2.28 \times 19.33 = 654.8$
  - residual: $\hat{u}_{Antelope} = 657.8 - 654.8 = 3.0$

The OLS regression line is an estimate, computed using our sample of data; a different sample would have given a different value of $\hat{\beta}_1$.

How can we:

- quantify the sampling uncertainty associated with $\hat{\beta}_1$?
- use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$?
- construct a confidence interval for β_1?

Like estimation of the mean, we proceed in four steps:

1. The probability framework for linear regression
2. Estimation
3. Hypothesis Testing
4. Confidence intervals

The Least Squares Assumptions

1. The conditional distribution of u given X has mean zero, that is,

$$E(u \mid X = x) = 0.$$

2. (X_i, Y_i), $i = 1,\ldots,n$, are i.i.d.

3. X and u have four moments, that is:

$$E(X^4) < \infty \text{ and } E(u^4) < \infty.$$

Randomised Control Trials

What is an RCT?

- Trial population is *randomly* divided into two groups, experimentals and controls
  - Experimentals get treatment
  - Controls get none
  - Average outcome in experimental group minus average outcome in control group tells us if the treatment works, and by how much on average

- An RCT estimates an *average* treatment effect
  - In general, each person (unit) will have a different treatment effect
  - We cannot observe these for each individual
  - But RCT gives us the average for the group, which is a lot!

- Minimal assumptions, absence of bias, establishing causality are big advantages
  - But is this really the only “rigorous” evaluation?

Example

- Conditional cash transfers (CCTs), some villages got CCTs, some did not
  - Better average outcomes for treatment villages
  - Random selection means it must have been the CCT, not something else
- What do we learn?
  - Will it work in India? External validity.
  - Will it work for a *specific* village in Mexico?
  - Why did it work? If we knew, we could answer two questions?
  - Controls knew they were going to get CCTs later? Does that matter?
- Dams: not possible to do randomized dam construction!!
  - So RCTs cannot be done in all cases
  - Some have argued that policies should not be implemented in these cases
  - Do many things routinely for which there have been no RCTs!

Small RCTs

- Are often not large enough to be reliable
  - Expensive to do, so this is not a matter that is easily fixed
  - In a small trial, a few outliers can drive the results
  - Example might be microfinance, where one or two women might be able to do really well, and the rest not at all
  - Get lots of weird and counterintuitive results
  - No idea if they are real, or method is just broken
  - Doubt one can learn *anything* from a trial of 10 experimental villages and 10 control villages in CCT experiment
- Experiment is often conducted on a convenience sample
  - Not easy to get cooperation from all relevant units: e.g. in looking at CCT, those opposed to the idea might be less willing to cooperate
  - Results are correct *only* for the convenience population
  - Not for population that would be affected by the policy

Large scale RCTs

- Use *all* of the units in a country
  - PDS/CCT experiment for *all* of rural India
- Comparable to large social experiments in the US in the 70s
  - NJ income tax experiment, SIME/DIME
  - Rand Health experiment
- Rand experiment is an important part of the debate today, others not
- Ex post data mining
  - Null result is *never* acceptable to the sponsors
  - Enormous pressure on investigators to find *something*
  - Usually by subgroup analysis, or looking for other outcomes
- MTO has now examined *thousands* of outcomes
  - Some of the statistically significant ones are spurious
  - And we are back to the small sample problem again
- Large experiments not decisive either

Dynamic effects

- Many policies take time to work out
  - Lots of things work as intended in the short-run, fail later
  - People learn to “work the system”
  - Procurement and supply effects in food policy
- What would an RCT show?
  - It works! Expensive and unethical to continue the experiment
  - We get the wrong answer, or only part of the answer
  - Issue in medicine too

What does an RCT *not* tell us?

- Informative about the mean, not of *any* other characteristic of the distribution of treatment effects, e.g. the median, or the fraction of people who benefit, or lose
  - Policymakers are often interested in these
- Does yield the full distribution of outcomes for both treatments and controls
  - For some purposes, this might be enough
  - If one distribution first-order stochastically dominates the other
- Another aspirin example: low dose regime
  - RCTs show a net reduction in mortality
  - But it kills some and saves some
  - Public health perspective says do it
  - Individual doctor or patient perspective is much less clear

Taking Evidence to Policy

Using a perfect evaluation

- Suppose we have a result, e.g.
  - On average, CCTs make people happier than Public Distribution System (PDS)
  - On average, dams increase poverty
  - On average, reducing open defecation improves child health and reduces mortality
- Suppose also that these were all done perfectly, so there is no dispute about the conclusions
  - Which, of course, never happens!
- What use can we make of those results in policy?
  - Should the Planning Commission ban new dams?
  - Should MRD encourage better sanitation?
  - Should we replace PDS by CCTs?
- That dams don't work on average tells us little about any individual dam
  - It is an individual dam that comes up for approval, not all dams!
  - We needs to know more, *why* dams cause poverty, *under what circumstances*, none of which comes from an RCT

What should a village do?

- Or any local authority that decides
  - Given an RCT about CCT v PDS
- Again, the average is useful but not decisive
  - Will it have the same effect for us?
  - We are not the average village
  - Again, we need to know *why* it works, not *whether* it works
- Neighboring village tried and is happy with the outcome
  - Perhaps this is just an anecdote (“your uncle likes his new TV”)
  - But for the village, the average outcome is an anecdote too
  - Perhaps the authorities should visit their neighbors and see what is going on, see if it would work for them
- Average is more useful for a public health policy that will be applied to the whole country
  - Sanitation?

Causality & helping factors

- Do not RCTs reveal causality?
  - It was the treatment that did it! Not something else
- Is this not particularly helpful in policy? Yes and no.
- Causality, by itself, is not always useful
  - The house burned down because the TV was left on
  - Causal, but not general: TVs do not usually burn down houses
  - RCT would show this causal effect
  - But TVs need “helping factors” like bad wiring, or inflammable material left nearby
- We have to think about what are the helping factors, how they work, and whether they will work for us
  - Will a CCT work in a particular village, or during food price inflation, or in a competent v a corrupt state
  - Does it need banks, or clinics to make it work?
  - Does it matter who gets it? Men and women: gender issues in India v Latin America
- Replication of an RCT is not useful, because get different results in different contexts with or without helping factors
  - Causality is “local”

Other Approaches

- Dif-in-Dif
- RDD

Any Other Questions?

- Talk during the coffee break
- Email: klymakm@tcd.ie