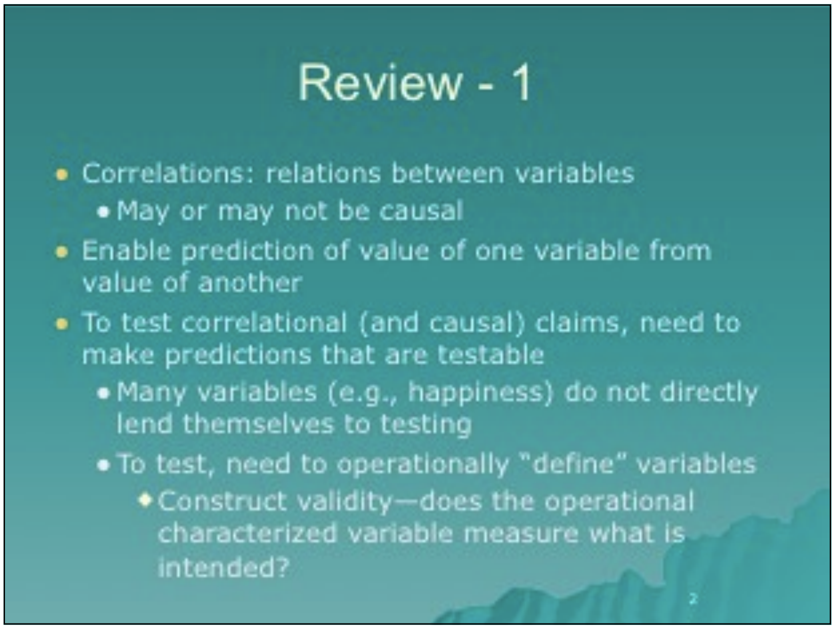
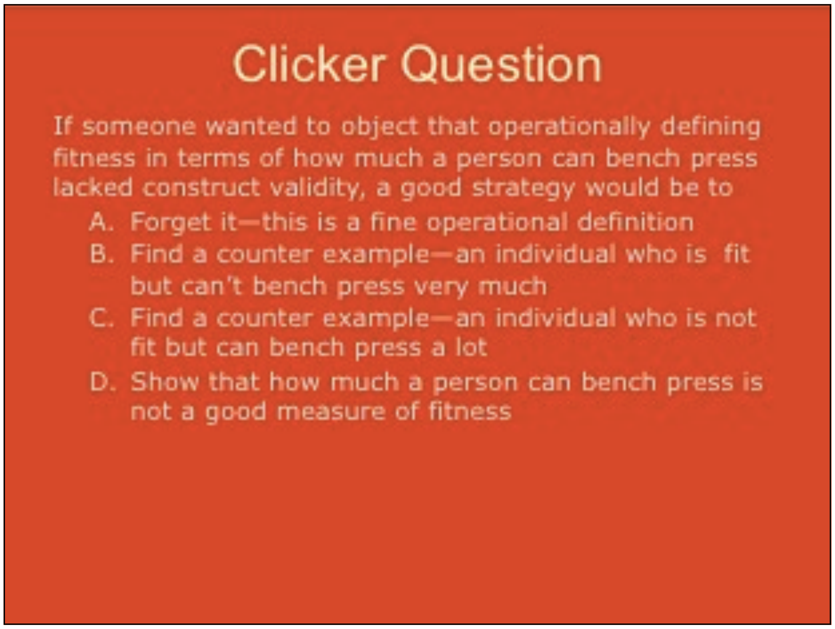
Predicting from Correlations



Review - 1

- Correlations: relations between variables
  - May or may not be causal
- Enable prediction of value of one variable from value of another
- To test correlational (and causal) claims, need to make predictions that are testable
  - Many variables (e.g., happiness) do not directly lend themselves to testing
  - To test, need to operationally "define" variables
    - Construct validity—does the operational characterized variable measure what is intended?



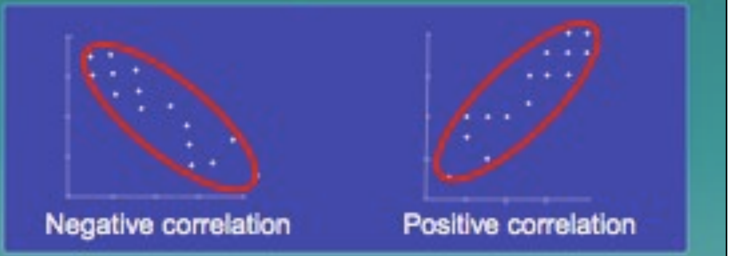
Clicker Question

If someone wanted to object that operationally defining fitness in terms of how much a person can bench press lacked construct validity, a good strategy would be to

A. Forget it—this is a fine operational definition
B. Find a counter example—an individual who is fit but can't bench press very much
C. Find a counter example—an individual who is not fit but can bench press a lot
D. Show that how much a person can bench press is not a good measure of fitness

Review - 2

- Use scatterplots to diagram correlations

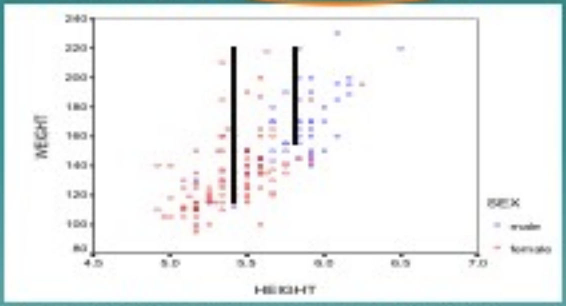


Review - 3

- **Measure strength of correlation in terms of Pearson co-efficient (r)**
- **-1.0____________________0____________________1.0**
- **Perfect negative No Correlation Perfect Positive**
  - With a high absolute value of r, you can predict accurately the value on the second variable from the value on the first
  - With a value of r near 0, you cannot make an accurate prediction of the second variable from the value of the first

Correlation Coefficients

- Height and weight are positively correlated
  - In this graph, Pearson r=.67



Contains two subgroups: men and women

May exhibit different correlations

- For females (red) only, r =.47
- For males (blue) only, r = .68

Clicker Question

From the information that the Pearson (r) for correlation between height and weight is .47 for females and .68 for males, one can conclude that

A. It is possible to make a more accurate prediction of weight from height for males

B. It is possible to make a more accurate prediction of weight from height for females

C. Males are taller and weigh more than females

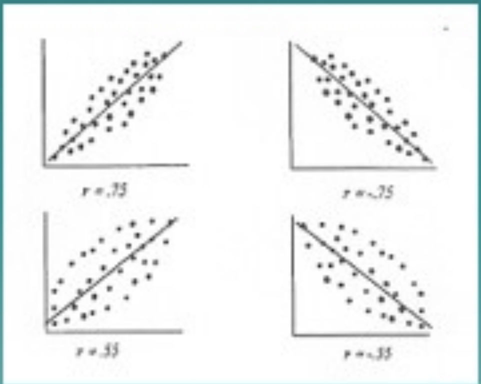
D. Since both correlations are close to .5, there is nothing that can be predicted

How much does the correlation account for?

- Correlations are typically not perfect (r=1 or r=-1)
  - Evaluate the correlation in terms of how much of the *variance in one variable is accounted for by the variance in another* [variance=Σ (X-mean)^2/N]
- Amount of variance accounted for (on the variable whose value is being predicted) equals:
  - Variance explained/total variance
- This turns out to be the square of the Pearson coefficient: r^2
- So:
  - if r=.80, then we can say that 64% of the variance is explained.
  - If r=.30, then we can say that 9% of the variance is explained.

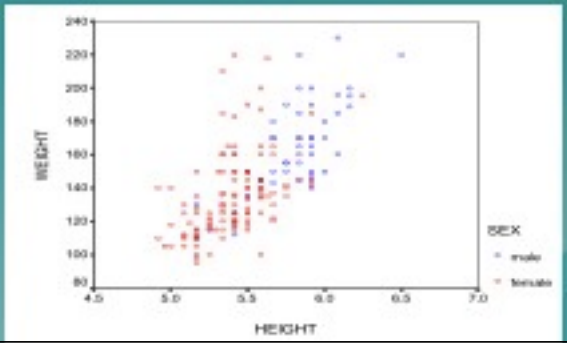
Variance Accounted for

- r^2 = .56
- r^2 = .30



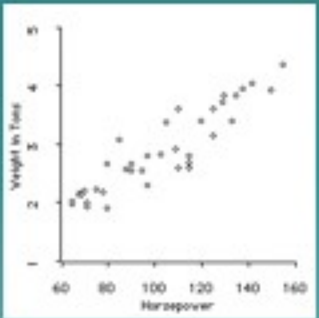
Variance accounted for - 2

- Height only partially accounts for weight
  - For females, r =.47, so r^2=.22
  - For males, r = .68, so r^2=.46



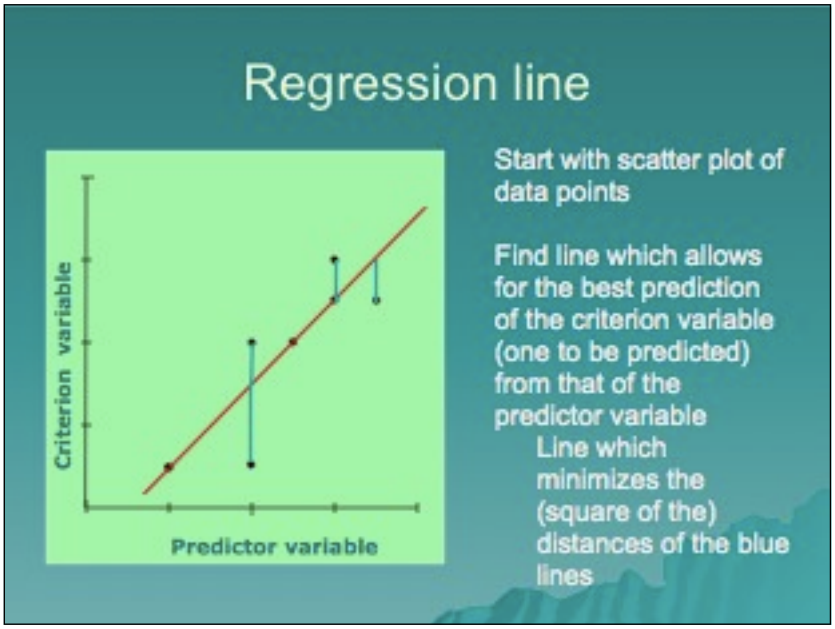
Variance accounted for - 3

- Correlating automobile horsepower and weight
  - r = .92
  - r^2 = .81
- Horsepower accounts for 81% of the variance in car weight
  - Given only the horsepower of a car, you can make a quite reliable estimate of the car's weight



Prediction

- A major reason to be interested in correlation
  - If two variables are correlated, we can use the value of an item on one variable to predict the value on another
    - Employment prediction: *future job performance* based on *years of experience*
    - Actuarial prediction: *how long one will live* based on *how often one skydives*
    - Risk assessment: prediction of *how much risk* an activity poses in terms of its values on *other variables*
- Prediction employs the ***regression line***

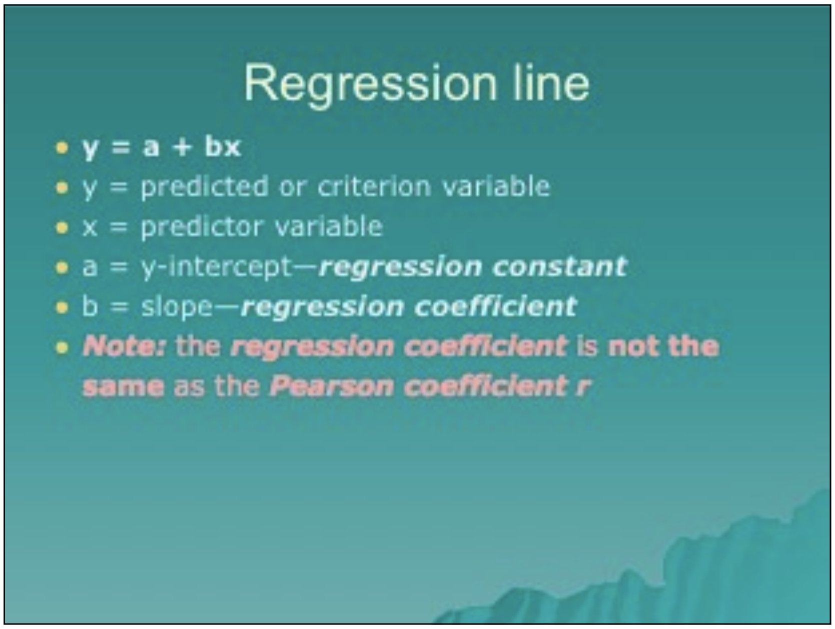


Regression line

Start with scatter plot of data points

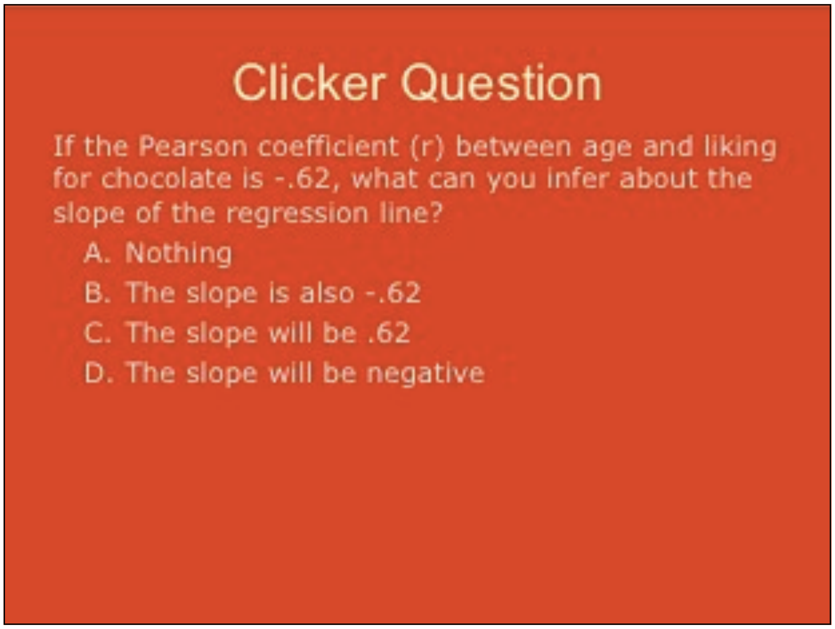
Find line which allows for the best prediction of the criterion variable (one to be predicted) from that of the predictor variable

Line which minimizes the (square of the) distances of the blue lines



Regression line

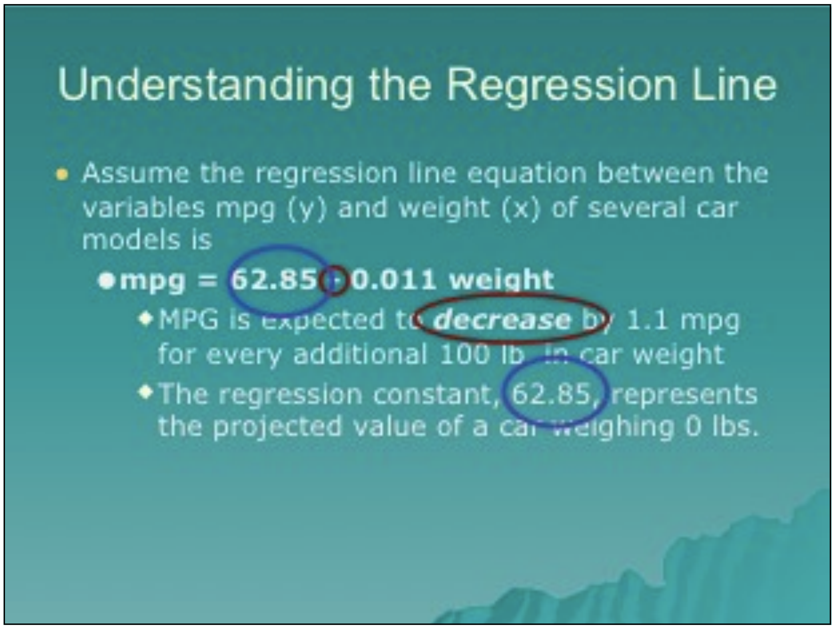
- **y = a + bx**
- y = predicted or criterion variable
- x = predictor variable
- a = y-intercept—***regression constant***
- b = slope—***regression coefficient***
- ***Note:*** the ***regression coefficient*** is **not the same** as the ***Pearson coefficient r***



Clicker Question

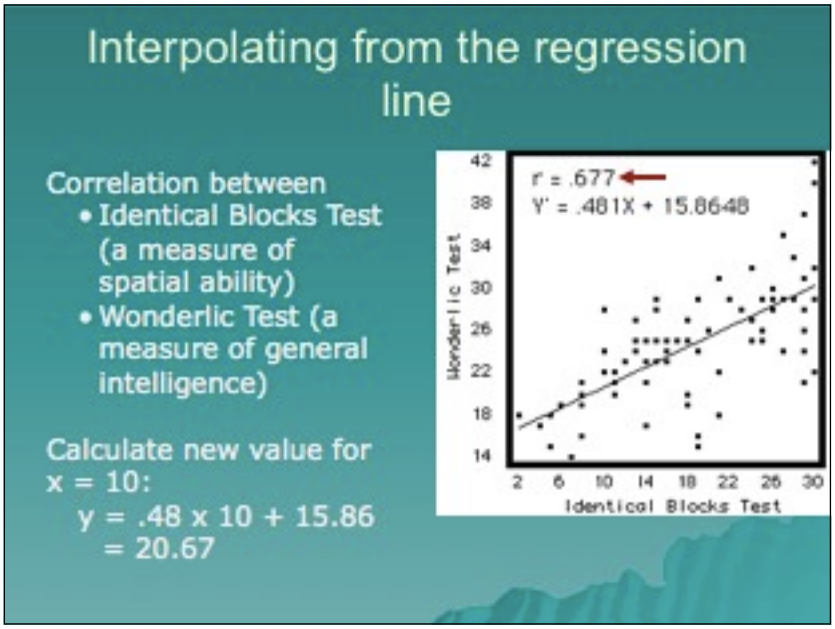
If the Pearson coefficient (r) between age and liking for chocolate is -.62, what can you infer about the slope of the regression line?

A. Nothing
B. The slope is also -.62
C. The slope will be .62
D. The slope will be negative



Understanding the Regression Line

- Assume the regression line equation between the variables mpg (y) and weight (x) of several car models is
  - mpg = 62.85 - 0.011 weight
    - MPG is expected to ***decrease*** by 1.1 mpg for every additional 100 lb. in car weight
    - The regression constant, 62.85, represents the projected value of a car weighing 0 lbs.



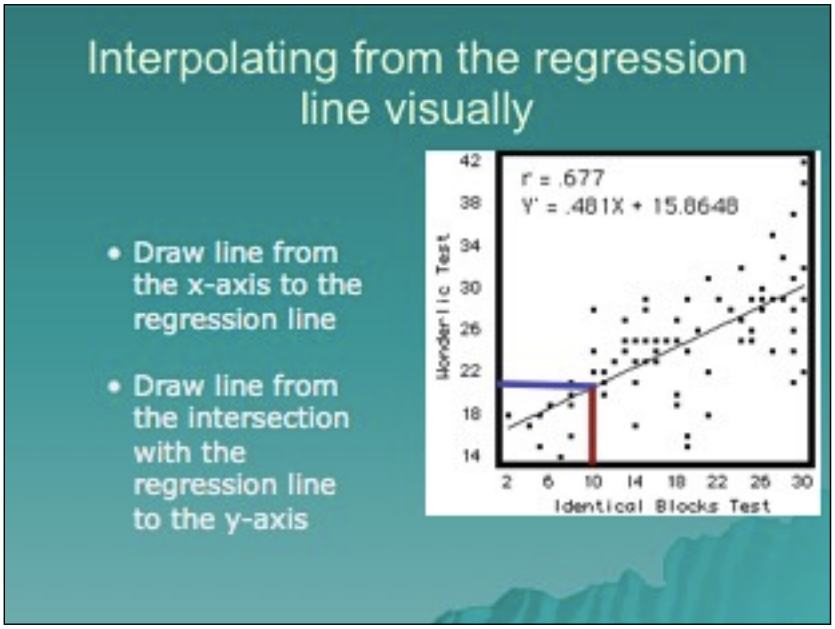
Interpolating from the regression line

Correlation between

- Identical Blocks Test (a measure of spatial ability)
- Wonderlic Test (a measure of general intelligence)

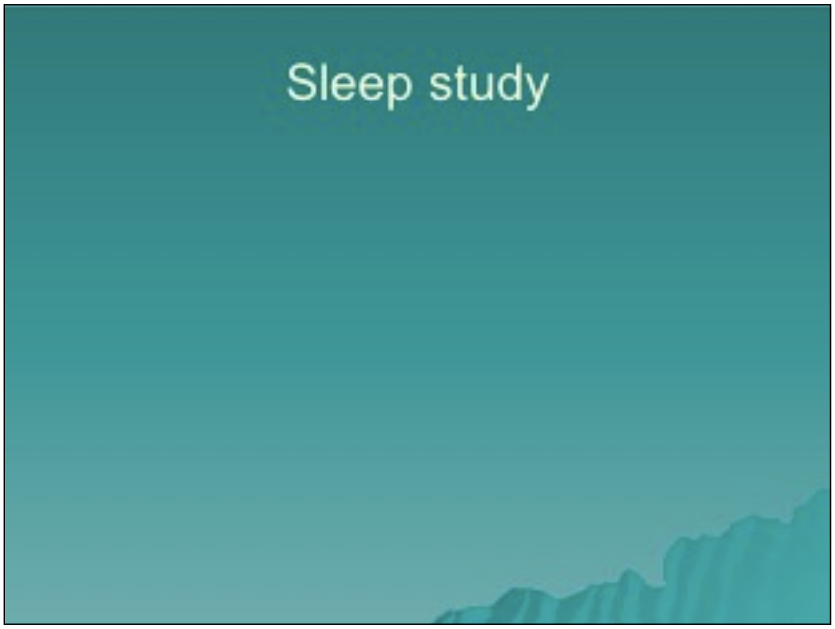
Calculate new value for x = 10:

y = .48 x 10 + 15.86
= 20.67

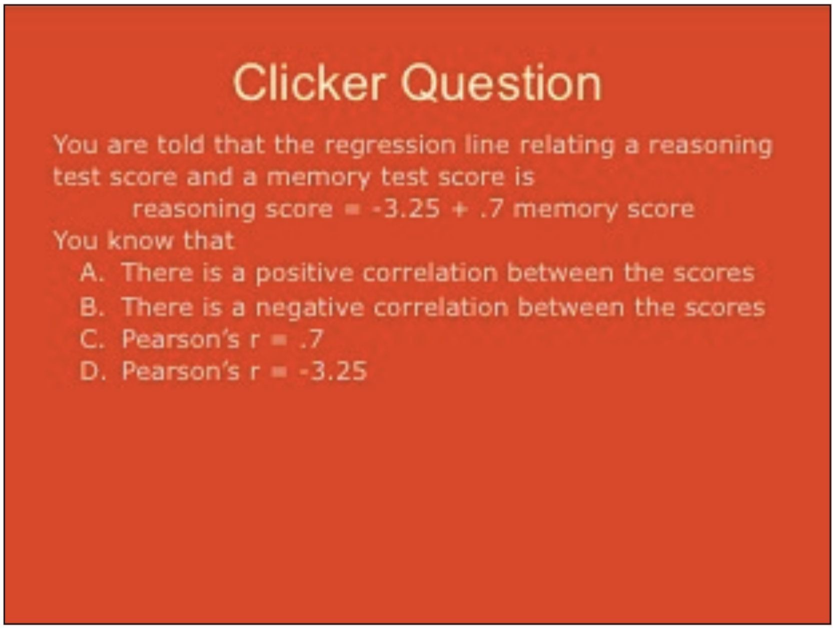


Interpolating from the regression line visually

- Draw line from the x-axis to the regression line
- Draw line from the intersection with the regression line to the y-axis



Sleep study



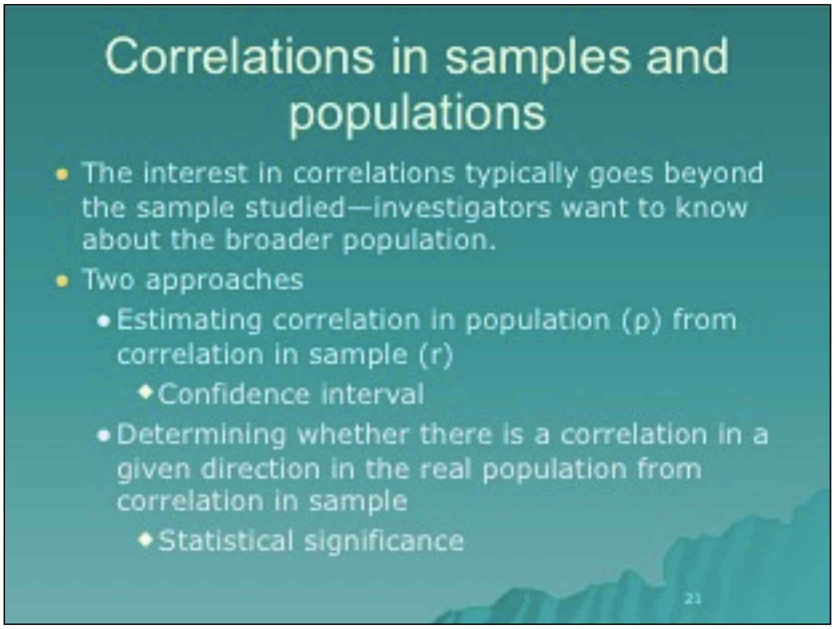
Clicker Question

You are told that the regression line relating a reasoning test score and a memory test score is

reasoning score = -3.25 + .7 memory score

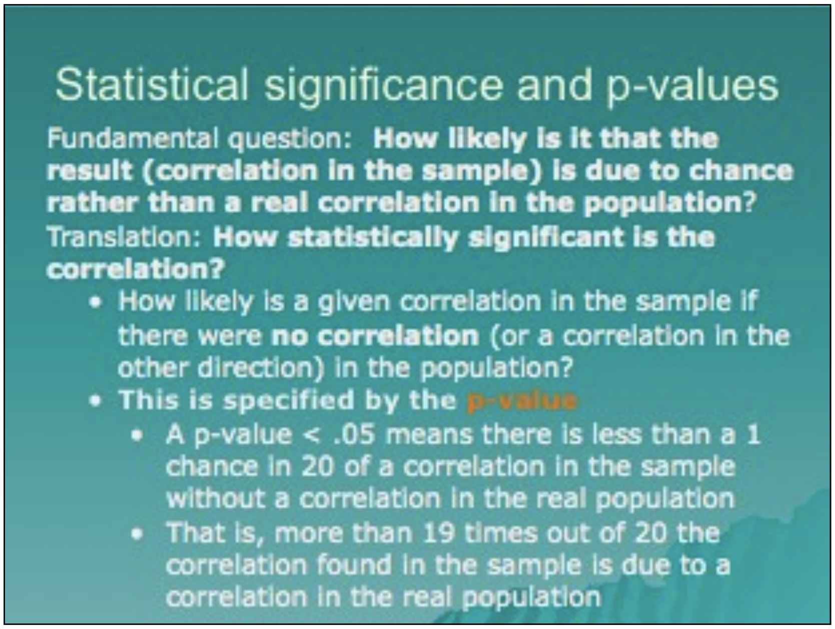
You know that

A. There is a positive correlation between the scores
B. There is a negative correlation between the scores
C. Pearson's r = .7
D. Pearson's r = -3.25



Correlations in samples and populations

- The interest in correlations typically goes beyond the sample studied—investigators want to know about the broader population.
- Two approaches
  - Estimating correlation in population (ρ) from correlation in sample (r)
    - Confidence interval
  - Determining whether there is a correlation in a given direction in the real population from correlation in sample
    - Statistical significance

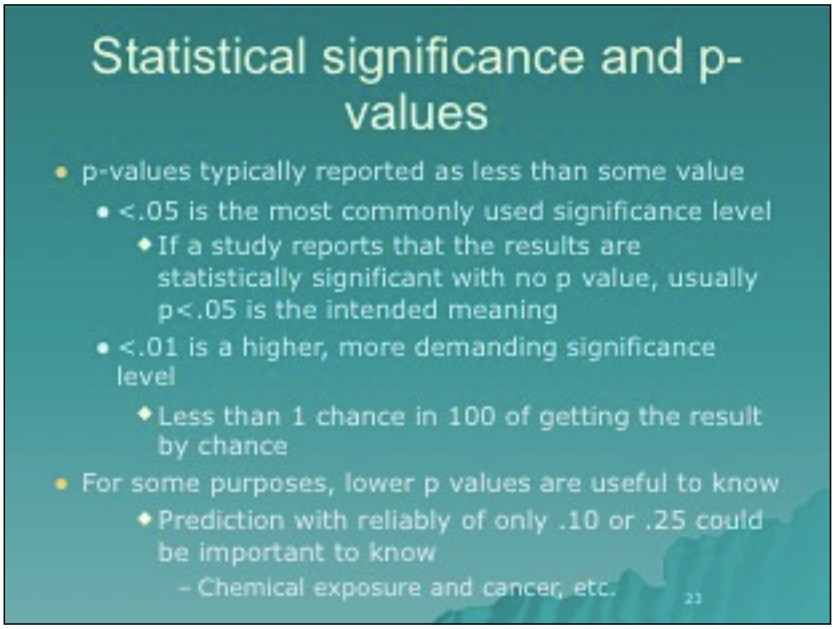


Statistical significance and p-values

Fundamental question: **How likely is it that the result (correlation in the sample) is due to chance rather than a real correlation in the population?**

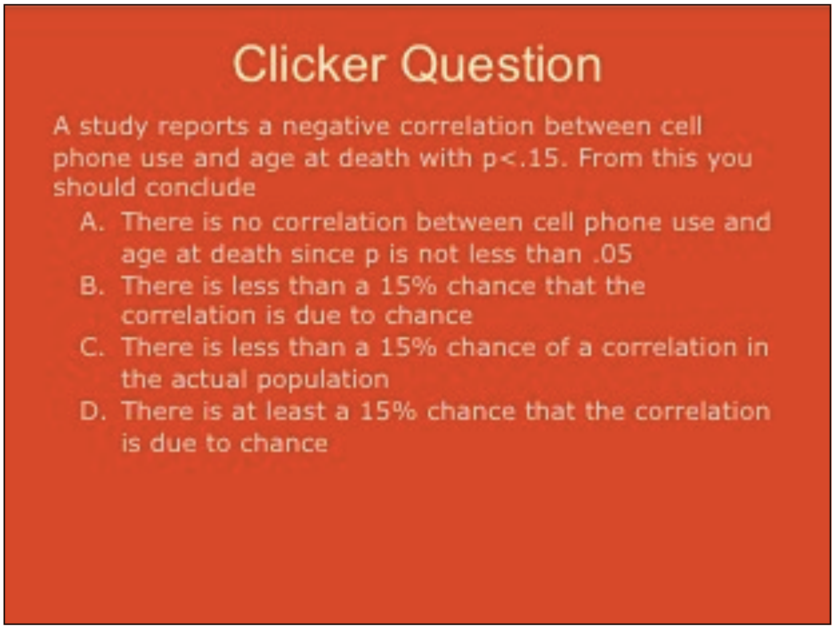
Translation: **How statistically significant is the correlation?**

- How likely is a given correlation in the sample if there were **no correlation** (or a correlation in the other direction) in the population?
- **This is specified by the p-value**
  - A p-value < .05 means there is less than a 1 chance in 20 of a correlation in the sample without a correlation in the real population
  - That is, more than 19 times out of 20 the correlation found in the sample is due to a correlation in the real population



Statistical significance and p-values

- p-values typically reported as less than some value
  - <.05 is the most commonly used significance level
    - If a study reports that the results are statistically significant with no p value, usually p<.05 is the intended meaning
  - <.01 is a higher, more demanding significance level
    - Less than 1 chance in 100 of getting the result by chance
- For some purposes, lower p values are useful to know
    - Prediction with reliably of only .10 or .25 could be important to know
      - – Chemical exposure and cancer, etc.



Clicker Question

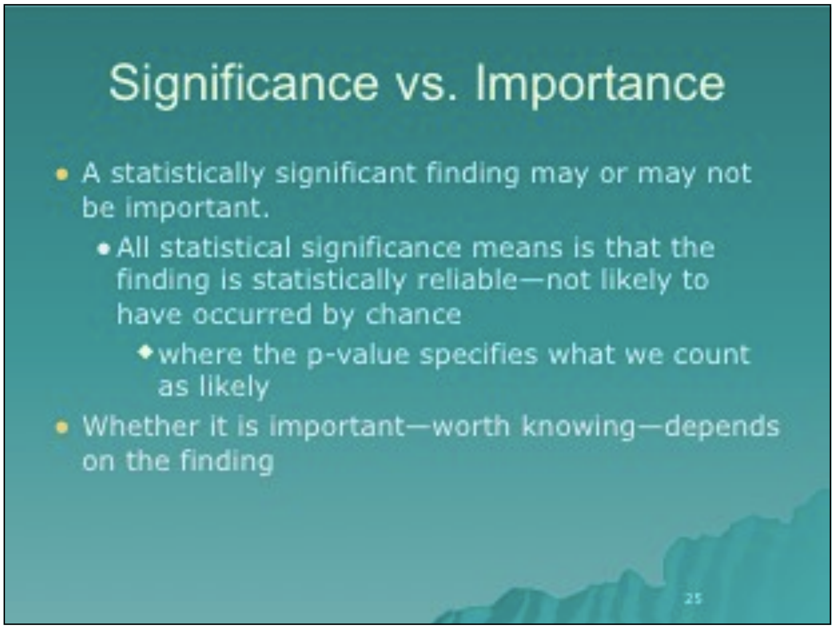
A study reports a negative correlation between cell phone use and age at death with p<.15. From this you should conclude

A. There is no correlation between cell phone use and age at death since p is not less than .05

B. There is less than a 15% chance that the correlation is due to chance

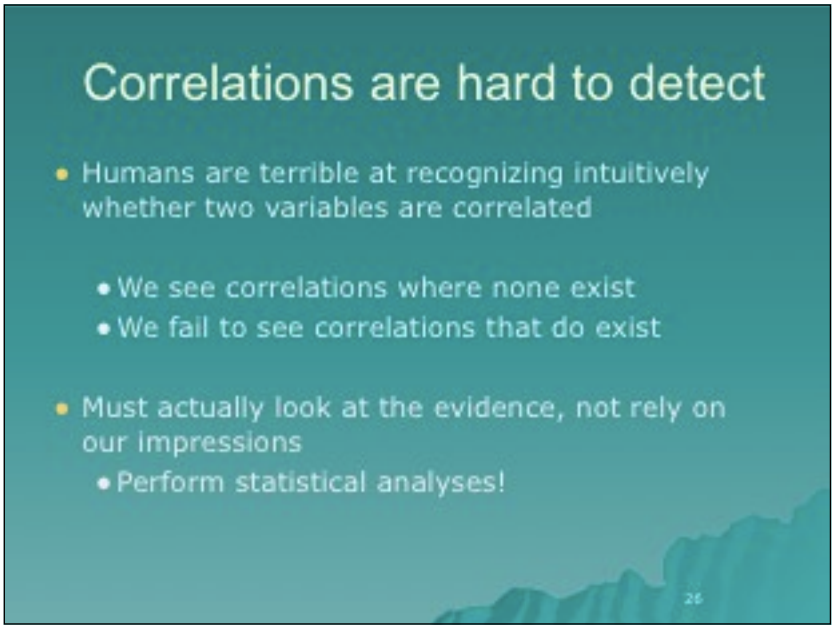
C. There is less than a 15% chance of a correlation in the actual population

D. There is at least a 15% chance that the correlation is due to chance



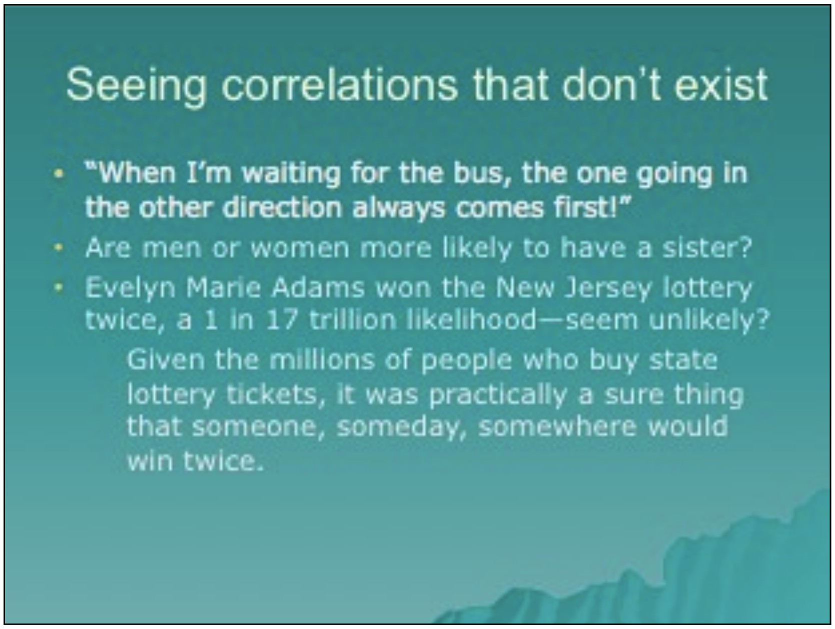
Significance vs. Importance

- A statistically significant finding may or may not be important.
  - All statistical significance means is that the finding is statistically reliable—not likely to have occurred by chance
    - where the p-value specifies what we count as likely
- Whether it is important—worth knowing—depends on the finding



Correlations are hard to detect

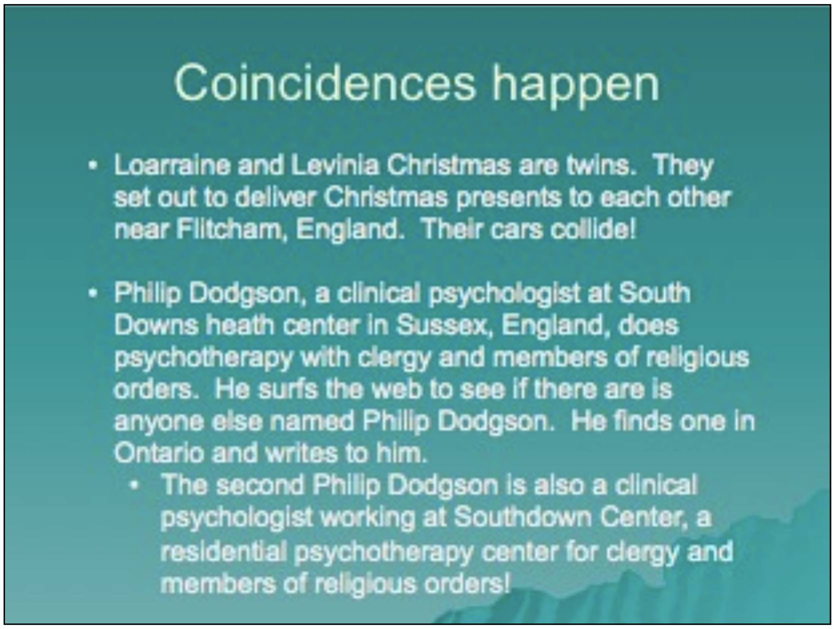
- Humans are terrible at recognizing intuitively whether two variables are correlated
  - We see correlations where none exist
  - We fail to see correlations that do exist
- Must actually look at the evidence, not rely on our impressions
  - Perform statistical analyses!



Seeing correlations that don't exist

- **"When I'm waiting for the bus, the one going in the other direction always comes first!"**
- Are men or women more likely to have a sister?
- Evelyn Marie Adams won the New Jersey lottery twice, a 1 in 17 trillion likelihood—seem unlikely?

  Given the millions of people who buy state lottery tickets, it was practically a sure thing that someone, someday, somewhere would win twice.

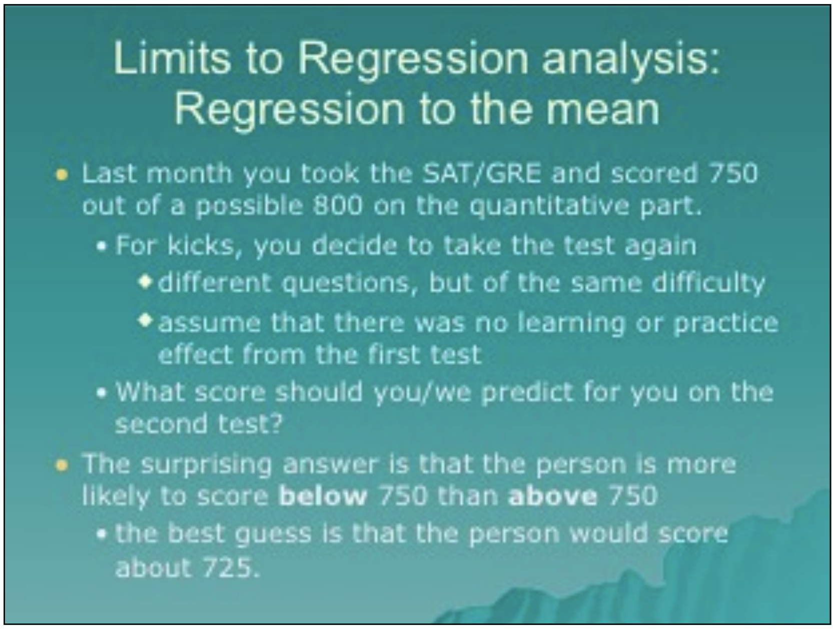


Coincidences happen

- Loarraine and Levinia Christmas are twins. They set out to deliver Christmas presents to each other near Flitcham, England. Their cars collide!
- Philip Dodgson, a clinical psychologist at South Downs heath center in Sussex, England, does psychotherapy with clergy and members of religious orders. He surfs the web to see if there are is anyone else named Philip Dodgson. He finds one in Ontario and writes to him.
  - The second Philip Dodgson is also a clinical psychologist working at Southdown Center, a residential psychotherapy center for clergy and members of religious orders!

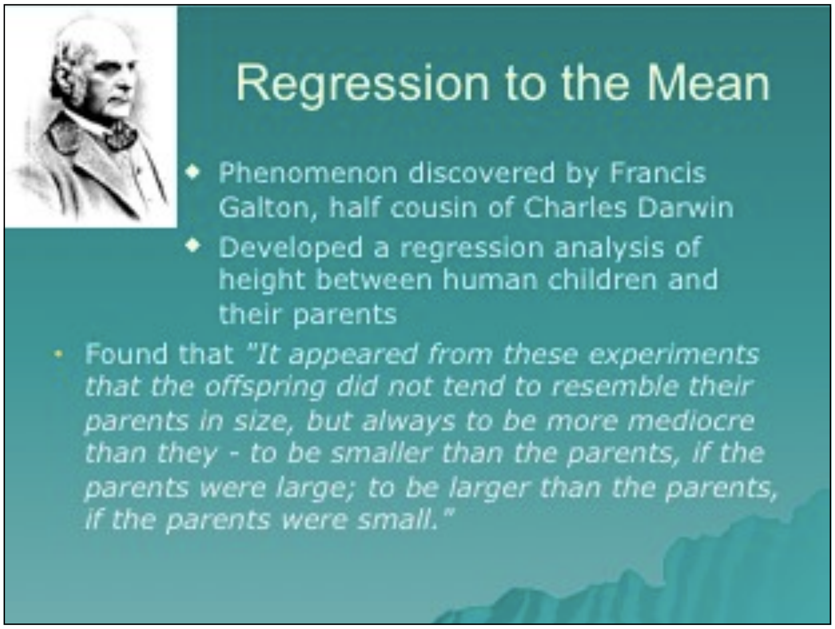
Coincidences happen

- Adams, Jefferson, and Monroe, three of the first five presidents of the US, died on the same date —July 4!
- Charles Schulz died of a heart attack on the day his last published Peanuts cartoon!



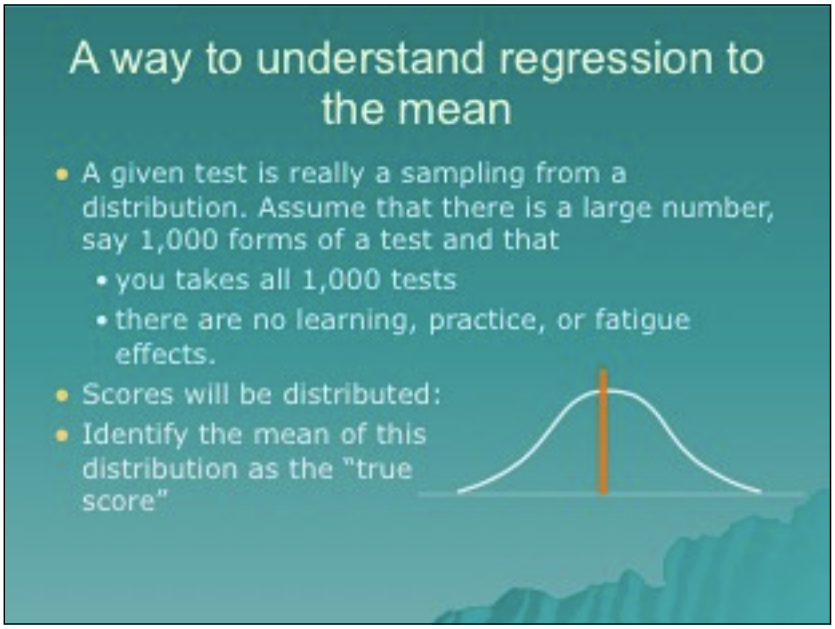
Limits to Regression analysis: Regression to the mean

- Last month you took the SAT/GRE and scored 750 out of a possible 800 on the quantitative part.
  - For kicks, you decide to take the test again
    - different questions, but of the same difficulty
    - assume that there was no learning or practice effect from the first test
  - What score should you/we predict for you on the second test?
- The surprising answer is that the person is more likely to score **below** 750 than **above** 750
  - the best guess is that the person would score about 725.



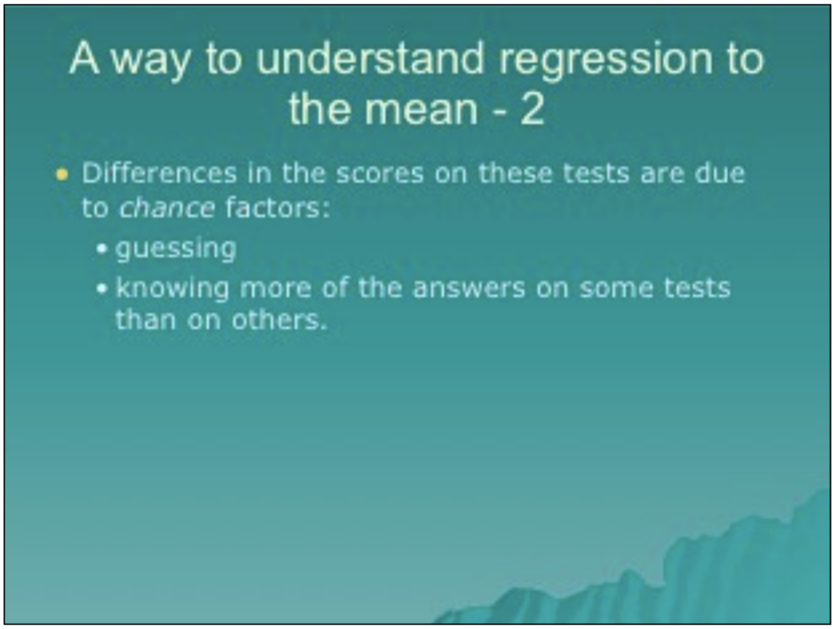
Regression to the Mean

- Phenomenon discovered by Francis Galton, half cousin of Charles Darwin
- Developed a regression analysis of height between human children and their parents
- Found that *"It appeared from these experiments that the offspring did not tend to resemble their parents in size, but always to be more mediocre than they - to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were small."*



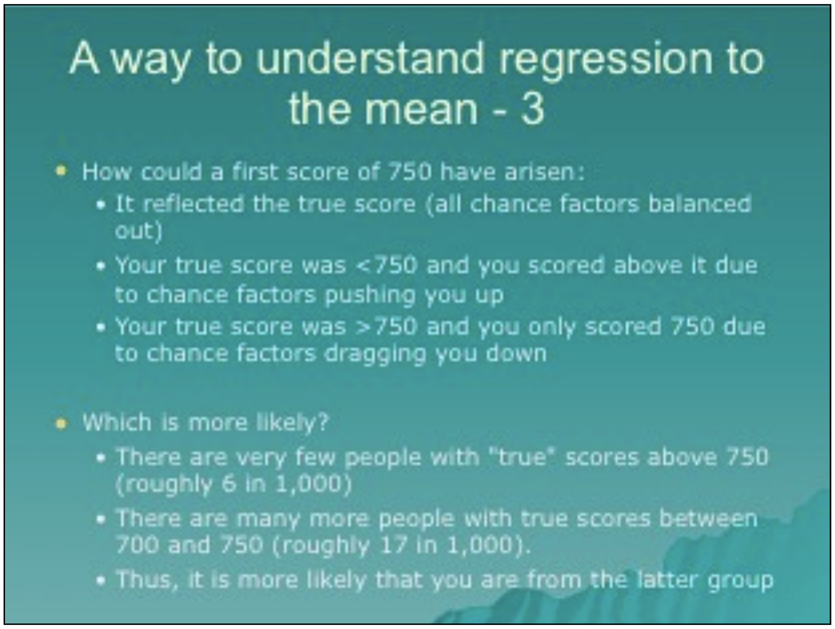
A way to understand regression to the mean

- A given test is really a sampling from a distribution. Assume that there is a large number, say 1,000 forms of a test and that
  - you takes all 1,000 tests
  - there are no learning, practice, or fatigue effects.
- Scores will be distributed:
- Identify the mean of this distribution as the "true score"



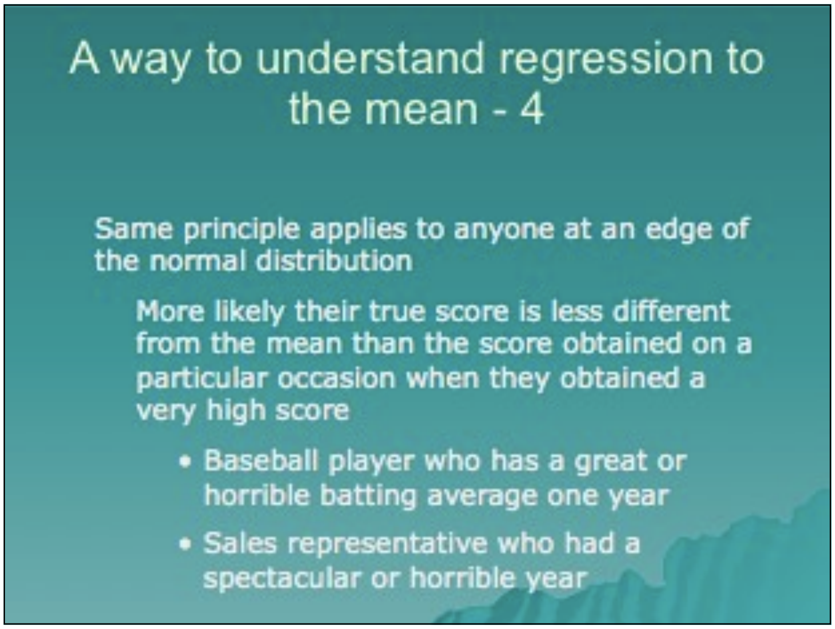
A way to understand regression to the mean - 2

- Differences in the scores on these tests are due to *chance* factors:
  - guessing
  - knowing more of the answers on some tests than on others.



A way to understand regression to the mean - 3

- How could a first score of 750 have arisen:
  - It reflected the true score (all chance factors balanced out)
  - Your true score was <750 and you scored above it due to chance factors pushing you up
  - Your true score was >750 and you only scored 750 due to chance factors dragging you down
- Which is more likely?
  - There are very few people with "true" scores above 750 (roughly 6 in 1,000)
  - There are many more people with true scores between 700 and 750 (roughly 17 in 1,000).
  - Thus, it is more likely that you are from the latter group

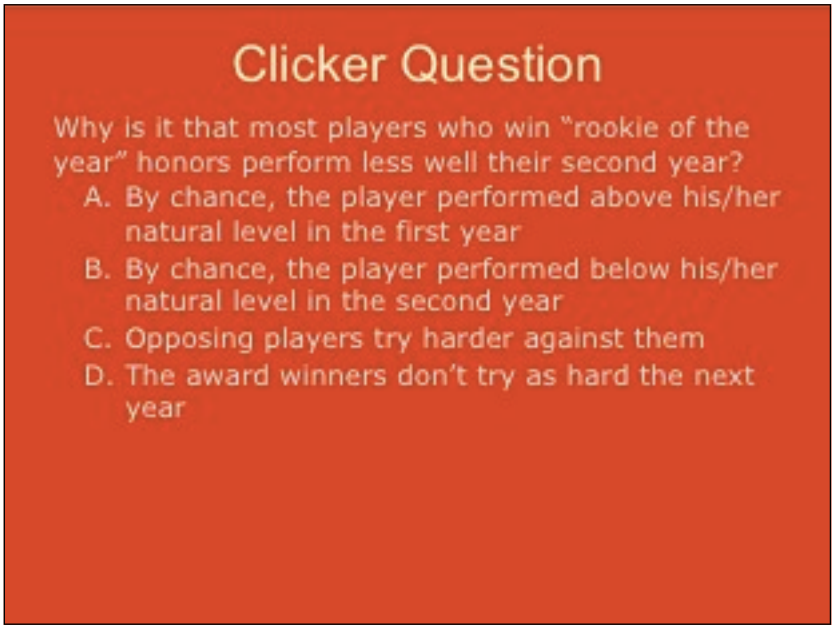


A way to understand regression to the mean - 4

Same principle applies to anyone at an edge of the normal distribution

More likely their true score is less different from the mean than the score obtained on a particular occasion when they obtained a very high score

- **Baseball player who has a great or horrible batting average one year**
- **Sales representative who had a spectacular or horrible year**



Clicker Question

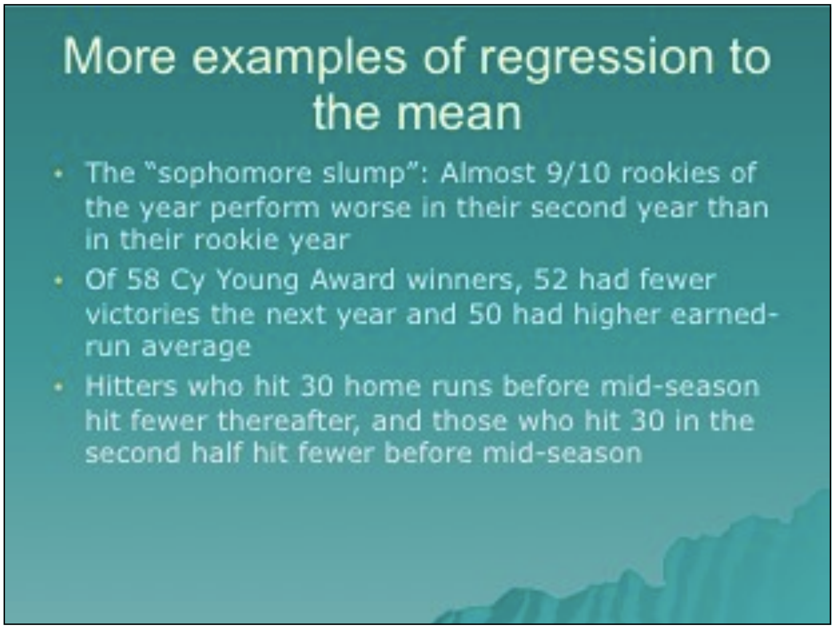
Why is it that most players who win "rookie of the year" honors perform less well their second year?

A. By chance, the player performed above his/her natural level in the first year

B. By chance, the player performed below his/her natural level in the second year

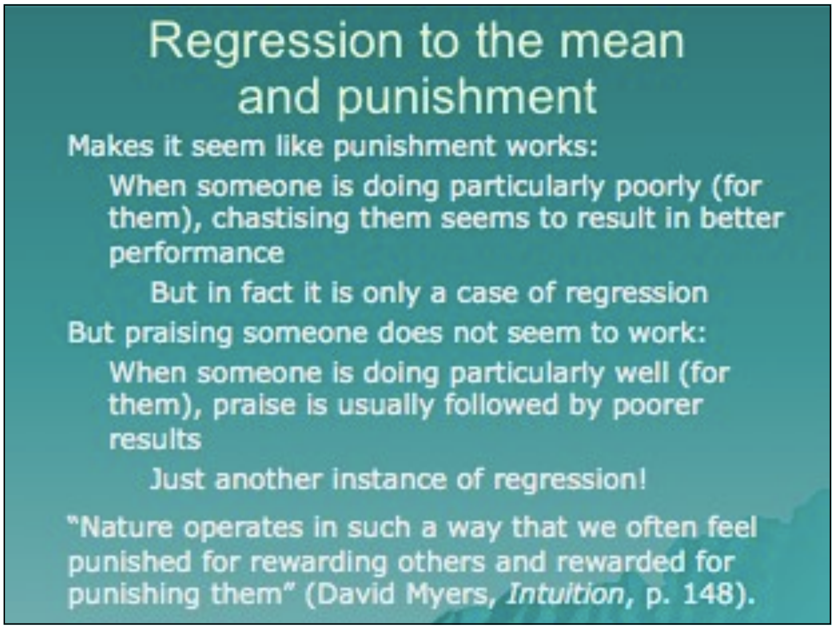
C. Opposing players try harder against them

D. The award winners don't try as hard the next year



More examples of regression to the mean

- The "sophomore slump": Almost 9/10 rookies of the year perform worse in their second year than in their rookie year
- Of 58 Cy Young Award winners, 52 had fewer victories the next year and 50 had higher earned-run average
- Hitters who hit 30 home runs before mid-season hit fewer thereafter, and those who hit 30 in the second half hit fewer before mid-season



Regression to the mean and punishment

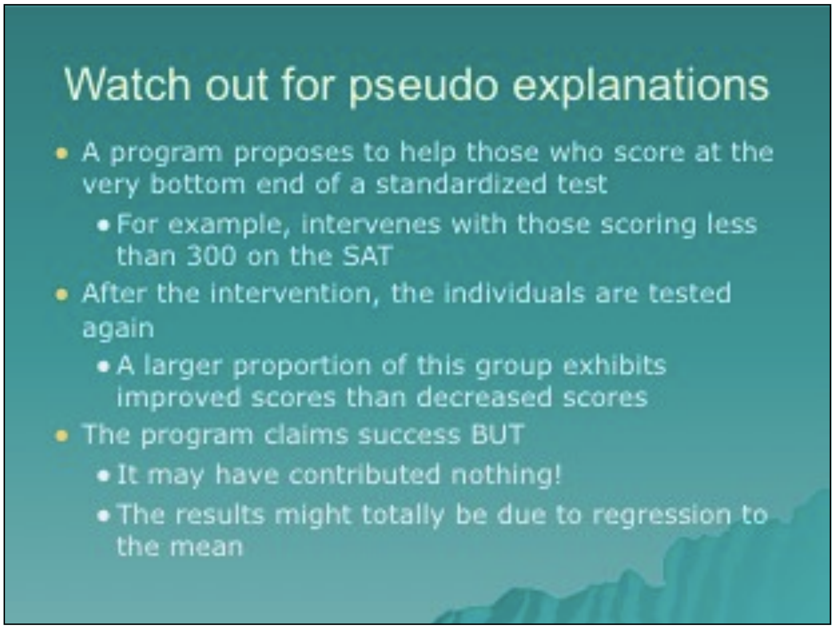
Makes it seem like punishment works:

- When someone is doing particularly poorly (for them), chastising them seems to result in better performance
  - But in fact it is only a case of regression

But praising someone does not seem to work:

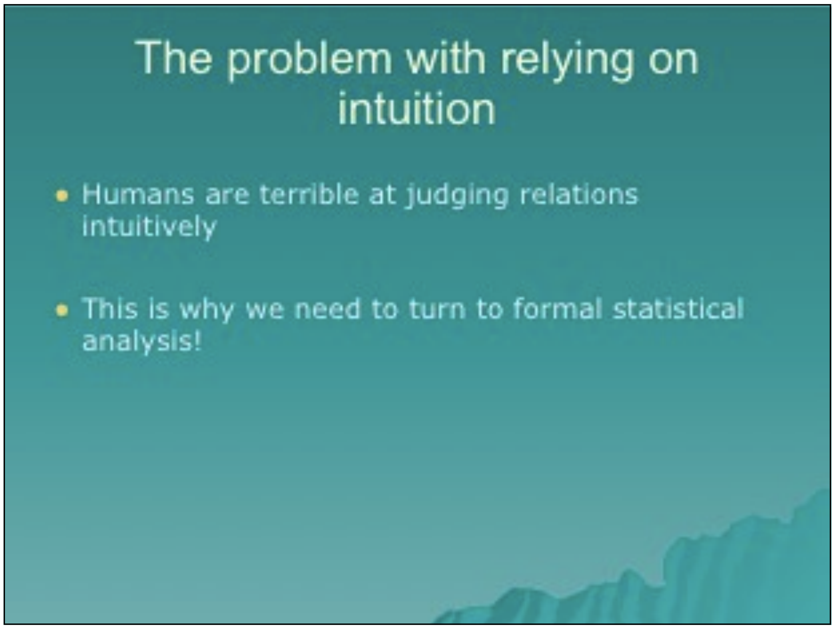
- When someone is doing particularly well (for them), praise is usually followed by poorer results
  - Just another instance of regression!

"Nature operates in such a way that we often feel punished for rewarding others and rewarded for punishing them" (David Myers, *Intuition*, p. 148).



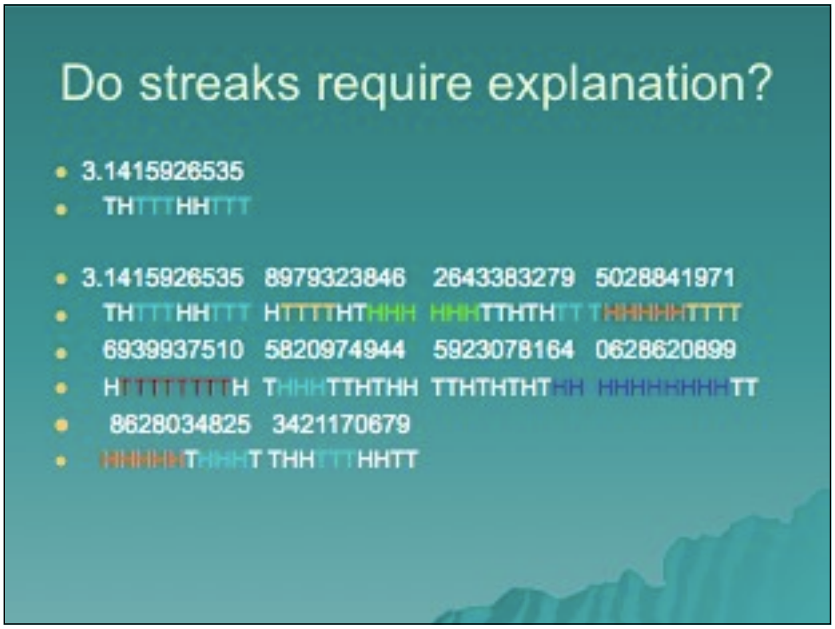
Watch out for pseudo explanations

- A program proposes to help those who score at the very bottom end of a standardized test
  - For example, intervenes with those scoring less than 300 on the SAT
- After the intervention, the individuals are tested again
  - A larger proportion of this group exhibits improved scores than decreased scores
- The program claims success BUT
  - It may have contributed nothing!
  - The results might totally be due to regression to the mean



The problem with relying on intuition

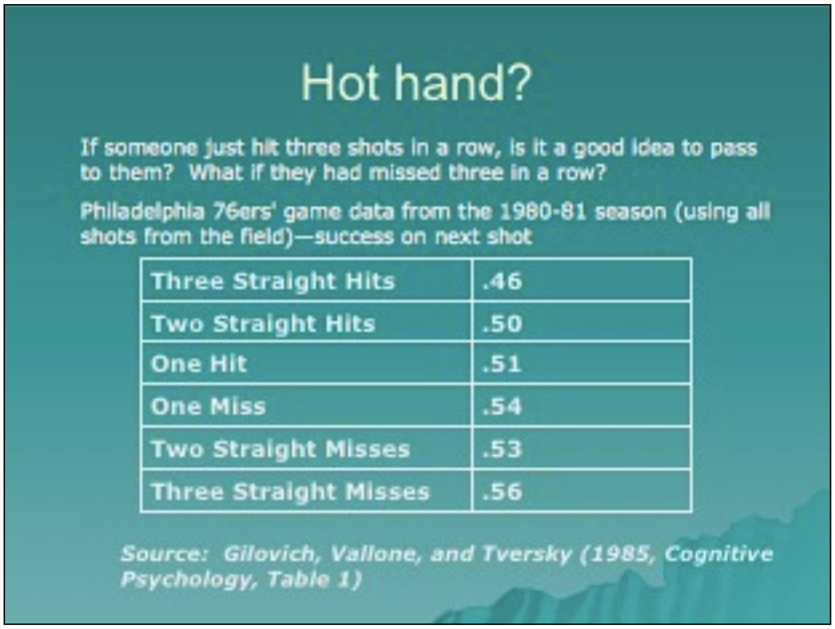
- Humans are terrible at judging relations intuitively
- This is why we need to turn to formal statistical analysis!



Do streaks require explanation?

- 3.1415926535
- THTTTHHTTT

- 3.1415926535 8979323846 2643383279 5028841971
- THTTTHHTTT HTTTTHTHHH HHHTTHTHTT THHHHHTTTT
- 6939937510 5820974944 5923078164 0628620899
- HTTTTTTTTH THHHTTHTHH TTHTHTHTHH HHHHHHHHTT
- 8628034825 3421170679
- HHHHHTHHHT THHTTTHHTT



Hot hand?

If someone just hit three shots in a row, is it a good idea to pass to them? What if they had missed three in a row?

Philadelphia 76ers' game data from the 1980-81 season (using all shots from the field)—success on next shot

Three Straight Hits	.46
Two Straight Hits	.50
One Hit	.51
One Miss	.54
Two Straight Misses	.53
Three Straight Misses	.56

Source: Gilovich, Vallone, and Tversky (1985, Cognitive Psychology, Table 1)