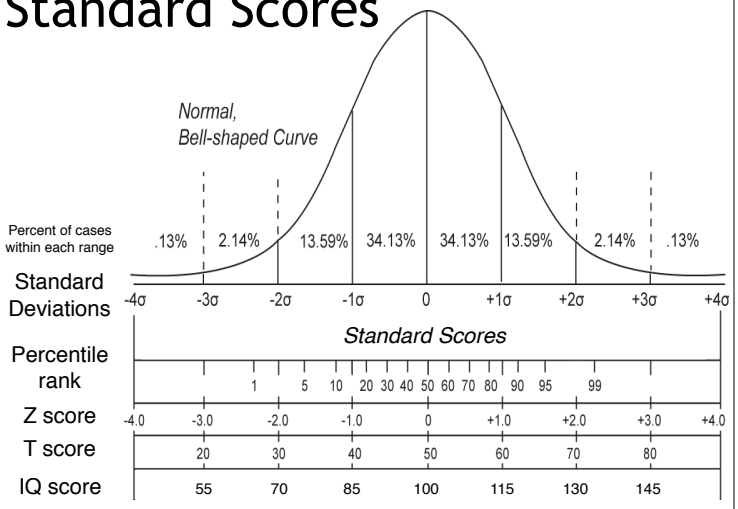
Ch. 3: Correlation & Linear Regression

Review

- Norms and Standard Scores:
  - Rank, Percentile Rank
  - Z, IQ, T

Standard Scores



Ch. 3: Correlation & Linear Regression

- Relationships between 2 variables
- Scatterplots
- Linear Regression
- Exercise 2
- Correlation
- Race / DNA

Number of variables

- One variable, one dimension
- Number Line
- Frequency Distribution / Histogram
  - 2 dimensional graph of 1D data

- Difference Score
  - 1 dimension
  - 2 dimensions

Bivariate relationships

- “is factor A related to factor B”?
- Methods of analysis…
  - Anecdotal / Clinical
  - Numerical : simple 2x2 analysis
  - Visually -- scatterplots
    - see relationships and problems w/data
    - can’t test hypothesis
  - Statistically -- correlation & regression
    - hard to detect problems w/data
    - easy to test hypothesis

Anecdotal / Clinical

- Many interesting findings began from non-scientific approaches
- "Intuition" that something is related through experiencing multiple situations
- Pattern recognition - Good and Bad
- Problems -- faulty memory, confirmation biases, prejudice, etc…
- Next step after a "gut" feeling : design experiment and collect data.

Simple numerical analysis

- Simplify:
  - use categorical variables
  - or convert continuous variables to categorical
- Use extreme cases to maximize effect
- Compute percentages in a 2x2 matrix
- Do the results suggest an effect?

- Compute Chi-square statistic to judge significance

Dichotomous Variables

- The simplest form of categorical
- Aka "Binary"
- Examples:
  - 1/0
  - yes/no
  - pass/fail
  - true/false
  - healthy/sick
  - normal/impaired
  - etc.

Example

- "I think there is brain dysfunction in HIV disease" as measured by neuropsychological testing
- Medical status: control vs. HIV+ symptomatic
- NP test results: normal vs. impaired

		Medical Status: Control	Medical Status: HIV+
NP Status	Normal	85%	52%
	Impaired	15%	48%

2x2 Analysis

- Pro: easy to understand
- Con: using binary categories reduces power

- Conclusion: other Graphical and Statistical methods should be used as well.

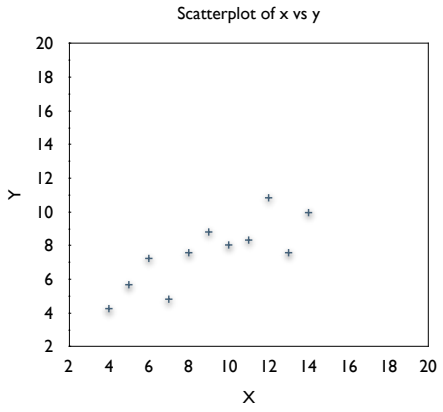
Scatterplots

- Graph two variables in relation to each other on two-dimensional X, Y axis
- Easy to see
  - relations
  - problems

- Can't prove relationship is "significant"
- Difficult to interpret clinically or in "common sense" terms

Scatterplots

x	y
10	8.04
8	7.58
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68



Linear Regression

- Assume that two variables are related, and that this relationship is linear -- model the data by a simple straight line for the data.
- For any given data set, we pick the line that best "fits" our data
- Similar terms: linear regression, fitting a line, finding the trend, creating a trendline, best fit line, etc.
- Residuals = difference between prediction and actual value
- Linear Regression minimizes the square of the residuals, often called "Ordinary Least Squares"

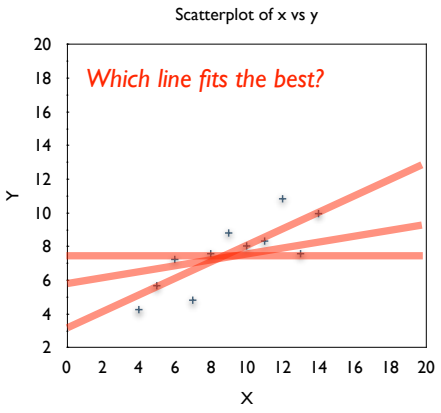
Why "Regression"

- Frances Galton
- Height of children vs parents.
- Tall parents have tall children (and vice versa)
- But children are closer to the mean than their parents (by a factor of ~2/3)
- Galton called this "Regression to the Mean"
- His paper fit** straight lines to data points.
- The technique has been called "regression" ever since
- ** He never calculated the lines, he just eyeballed them

Linear Regression

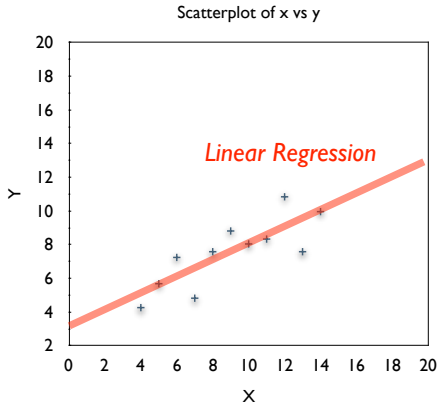
Equation:
$y = 3.0 + 0.5x$

Correlation
$r_{x,y} = 0.816$



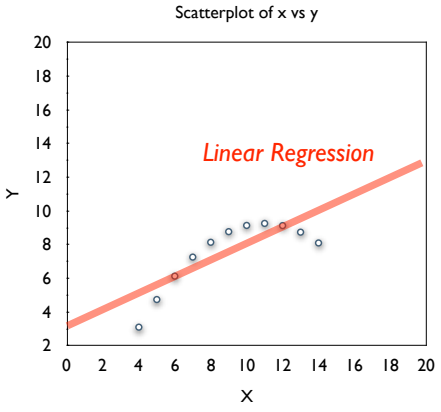
Anscombe's Quartet I

x	y
10	8.04
8	7.58
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68



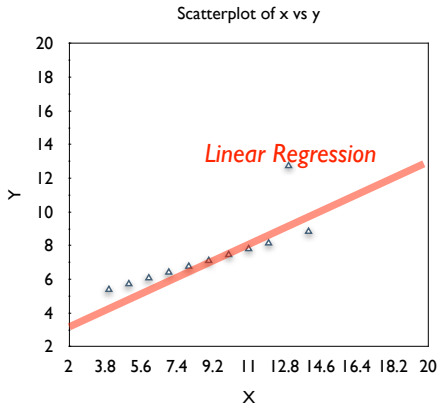
Anscombe's Quartet II

x	y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74



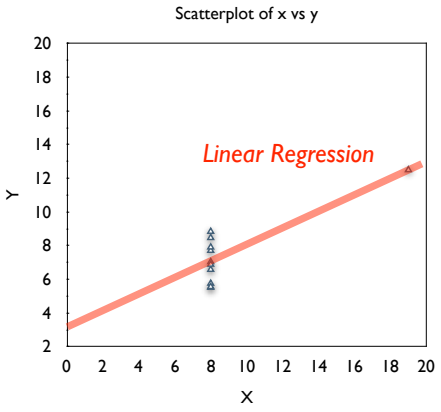
Anscombe's Quartet III

x	y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

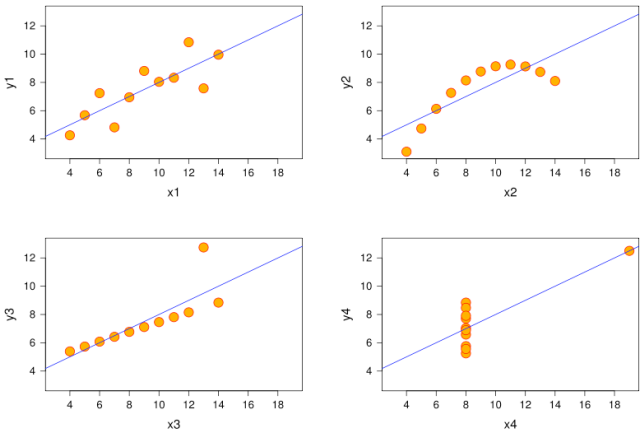


Anscombe's Quartet IV

x	y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.52
19	12.5
8	5.56
8	7.91
8	6.89



Anscombe's Quartet



Anscombe's Quartet Summary

- Each series has the same Quantitative stats:
  - linear regression equations
  - correlations
- Each one is Qualitatively different
- Each series needs special handling
- Lesson
  - Graph Your Data

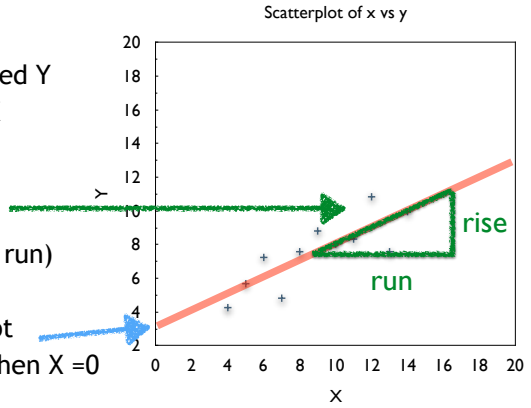
Linear Regression Equation

Y' = a + bX

Y' = predicted Y
X = actual X

b = slope
DY/DX
(rise over run)

a = intercept
Y value when X =0



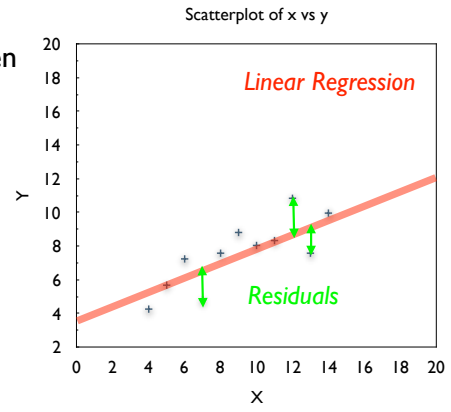
Residuals in Linear Regression

- X : independent variable
- Y : dependent variable
- Model: predict Y from X
- Y' : (Y prime) : predicted Y
- Y' = a + bX
- Prediction is imperfect.
- Difference between predicted (Y') and actual (Y) is called a "Residual" = (Y - Y')
- Calculation of best fit line minimizes the sum of the squared residuals $\Sigma(Y-Y')^2$

Residuals in Linear Regression

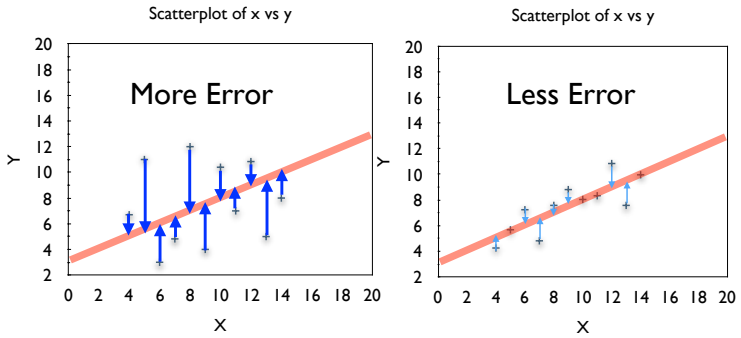
Residuals are difference between actual Y and predicted Y' (Y - Y')

Graphically it is equal to how far away (vertically) a point is from the linear regression line



Residuals and Error

Residuals (error) are greater when Y values are further from prediction.



Residuals

$$d_i = y_i - y_i'$$

- Difference between predicted and actual y value

Sum of Squares

$$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{N} d_i^2$$

$$SSR = \sum_{i=1}^{N} (y_i - y_i')^2$$

Sum of Squared Residuals

- Residual = (Y - Y')
- Squared residual = $(Y-Y')^2$
- SSR: Sum of squared residuals
  - Linear regression minimizes this value
- SSR is hard to interpret

R^2

$$R^2 = 1 - \frac{SSR}{SST}$$

- R^2 = 1 - (SSR/SST)
- Ranges from 0 to 1 (0% to 100%)

R^2

- Terminology
  - Coefficient of *Determination*
  - *Explained* Variance
  - Shared Variance
  - Covariance
- Meaning
  - what % of variation in Y values can we predict from the X values
- Careful: *Correlation* is not causation

Review

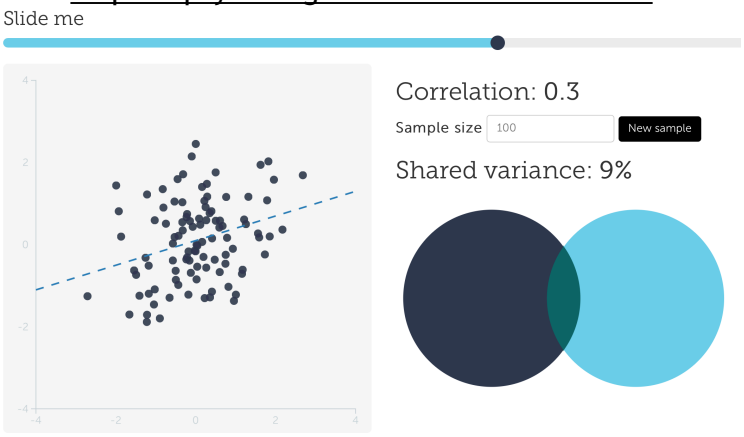
- # of variables / dimensions
  - 1
  - 2
- Kinds of statistics
  - descriptive
  - inferential
- Linear Regression

Residuals, Variance, R^2

- Residual = (Y - Y')
- Squared residual = $(Y-Y')^2$
- Sum of squared residuals = $\Sigma(Y-Y')^2$
  - Linear regression minimizes this value
- SSR is hard to interpret
- R^2
  - R^2 = 1 - (SSR/SST)
  - Coefficient of Determination
  - Explained Variance
  - Ranges from 0 to 1 (0% to 100%)

Interactive Correlation Demo

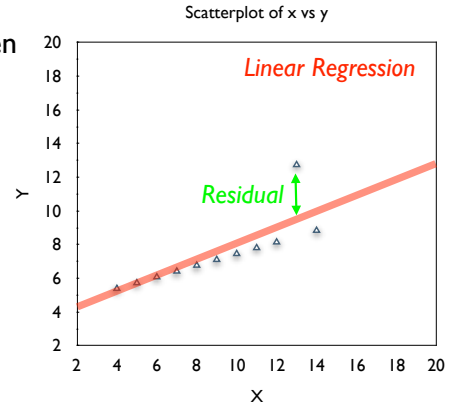
- http://rpsychologist.com/d3/correlation/



Residuals in Linear Regression

Residual is difference between actual Y and predicted Y' (Y - Y')

Graphically it is equal to how far away (vertically) a point is from the linear regression line



Residuals, Variance, R^2

- Residual = (Y - Y')
- Squared residual = $(Y-Y')^2$
- Sum of squared residuals = $\Sigma(Y-Y')^2$
  - Linear regression minimizes this value
- SSR is hard to interpret
- R^2
  - R^2 = 1 - (SSR/SST)
  - Coefficient of Determination
  - Explained Variance
  - Ranges from 0 to 1 (0% to 100%)

Standard Error of Estimate

- Residual = (Y - Y')
- Standard Deviation of residuals
  - measure of "average" error
  - aka "Standard Error of Estimate"
  - In Prism: $S_{y.x}$

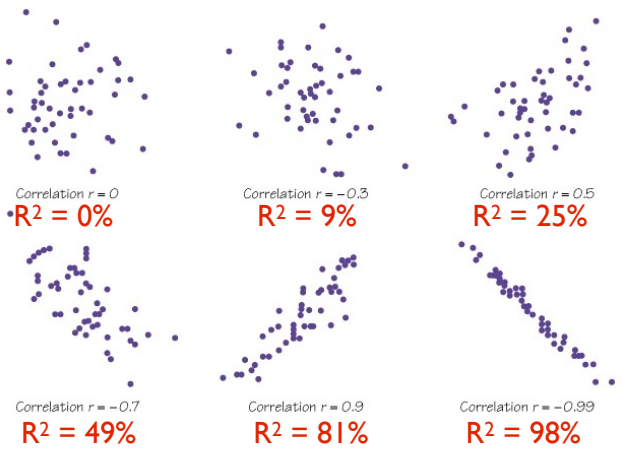
Correlation (r) Pearson's r

- Pearson's Product-Moment Correlation
- Measures the strength of the linear relationship between two variables
- Ranges between -1.0 and +1.0
- Is a special case of linear regression, when both X and Y have been turned into Z scores.
- r is ~~transitive~~ **commutative** (correlation between X and Y is same as correlation between Y and X)
- R^2 = "explained variance" is the proportion of variation in the data explained by the model.
- R^2 ranges from 0 to 1.0 (0% to 100%)

Regression vs. Correlation

	Linear Regression	Correlation
Scores	Raw	Z
Mean, Std Dev	sample means sample Std Dev	0 1
Equation	Y' = a + bX	Y' = r X
Slope	b = change in Y per change in X	r = correlation coefficient
Slope²	meaningless	R^2 = % variance explained
Commutative ?	no	yes, $R_{xy} = R_{yx}$

Correlations



Interactive Correlation Example

- http://rpsychologist.com/d3/correlation/

- R^2 or "Explained Variance" is sometimes called "Shared Variance"

Other Correlation Coefficients

- Continuous (interval & ratio): Pearson's r

- Ordinal (Ranked): A B C D… 1st, 2nd, 3rd...
  - Spearman's Rho: correlation between two ordinal / ranked variables.

- Dichotomous (yes/no, one/zero, T/F, Male/Female, Pass/Fail...)
  - True vs. Artificial?

Continuous vs. Dichotomous

Type of X / Type of Y	Continuous	Artificial Dichotomous	True Dichotomous
Continuous	Pearson r	Biserial r	Point biserial r
Artificial Dichotomous	Biserial r	Tetrachroic r	Phi
True Dichotomous	Point biserial r	Phi	Phi

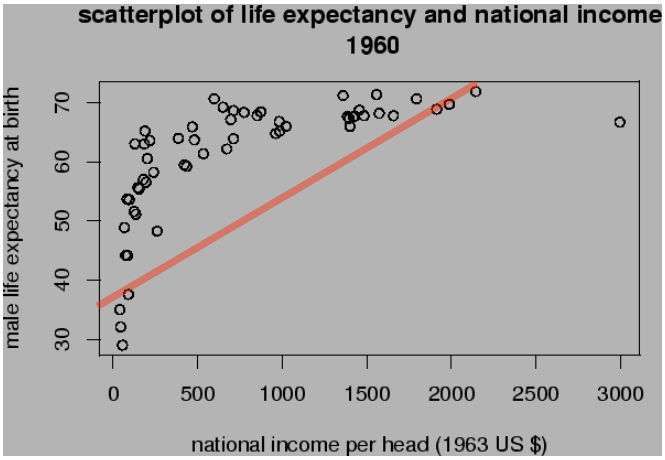
Correlation : Issues

- Technical / Calculation :
  - Non-normal distribution
  - Non-linear data and relationships
  - Outliers, data errors
  - Restricted Range
- Interpretation:
  - Correlation =? Causation
  - Third variable explanations

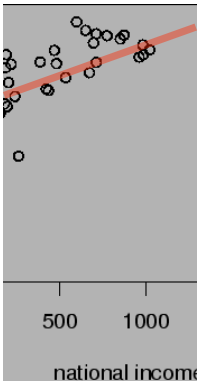
Non-linearity

- Linear Regression & Correlation assume a linear relationship between X and Y
- When it's not linear:
  - Restrict the range of X
  - Transform (log, square root, etc.)
  - other statistical analyses (Spearman's Rho…)

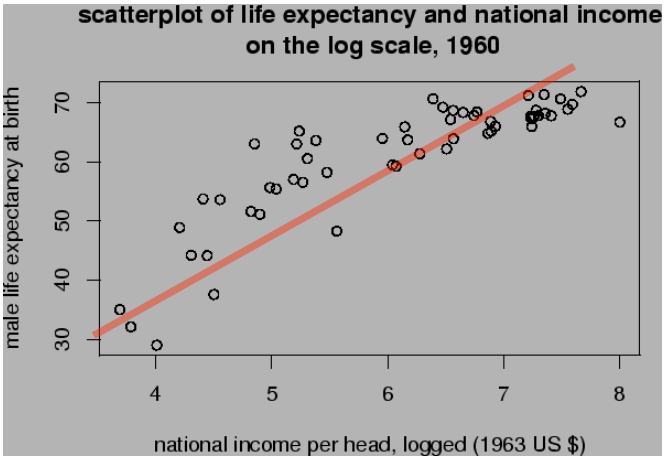
Life expectancy / national income



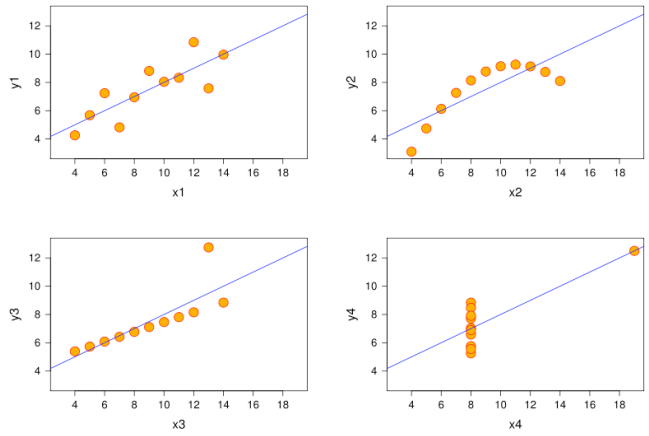
Restrict range of X



log transform X (or Y)



Outliers & Data Errors?



Correlation = Causation?

- A relationship (linear or otherwise) between X and Y tells us nothing about whether X causes Y
- Lack of correlation between X and Y does not mean that X doesn't cause Y

- Ice cream sales are positively related to increases in drowning deaths

Hypothesis Testing

- All parameters estimated from sample data have inherent error
- How do we know if a given estimate is correct?
- How big is the error likely to be (confidence intervals)?
- Inferential Statistics - covered later
  - Formulas to calculate probability, confidence intervals.
  - Higher N is better
  - "statistical significance" not the same as "clinical significance"

Statistical vs Clinical Significance

- Regarding the change in the Dependent Variable (DV)
- Statistical Significance:
  - Could the change be due to chance?
  - P value ($p < .05$: less than 5% probability)
- Clinical Significance
  - Was the change big enough to matter?
  - Effect Size (R^2)
  - Depends on context

Significance vs. Effect Size

- Two coin flips : both heads (100%)
  - big effect (50%)
  - not statistically significant (p=0.25)
- 1000 coin flips, 490 heads (49.0%)
  - small effect (1%)
  - statistically significant (p=0.02)
- 1000 coin flips, 350 heads (35%)
  - big effect (15%)
  - statistically significant (p<.00000001)

Lies, damned lies, and statistics

- Statistical significance (P) is a function of…
  - Errors of measurement (E)
  - Effect Size (D)
  - Sample Size (N)

- $p \sim E / (D \times N)$

Reporting Results

- "Men had higher IQ than women. Results were statistically significant p < .001"
- P-value : yes
- Effect Size : ?

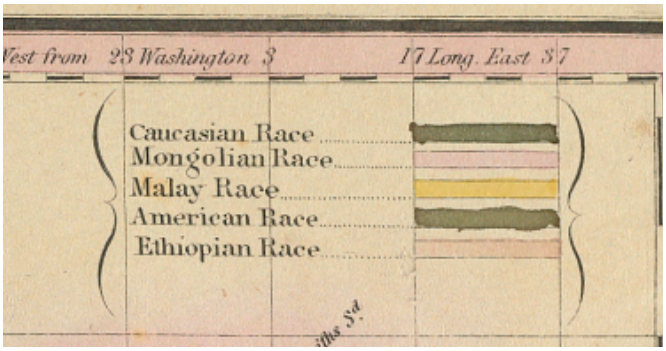
Review : Is race "real"?

- Pre-DNA theory
- Post-DNA theory

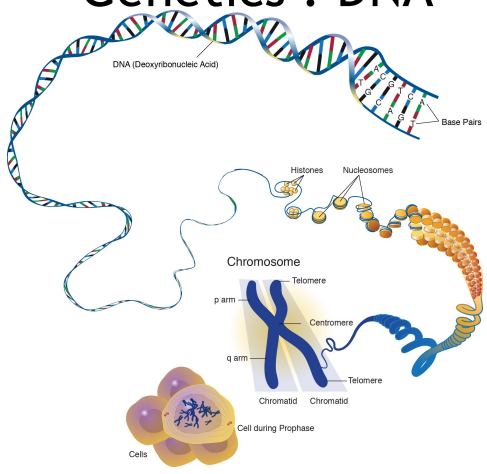
Pre-DNA

- Gold, Silver, Brass, Iron -- Plato
- "There is a physical difference between the white and black races which I believe will for ever forbid the two races living together on terms of social and political equality." -- Abraham Lincoln

Five Races?



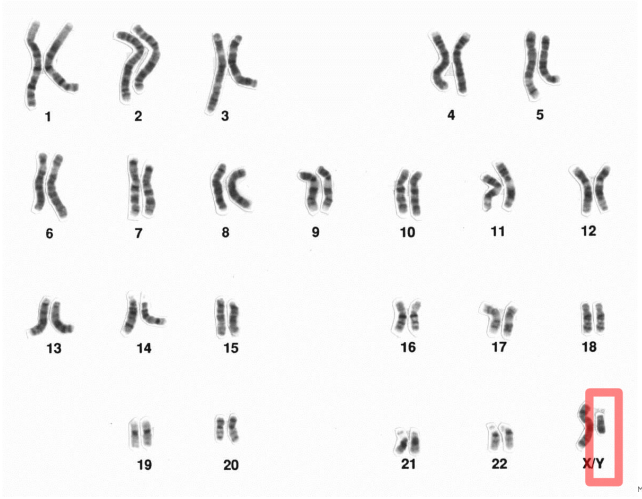
Genetics : DNA



Genetics

- Human genome contains about 4 billion pairs of deoxyribonucleic acid (DNA)
- DNA is Transcribed into RNA
- RNA is Translated into Proteins
- Proteins
  - serve as structural components
  - function as enzymes to catalyze biochemical reactions
- Human DNA is grouped into 46 chromosomes
  - 23 pairs, one of each pair comes from each parent
  - 22 pairs in both males and females (autosomes)
  - 1 pair determines sex: either "XX" (females) or "XY" (males)

Humans: 46 Chromosomes - 23 pairs



Genetics : Species Differences

organism	estimated size (base pairs)	# genes	gene size	# chromosomes
Homo sapiens (human)	3.2 billion	~25,000	1 gene per 100,000 bases	46
Mus musculus (mouse)	2.6 billion	~25,000	1 gene per 100,000 bases	40
Drosophila melanogaster (fruit fly)	137 million	13,000	1 gene per 9,000 bases	8
Arabidopsis thaliana (plant)	100 million	25,000	1 gene per 4000 bases	10
Caenorhabditis elegans (roundworm)	97 million	19,000	1 gene per 5000 bases	12
Saccharomyces cerevisiae (yeast)	12.1 million	6000	1 gene per 2000 bases	32
Escherichia coli (bacteria)	4.6 million	3200	1 gene per 1400 bases	1
H. influenzae (bacteria)	1.8 million	1700	1 gene per 1000 bases	1

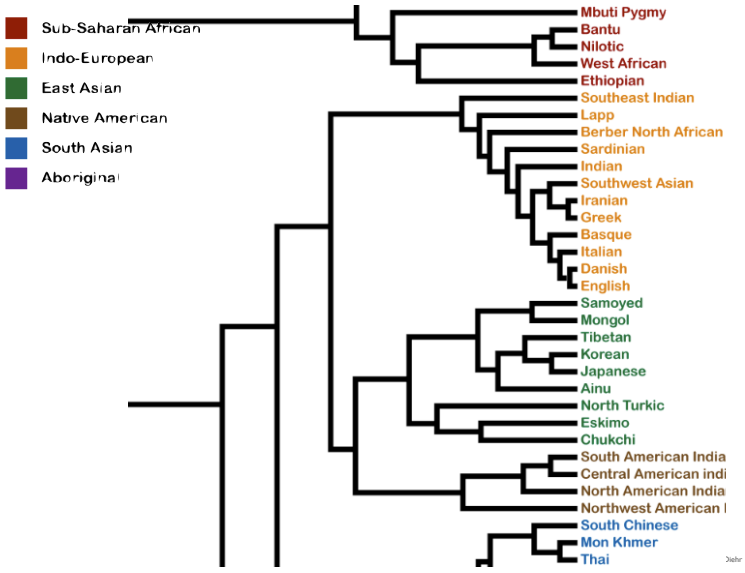
Visible differences?

Indigenous Australian
Melanesia
African
European

Australian and Africans are most genetically different

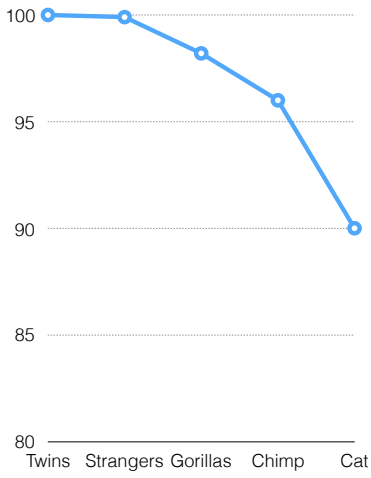
Post-DNA theory

- Variance
  - variation between individuals
    - aka variation *within* ~~races~~ *population groups*
  - variation *between population groups*
- Variance
  - variation between individuals : 3mbp / person
  - variation within groups : 85%
  - variation between groups: 15%
    - 5% - within *population groups*
    - 10% - between *population groups*



DNA Differences

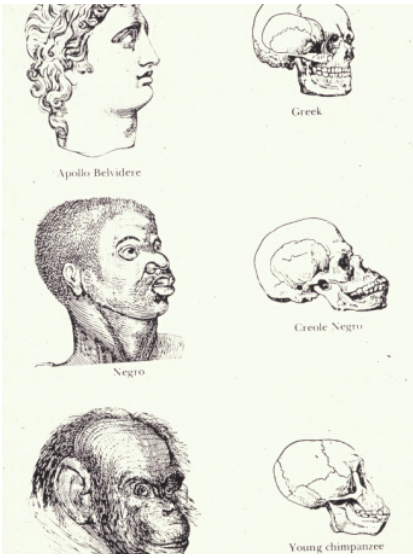
- Identical Twins
  - 0.0%
- Human vs. Human
  - 0.1%
- Humans vs Gorillas
  - 1.6%
- Humans vs Chimps:
  - 4.0%
- Humans vs. Cats
  - 10.0%



Indigenous Races of the Earth

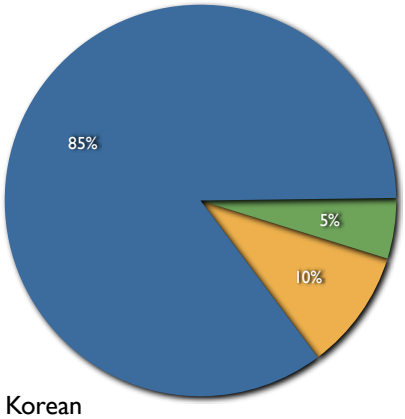
note the false exaggeration of the chimp and negro skills to suggest that "blacks might even rank lower than the apes"

(Gould, p. 65, citing Nott & Gliddon, 1868)



Variance: Genetic Variation

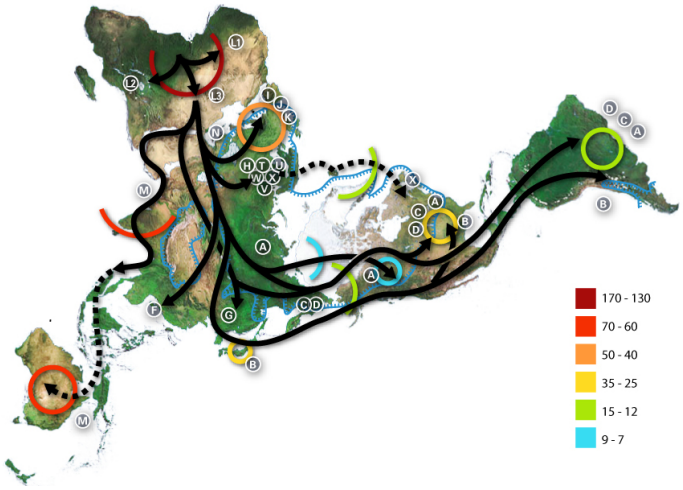
- Within local populations
- Within "race"
- Between "race"



For example:

- 85% within Japanese
- 5% between Japanese & Korean
- 10% between Asian and Caucasian

Prehistorical Migration



'Cheddar Man,' Britain's Oldest Skeleton, Had Dark Skin, DNA Shows

By CEYLAN YEGINSU and CARL ZIMMER FEB. 7, 2018

A likeness of "Cheddar Man," Britain's oldest complete human skeleton, at the Natural History Museum. Genetic evidence showed that he was dark-skinned and blue eyed, scientists said.
London Natural History Museum

"Fair skin, long considered a defining feature of Europe, only goes back less than 6000 years…"