Review and Introduction to Biostatistics

Advanced Biostatistics

Dean Adams

Lecture 1

EEOB 590C

Lectures: Mondays 11:00 – 12:30, 145 Bessey Hall

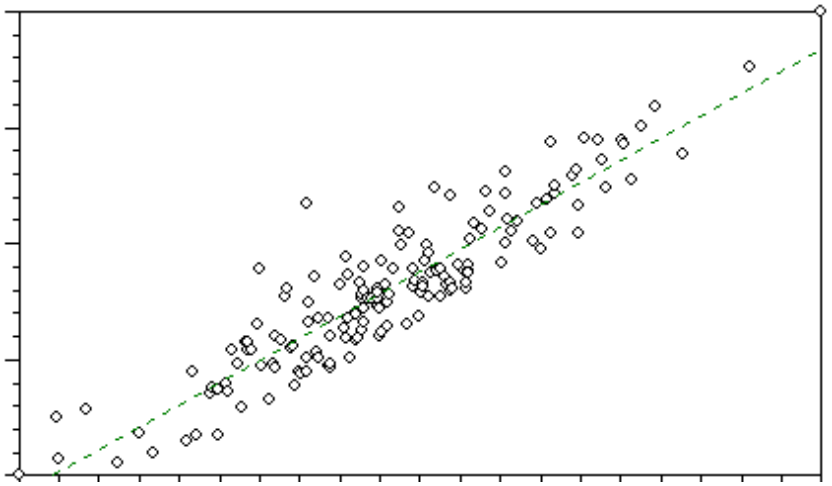
Lab: Mondays 12:30 – 1:30, 145 Bessey Hall

Readings: no assigned readings, but 3 references are useful

- Legendre, P., and L. Legendre. 2012. *Numerical Ecology. 3rd* edition. Elsevier Science, Amsterdam.
- Manly, B. F. J., and J.A. Navarro Alberto 2017. *Multivariate Statistical Methods: A Primer. 4th edition*. Chapman & Hall, London.
- Sokal, R. R. and F. J. Rohlf. 2012. *Biometry*. 4th Edition. Freeman Associates, San Francisco.

All data are dots in space

Axes of the space change depending on data type (e.g., categorical vs. continuous data)



Statistical methods represent a set of tools for visualizing and describing patterns of dots in data space

Secret of statistics: knowing which method to use for which data type (choice is combination of hypotheses and data types)

Independent Variable: specifies the hypothesis; a *predictor* for other variables (e.g., sex, age). (*X-matrix*)

Dependent Variable: the *response* variable; its variation depends other variables. This is the 'data' (*Y-matrix*)

MANY ways of classifying statistical methods, I prefer…

Inferential Statistics: test for specific patterns in data using independent variables to generate hypotheses (Y vs. X)

Exploratory Statistics: describe patterns in data without independent variables or specific hypotheses (patterns in Y)

Some examples:

Inferential Methods	Exploratory Methods
T-test	Principal Components (PCA)
F-test	Principal Coordinates (PCoA)
ANOVA/MANOVA	Multi-Dimensional Scaling (MDS)
Chi-square	UPGMA
Regression	K-Means clustering
Some methods are 'hybrids' (e.g., CVA and Correspondence Analysis)	

Parametric Methods: Use parameters estimated from the data to test hypotheses. Parameters compared to theoretical distributions to assess significance (can be powerful approaches when assumptions are met).

Maximum Likelihood Methods: Obtain likelihood of the data given the model, and compare the fit of models using likelihood or some derivative of it (e.g., AIC). Approach obtains maximum likelihood estimates of parameters of the model. Can be very powerful. (Estimates from parametric methods such as least-squares frequently converge on maximum likelihood estimates of those parameters under large samples and when assumptions are met).

Bayesian Methods: Use data, model, and prior knowledge to assess hypotheses and obtain posterior probabilities of parameters. Can be a powerful approach, but also can be context dependent (e.g., different priors can lead to differing conclusions)

Nonparametric Methods: Depend on the distribution of the variates (the data); some use ranks of the data, others generate their own distribution from the data (e.g,. Randomization). Useful when data don't meet parametric assumptions.

Test for specific patterns in data using independent variables to generate hypotheses (Y vs. X)

Procedure:

-Obtain an estimate of the parameter(s) (i.e., the observed test value)

-Compare observed test value to expected values **under the null hypothesis** (or from the likelihood under the null hypothesis)

-Determine whether observed value is something interesting (i.e., it is unexpected relative to expected under the null hypothesis? Typically $P < 0.05$)

Where do the expected values come from?

Obtained by:

1: The type of data examined

2: A model of how the process works

Often a theoretical distribution can be derived

Example 1: How likely is it that you get 5 heads in a row?

-**Type of data**: Binary: head/tail (1/0)

-**Verbal Model**: a 'fair' coin (i.e., chance of heads = 0.5). Also, each event independent (flipping a head the 1st toss does not alter chance of heads on 2nd toss)

-Likelihood of 5 tails in a row is: 0.5*0.5*0.5*0.5*0.5 = 0.03125

Value is small, so result of 5 heads in a row is not likely

Example 2: Are male and female mandible lengths the same?

-**Type of data**: Continuous

-**Verbal Model**: Assess trend in male vs. female jackal mandible length (i.e., compare average values) --- -Assume independence of individuals measured

-Calculate test value $D_{obs} = \left(\overline{Y}_m - \overline{Y}_f\right)$ and generate expected values under the null hypothesis of no difference between males and females (i.e., no pattern)*

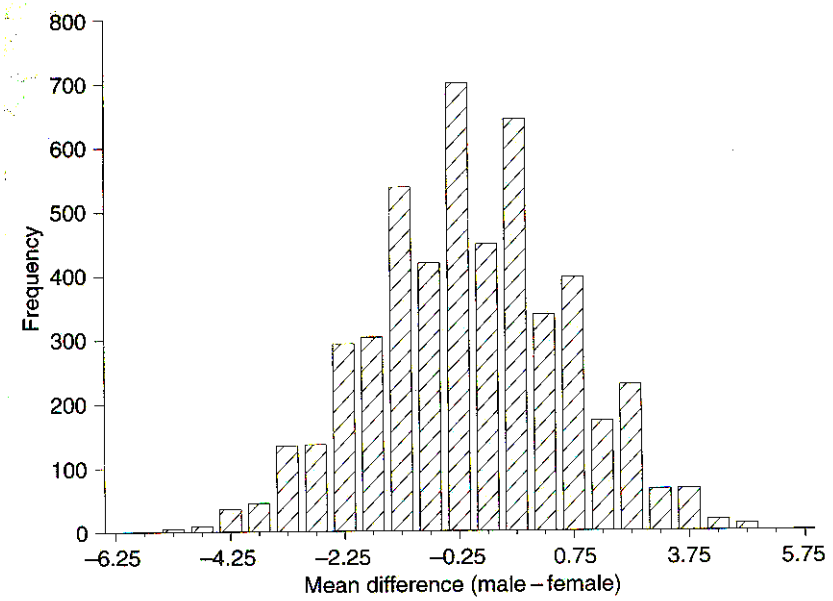
-Shuffle specimens among M/F groups, calculate D_{rand} and repeat many times, compare D_{obs} to distribution of D_{rand}

-Males: 120,107,110,116, 114, 111, 113, 117, 114, 112

-Females: 110, 111, 107, 108, 110, 105, 107, 106, 111, 111

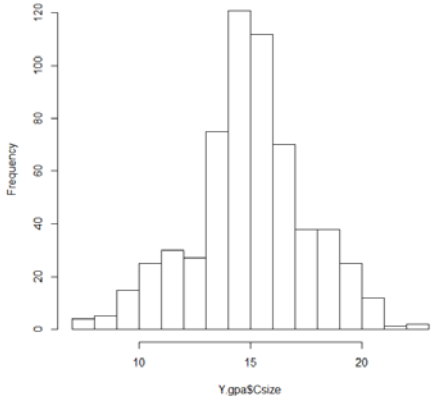
-D_{obs} = 4.8

-This is a randomization test

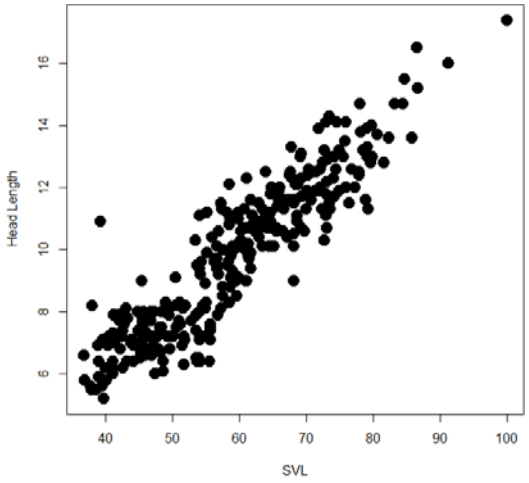


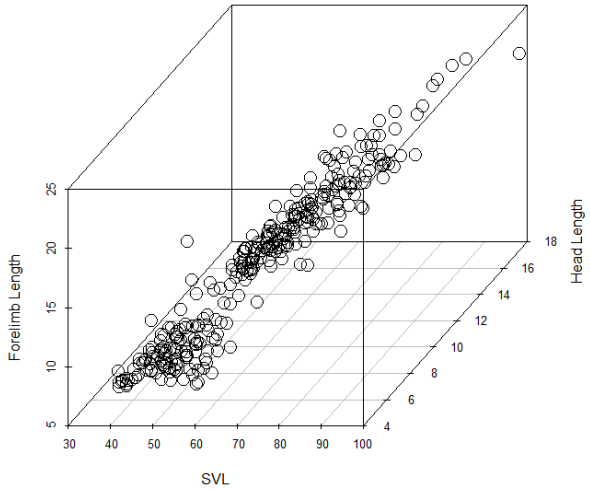
The null hypothesis and expected values derived from it are extremely important for designing proper permutation tests

Univariate statistics: Assess variation in single Y (obtain scalar result)



Multivariate statistics: Assess variation in multiple Y simultaneously

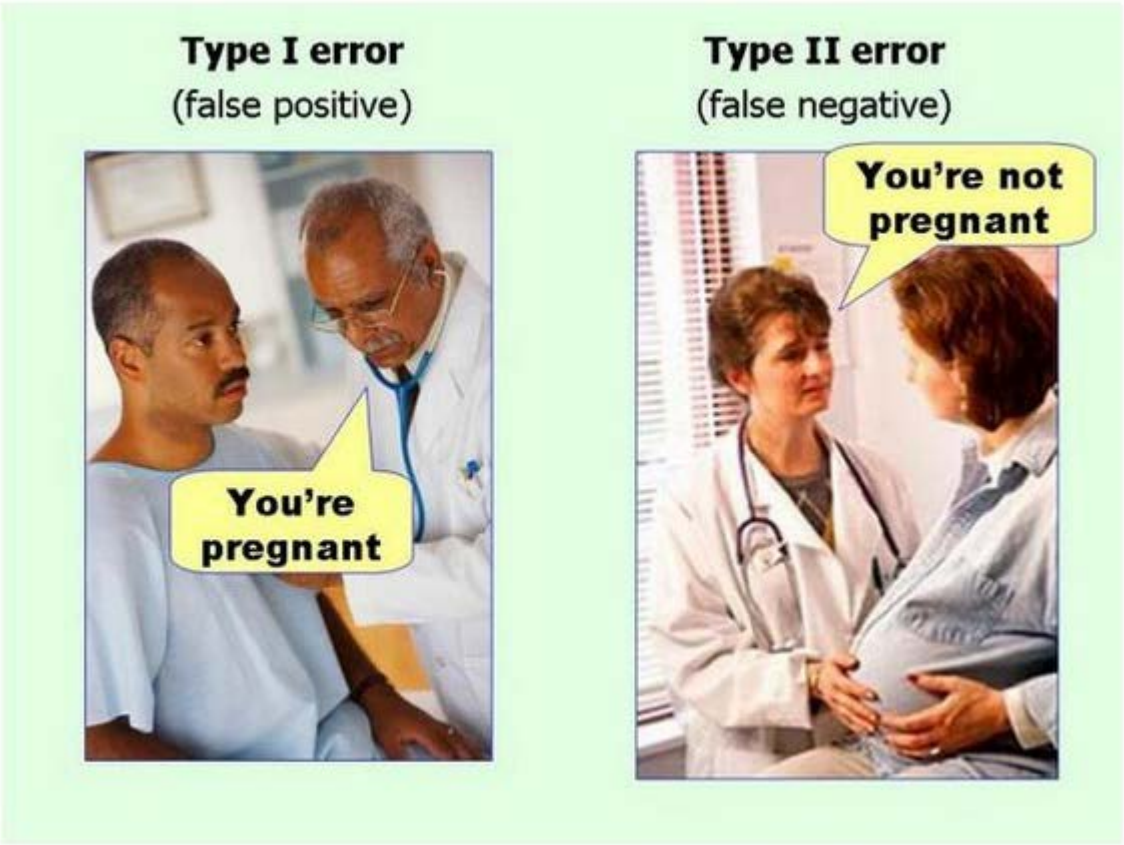




Multivariate methods are mathematical generalizations of univariate (ACTUALLY, univariate methods are special cases of multivariate!)

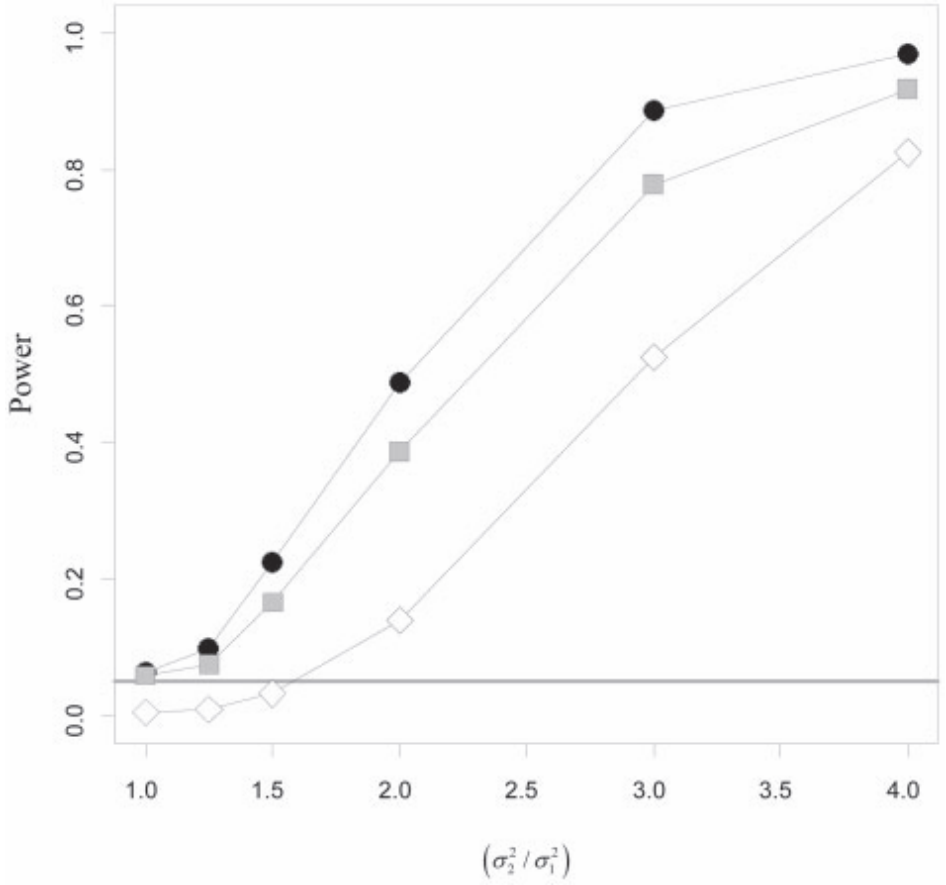
There is error associated with hypothesis testing

Type I error (α): reject H_0 when true (false positives)
Type II error (β): not rejecting H_0 when false (false negatives)



Power (1-β): Ability to detect significant effect when it is present (function of effect size, *N*, σ^2)

Power of test can be empirically determined in many instances

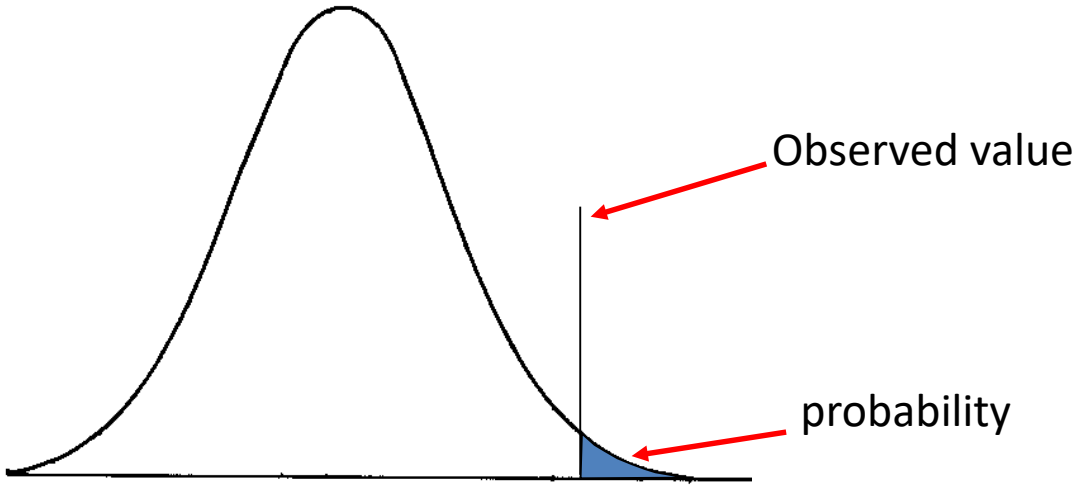


Type I error and power for several evolutionary test statistics. (Adams. 2013. *Syst. Biol.*)

Parametric statistics: estimate statistical parameters from data, and compare to a theoretical distribution of these parameters

Significance based upon how ‘extreme’ the observed value is relative to the distribution of values (under the null hypothesis of no pattern)

Different distributions used for different statistical parameters and tests



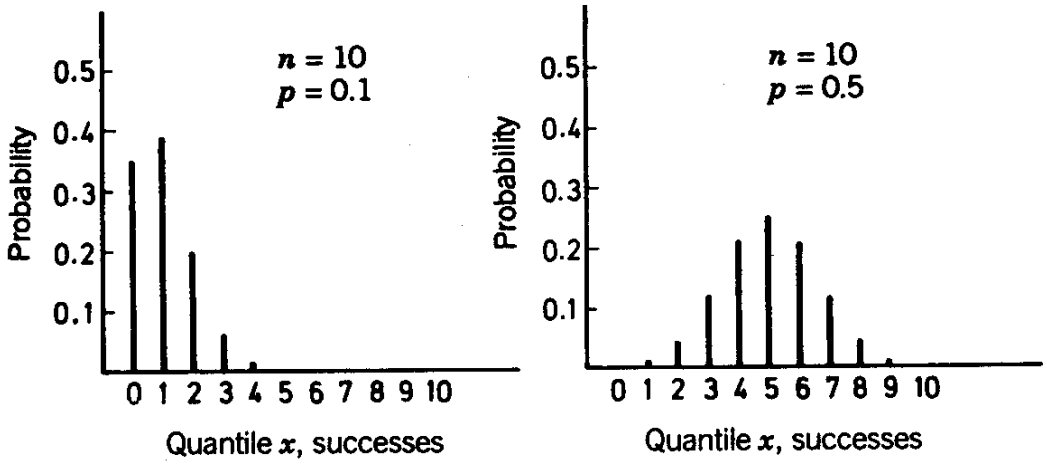
NOTE: theoretical aspects and general properties of these distributions is a HUGE and interesting area of statistical research (consult Stats. Dept. for details)

Distribution for binary events, calculate probabilities directly

Determine probability of obtaining x outcomes in *n* total tries

$$prob = \binom{n}{x} p^x q^{n-x}$$

n is the total # events, *x* is the # successes, and *p* & *q* are the probability of success and failure.



Example: prob. of obtaining 7 heads in 10 tries

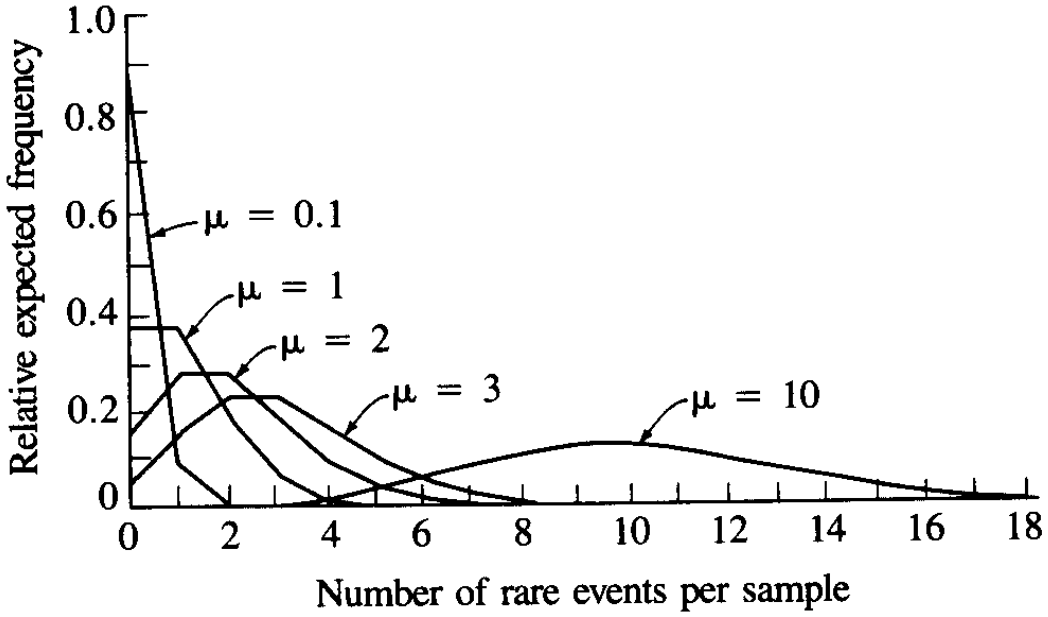
The 'coin-toss' distribution

$$prob = \binom{10!}{7!3!} 0.5^7 0.5^3 = 0.117$$

Models probability of rare, and independent events

For Poisson: $\mu_X = \sigma_X^2$

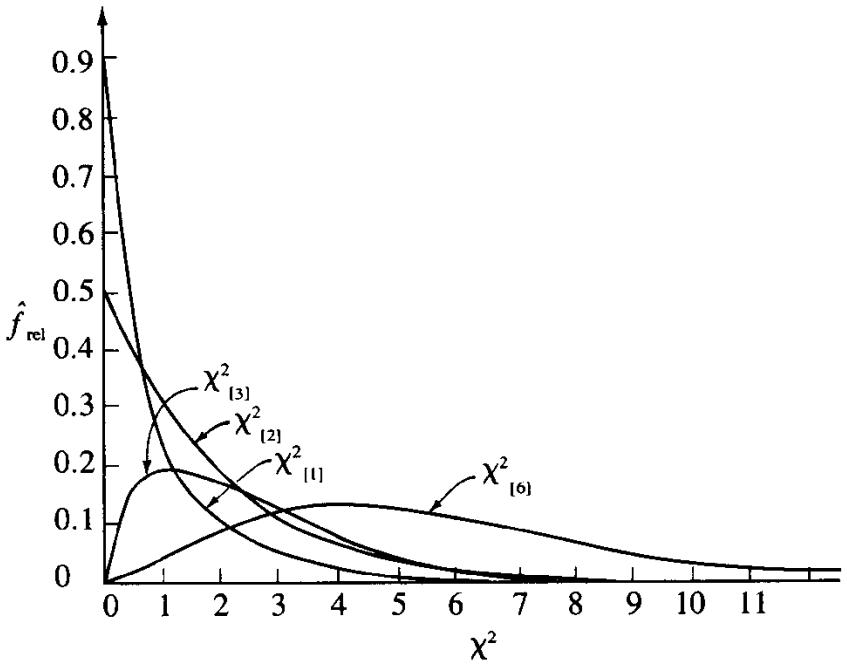
Multiple Poisson distributions exist (for different means)



Molecular mutations are typically modeled as Poisson

Ranges from 0 to + ∞: Multiple χ^2 distributions exist (for different df)

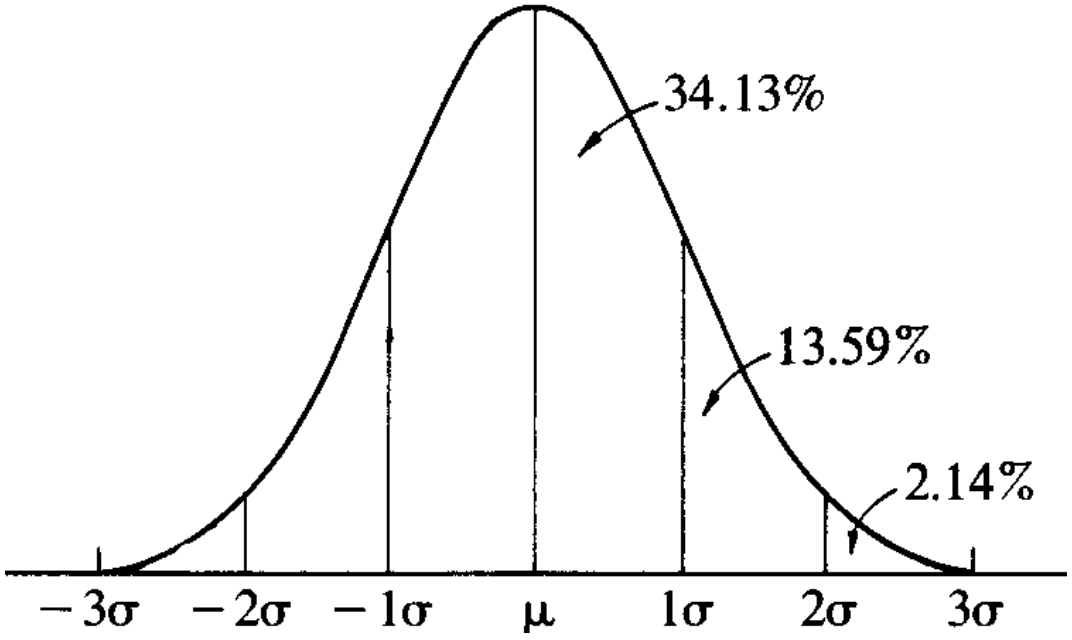
Many statistics (particularly for categorical data) are compared to χ^2 distributions (χ^2, G-test, Fisher's exact test, etc.)



The ‘bell curve’: data are symmetrical about the mean

Range: $-\infty$ to $+\infty$: $\pm 1\sigma$ = 68% of data, $\pm 2\sigma$ = 95% of data, $\pm 3\sigma$ = 99% of the data

T-distribution is a set of approximate normal distributions for finite sample sizes (used to compare to groups)

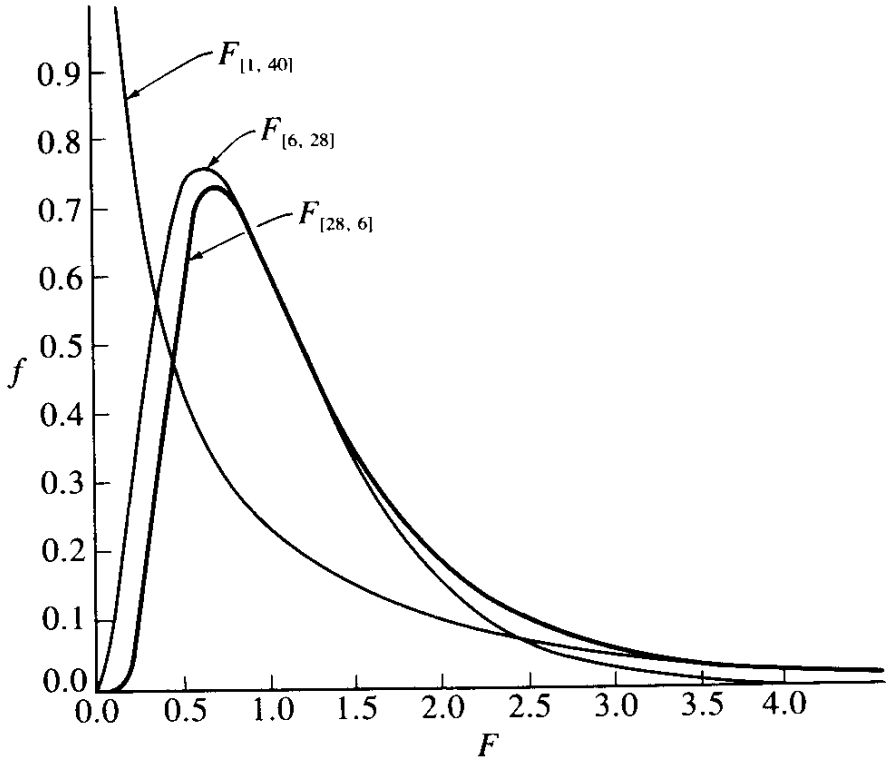


Models the ratio of variances: range: 0 to +∞

One of the most commonly used distributions (used for GLM)

Is the combination of 2 T-distributions (and has 2 df)

Multiple F-distributions for various df



Sometimes, our data and hypotheses match a type of analysis (e.g., ANOVA), but violate the assumptions (e.g., normality)

One solution: transform the data so that they more closely match the assumptions of the test (meeting the test assumptions is important, so that results can be attributed to true differences in the data vs. violations of the properties of the test)

Some common transformations for biological data are:

-**Log transformation**: when mean is positively correlated with variance (e.g., linear morphological measurements)

-**Square-root transformation**: used for frequency counts (note: need to add 0.5 to ALL counts if zeros exist in data)

-**Arcsine transformation**: used for percentage/proportion data

$$\theta = \arcsin\sqrt{p}$$

There are MANY other possible transformations

Summarize a sample of data points using parameters, and the measure of central tendency is one VERY important concept

Arithmetic Mean: average of the values: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

Weighted Mean: average of the weighted values: $\bar{Y}_W = \sum_{i=1}^{n} w_i Y_i \Big/ \sum w_i$

Geometric Mean: used for log-data: $\bar{Y}_{GM} = \sqrt[n]{Y_1 Y_2 \cdots Y_n}$ $\bar{Y}_{GM} = e^{\left(\frac{1}{n}\sum_{i=1}^{n} \ln Y_i\right)}$

Harmonic Mean: used for rates/speed: $\bar{Y}_H^{-1} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{Y_i}$

Median: value having an equal number of lower and higher-valued items (the 'middle' or 50th percentile value)

Mode: value found in highest frequency in the data

Moment Statistics: deviations around the mean, raised to powers

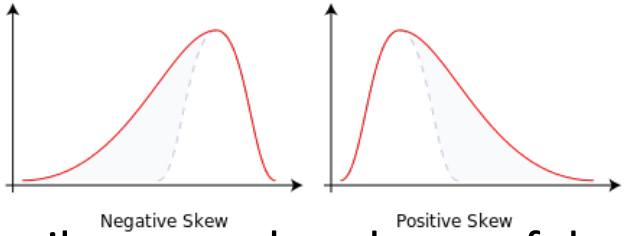
1st moment: sum of deviates (equals zero)

$$M^1 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^1 = 0$$

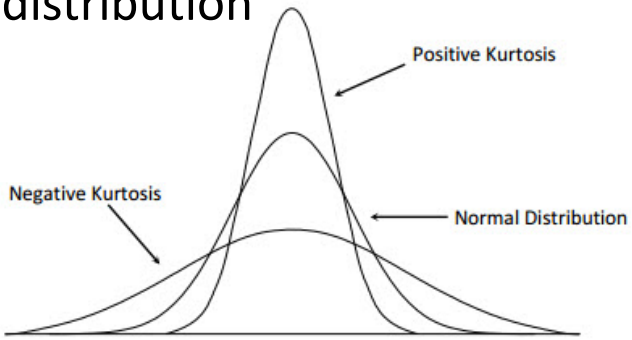
2nd moment (variance)*: sum of squared deviates, measures dispersion around mean

$$s^2 = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y - \bar{Y}\right)^2$$

3rd moment (skewness): describes the direction (skew) of the distribution



4th moment (kurtosis): describes overabundance of deviates at tails (platykurtic) or at center (leptokurtic) of distribution



*Note: for most statistical analyses, variance [of a sample] is calculated using (n-1)

Standard deviation: another measure of dispersion around the mean: square-root of variance (note that n-1 used for the sample)

$$s = \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y - \bar{Y}\right)^2}$$

Degrees of Freedom (df): describes the number of parameters in data that are free to vary after we've calculated some parameter (e.g., if you know the mean and all but 1 value from the data, you can figure out the remaining variate)

Becomes important when determining whether your sample size is sufficient for a particular test (each test has associated df based on how many parameters are estimated)

Basic roadmap depends on type of X & Y variables

	Categorical X	Continuous X
Categorical Y	Contingency tables	Logistic Regression
Continuous Y	**ANOVA**	**Regression** (correlation)

← Log-Linear Models

← General Linear Models (GLM)

The vast majority of biological analyses lie somewhere in this simple roadmap

Main categories of inferential models: General Linear Models (GLM) and Log-linear models

Maximum likelihood (ML) used to calculate all parameters

GLM (ANOVA, regression): Used when Y is continuous. Fitting procedure is Least Squares (minimize the sum of squares deviations). This is equivalent to ML when error is normally distributed

Log-Linear Models (logistic regression, contingency tables): Used when Y is categorical. Called log-linear because ML estimate of logs of variables is linear

ANOVA (Analysis of Variance): Y~X

-X is categorical (groups), Y is continuous

H_0: no relationship between X & Y (i.e., no difference among groups)

-Compare variation between groups to variation within groups (e.g., are males & female different in height?)

Model: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ (μ is grand mean, α_i is i^{th} group mean, and ε_{ij} is error)

Multiple factors can be tested simultaneously (e.g., sex & species)

Assumptions: independent ε_{ij}, normally distributed ε_{ij}, & homoscedastic (equal) variance

Regression: Y~X

-X & Y are continuous

H_0 : no relationship between X & Y

-Fit a line or curve to data that minimizes [vertical] deviations in the bivariate plot (e.g., relationship of age vs. height)

Model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (β_0 is intercept, β_i is slope, and ε_i is error)

Multiple linear and/or curvilinear X-variables may be included

Assumptions: independent X_i (measured without error), independent normally distributed ε_i, & homoscedastic variance along regression line

Logistic Regression: Y~X

-X is continuous, Y is binary

H_0 : no relationship between X & Y

Y is derived from a categorical variable (e.g., % males)

Model: $$\ln\left[\frac{\hat{p}}{1-\hat{p}}\right] = \beta_0 + \beta_1 X_i + \varepsilon_i$$ (β_0 is intercept, β_i is slope, and ε_i is error)

Best regression line is fit using ML

-Note: above formulation is a linear regression of the *logits* of the proportions

R x C Tests: Y~X

-X is categorical, Y is categorical

H_0 : no frequency differences among groups (rows & cols independent)

Chi-square tests and G-tests fit this type of model

Common in medical studies: Is smoking associated with cancer rates?

	Cancer	No Cancer
Smokers	#A	#B
Non-Smokers	#C	#D

Includes more X & Y possibilities

	1 Categorical X	>1 Categorical X	1 Continuous X	>1 Continuous X	Both
1 Categorical Y	R x C Tests	Multi-way tables	Logistic Regression	'multi-factor' logistic regression	
1 Continuous Y	ANOVA	Factorial ANOVA	Regression	Multiple Regression	ANCOVA
>1 Continuous Y	MANOVA	Factorial MANOVA	Multivariate Regression	Multivariate Multiple Regression	MANCOVA

ALL GLM models are derived from SAME equation, so conceptual leap to multivariate statistics is minimal

Note: 'multivariate' log-linear models exist, but are infrequently used in biology

Compare single sample to a known value, or compare 2 samples

Determine whether means are significantly different

H_0 : no difference in means

For a single sample calculate: $$t = \frac{\bar{Y} - \mu}{s_{\bar{Y}}}$$

For two samples calculate: $$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

compare to *t*-distribution with pooled df (or do resampling)

Other variations exist (e.g., for paired data)

*Note: 2-sample t-test (n1=n2) yields equivalent results to 2-sample ANOVA

- Compare male and female class loads (# course credits)

$$\mathbf{X} = [m \quad m \quad m \quad m \quad m \quad f \quad f \quad f \quad f \quad f]$$
$$\mathbf{Y} = [5 \quad 4 \quad 4 \quad 4 \quad 3 \quad 7 \quad 5 \quad 6 \quad 6 \quad 6]$$

- Obtain t-statistic:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{6-4}{.7071\sqrt{\frac{1}{5} + \frac{1}{5}}} = \frac{2}{.44721} = 4.47$$

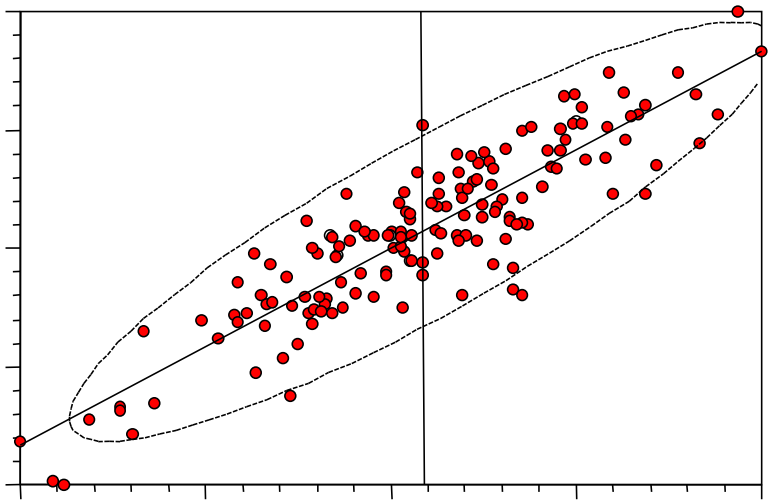
```
> t.test(y~x)

data:  y by x
t = 4.4721, df = 8, p-value = 0.002077
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9687236 3.0312764
sample estimates:
mean in group f mean in group m
              6               4
```

Determine amount of association (covariation) between two variables (H_o: no association)

Range: -1 to +1 (more extreme values = higher correlation)

$$r_{ij} = \frac{\text{cov}_{ij}}{s_i s_j} = \frac{\frac{1}{n-1}\sum_{i,j=1}^{n}\left(Y_i - \bar{Y}_i\right)\left(Y_j - \bar{Y}_j\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2 \frac{1}{n-1}\sum_{j=1}^{n}\left(Y_j - \bar{Y}_j\right)^2}} = \frac{\sum_{i,j=1}^{n}\left(Y_i - \bar{Y}_i\right)\left(Y_j - \bar{Y}_j\right)}{\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2 \sum_{j=1}^{n}\left(Y_j - \bar{Y}_j\right)^2}}$$

Measures ‘tightness’ of scatter of one variable relative to the other



Assess significance by converting r to t (or resampling)

Numerator of r_{ij} is covariance: describes deviations in 1 variable as they change with deviations in another variable (similar to variance, but for 2 variables)

$$\text{var}_i = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2}{N-1} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)\left(Y_i - \bar{Y}_i\right)}{N-1} \quad \text{vs.} \quad \text{cov}_{ij} = \frac{\sum_{i,j=1}^{n}\left(Y_i - \bar{Y}_i\right)\left(Y_j - \bar{Y}_j\right)}{N-1}$$

Very important concept*

Can also think of correlation as angle between vectors i and j in variable space: the tighter the angle, the higher the correlation

Thus, $r = \cos\theta$ (1 of MANY ways to visualize correlations)

*NOTE: N-1 drops out of numerator and denominator of r

Often used to summarize categorical data from contingency tables

Tests for independence of values in cells (i.e. between rows and columns)

For 2 x 2 table calculate: $$X^2 = \frac{\sum (O - E)^2}{E}$$

O = observed data in each cell, & E = expected data

Compare obs. X^2 to X^2-distribution with (n-1) df [frequently = (r-1)(c-1)]

Other derivations exist for particular data types, but this is general concept