Ch. 3: Correlation & Regression

- Exploring relationships between 2 variables
- Scatterplots
- Linear Regression
- Exercise 02
- Correlation
- Other Correlation Coefficients

Bivariate relationships

- “is factor A related to factor B”?
- Methods of analysis:
  - Anecdotal / Clinical -- often forms the basis for further systematic research & data collection
  - Numerically -- check values & % at extremes
  - Visually -- scatterplots
    - easy to see relationships and problems w/ data
    - hard to prove / test
  - Statistically -- correlation & regression
    - hard to detect problems w/data
    - easy to test hypothesis

Anecdotal / Clinical

- Many interesting findings in psychology first originate from non-scientific approaches
- “Intuition” that something is related through experiencing multiple situations
- Pattern recognition
- Human brains are both excellent and terrible pattern recognizers
- Problems -- faulty memory, confirmation biases, prejudice, etc…
- First step after a “gut” feeling is to begin collecting data.

Simple numerical analysis

- Simplify the situation by using Categorical variables (or reducing Continuous variables to Categorical variables)
- 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

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	HIV+
NP Status	Normal	85%	52%
	Impaired	15%	48%

Issues

- Pro: easy to understand
- Con: dividing continuous variables into binary reduces power

- Graphical and Statistical methods should be used as well.

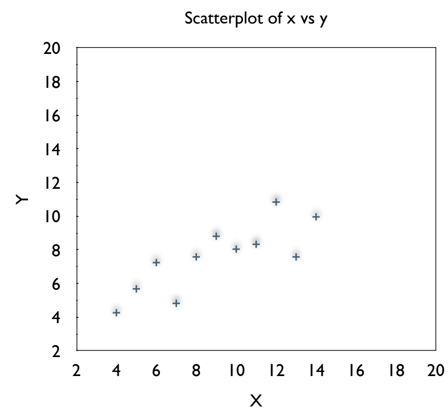
Scatterplots

- 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



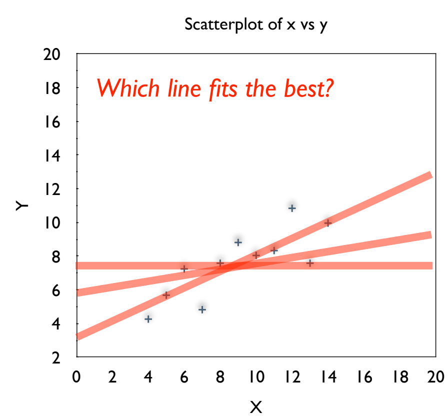
Linear Regression

- 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"

Linear Regression

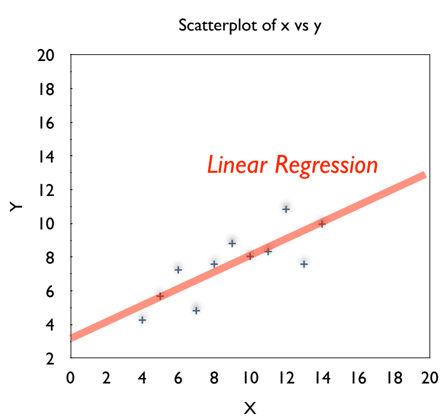
Equation:
$y = 3.0 + 0.5x$

Correlation
$r_{x,y} = 0.816$



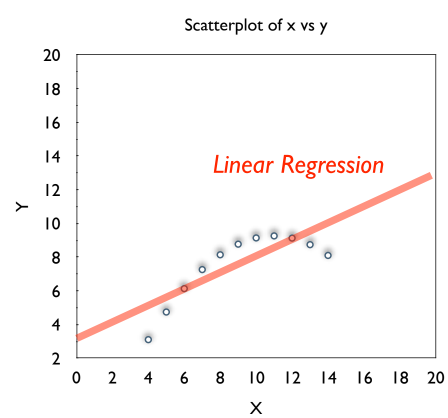
Anscombe's Quartet I

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



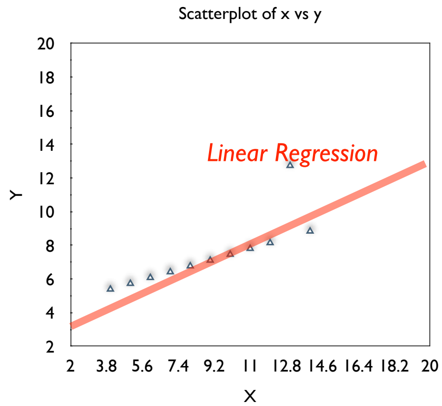
Anscombe's Quartet II

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



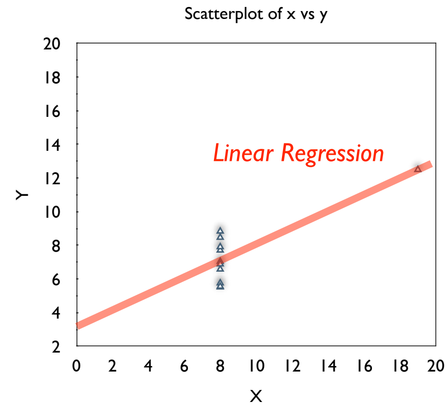
Anscombe's Quartet III

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

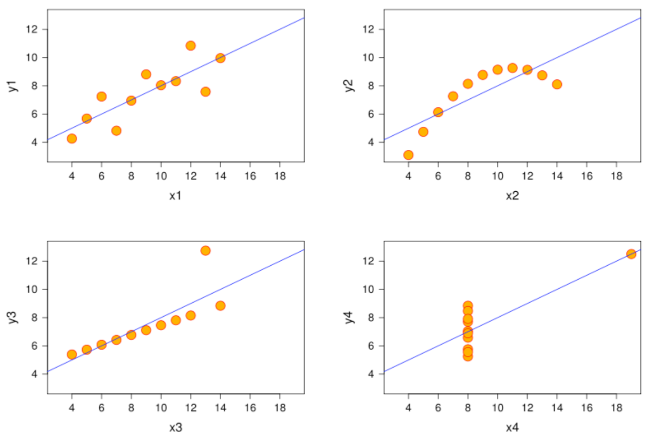


Anscombe's Quartet IV

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 Series 1-4



Anscombe's Quartet Summary

- The 4 series of data, though very different, have identical linear regression equations and identical correlations
- Each series has a Quantitative correlation, but it's clear (visually) that the relationships are Qualitatively different
- Each series should probably be handled differently, through techniques such as:
  - Trimmed Least Squares
  - Robust regression
- Graph Your Data!

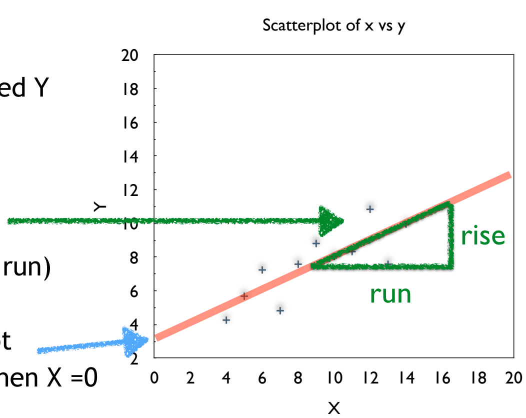
Linear Regression Equation

Y' = a + bX

Y' = predicted Y
X = actual X

b = slope
DY/DX
(rise over run)

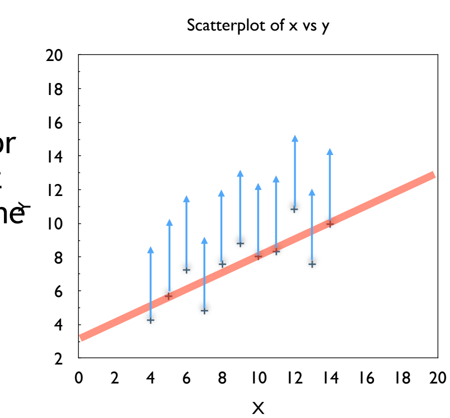
a = intercept
Y value when X =0



Intercept

Y' = a + bX
a = intercept

If all Ys go up (or down) intercept changes the same amount.

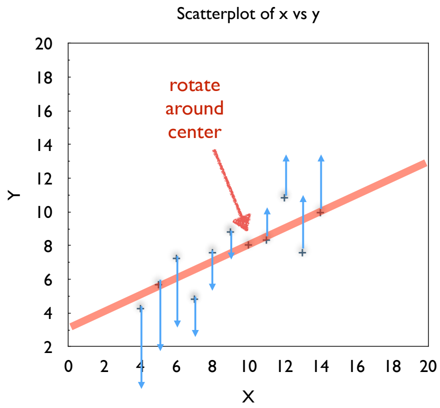


Slope

Y' = a + bX

b = slope

If points rotate around center, slope changes



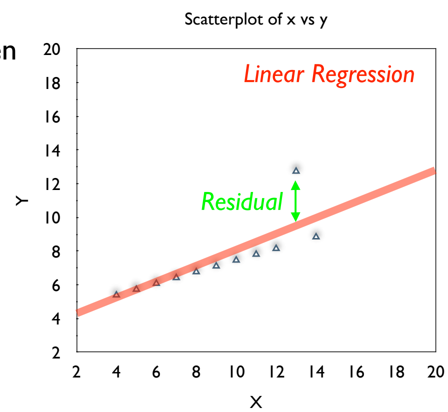
Residuals in Linear Regression

- X : independent variable
- Y : dependent variable
- Model: predict Y from X
- Y' : (Y prime) = predicted Y
- Y' = a + bX
- Prediction is (usually) incorrect. 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

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

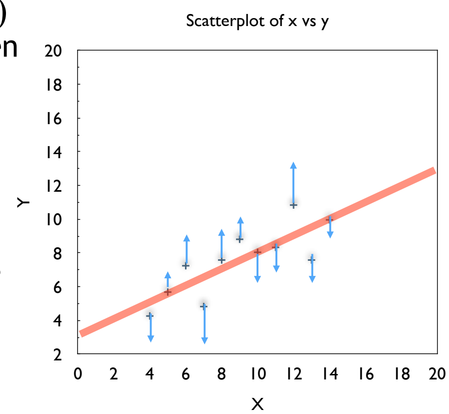
Graphically it is equal to how far away (vertically) a point is from the linear regression line



Residuals and Error

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

Error is lower when points are closer to 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%)

R^2

- Terminology
  - Coefficient of ***Determination***
  - ***Explained*** Variance
- Correlation: not causation

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 (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
Slope	meaningless	R^2
Transitive?	no	yes, R

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

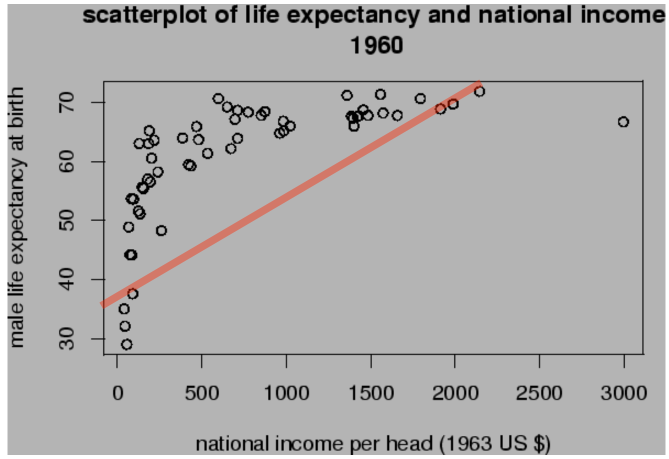
Correlation : Issues

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

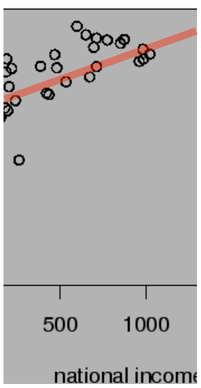
Non-linearity

- Linear regression & Correlation assume a linear relationship between X and Y
- Are most real-world relationships linear?
- Examples of non-linearity
- Solutions:
  - Intentionally restrict range of X
  - Rank variables then use Spearman's Rho
  - Transform variables (log, root, square, cube, etc.)
  - Use higher-order (polynomial) curve fitting, such as $Y = a + bX + cX^2 + dX^3$ ...

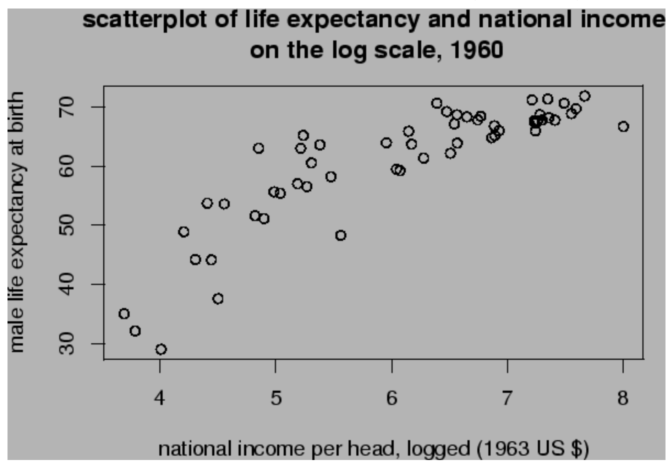
Life expectancy / national income



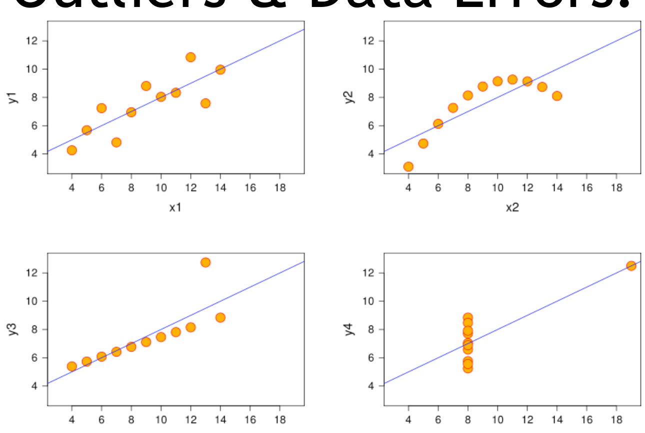
restrict range of X



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

- Sleeping with your shoes on is correlated with waking up with a headache
- Ice cream sales are positively related to increase in drowning

Shrinkage

- Least-squares regression attempts to fit the data set presented to it by reducing the observed residuals.
- This data set contains random errors.
- Thus, the parameters (equations) estimated for the linear regression line (and correlation coefficient) and residuals usually be higher than would be found in a separate data set.
- This reduction is called “Shrinkage”
- Cross-validation is best way to deal with it

Cross Validation

- Step 1: With a given data set, compute the linear regression line that fits this data.
- Step 2: Apply this linear regression equation to a different data set.
- Step 3: Calculate the observed error in step 2. This is typically higher than seen in step 1, and a much better measure of fit.

- Note: sometimes you may artificially “create” two data sets by splitting a single data set in half.

Hypothesis Testing

- All parameters (equations) we estimate from 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 and Clinical Significance

- These two terms are often confused and have very different meanings

- Statistical Significance: changes in DV are very unlikely to have been the result of random effects or chance. Often expressed as a P value ($p < .01$, or less than 1% chance to see these effects under H0)

- Clinical Significance : changes in DV are large enough to matter; the change was not trivial. If we accept H1, the conclusion is that H1’s effect size is important.
  - depends on context. often evaluated in terms of cost/benefit or risk/benefit tradeoff

Significance

- Example 1:
  - Two dice, Roll each once
  - Results: get a 3 and a 5
- Example B:
  - Two dice, Roll each 100 times
  - Results: Die A = 3.0, Die B = 3.10
- Example C:
  - Two dice, Roll each 100 times
  - Results: Die A = 3.0, Die B = 5.0

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"

- Effect Size (D)
- P-value

- Probability of Type I error (α)
- Probability of Type II error (β)

Review : Is race "real"?

- Phenomenology
- Pre-DNA views
- Post-DNA views

Pre-DNA views

- 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

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)

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

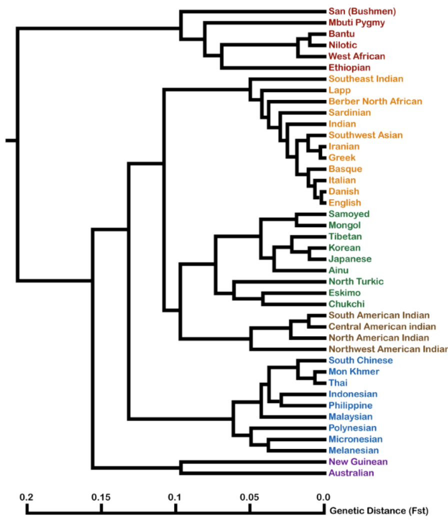
Australian and Africans are most genetically different

Post-DNA views

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

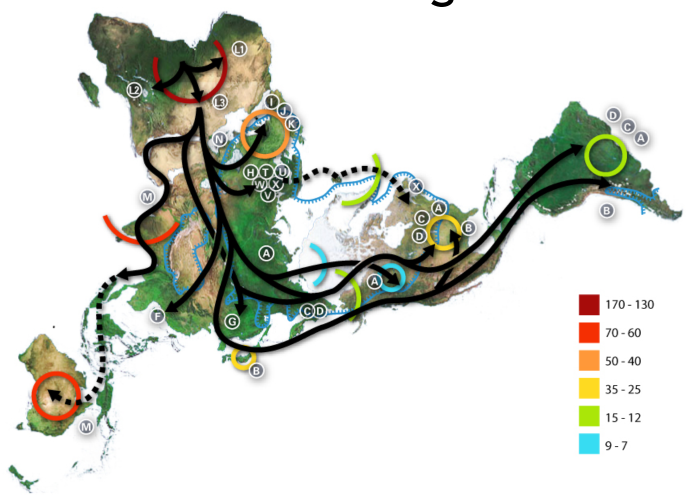
Genetic Differenc

Sub-Saharan African
Indo-European
East Asian
Native American
South Asian
Aboriginal



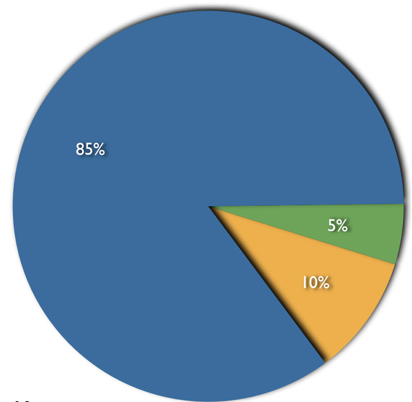
Fst = % of subpopulation variance

Prehistorical Migration



Variance: Genetic Variation

Within local populations
Within “race”
Between “race”



For example:

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