
A Brief (very brief) Overview of Biostatistics

Jody Kreiman, PhD
Bureau of Glottal Affairs

What We'll Cover

- Fundamentals of measurement
 - Parametric versus nonparametric tests
 - Descriptive versus inferential statistics
 - Common tests for comparing two or more groups
 - Correlation and regression
-

What We Won't Cover

- Most nonparametric tests
 - Measures of agreement
 - Multivariate analysis
 - Statistics and clinical trials
 - Anything in depth
-

Why You Should Care

- Without knowledge of statistics, you are lost.
- It's on the test.



I: Variables

- Independent versus dependent variables
 - Levels of measurement
 - Kinds of statistics
-

Levels of Measurement

- The kind of statistic that is appropriate depends on the way the dependent variable has been measured.
 - Four levels of measurement:
 - Categorical/nominal (special case: dichotomous)
 - Ordinal
 - Interval
 - Ratio
-

II. What Are Statistics?

- Methods for organizing, analyzing, and interpreting numerical data
 - Descriptive statistics: Organize and summarize data
 - Inferential statistics: Used to make an inference, on the basis of data, about the (non)existence of a relationship between the independent and dependent variables
-

Kinds of Statistics

- When data are measured at the categorical or ordinal level, nonparametric statistical tests are appropriate.
 - Unfortunately, time prohibits much discussion of this important class of statistics.
 - When data are interval or ratio, parametric tests are usually the correct choice (depending on the assumptions required by the test).
-

Kinds of Statistics

- It is always possible to “downsample” interval or ratio data to apply nonparametric tests.
 - It is sometimes possible to “upsample” ordinal or categorical data (e.g., logistic regression), but that is beyond the scope of this lecture.
 - Decisions about levels of measurement require careful consideration when planning a study.
-

Kinds of Statistics

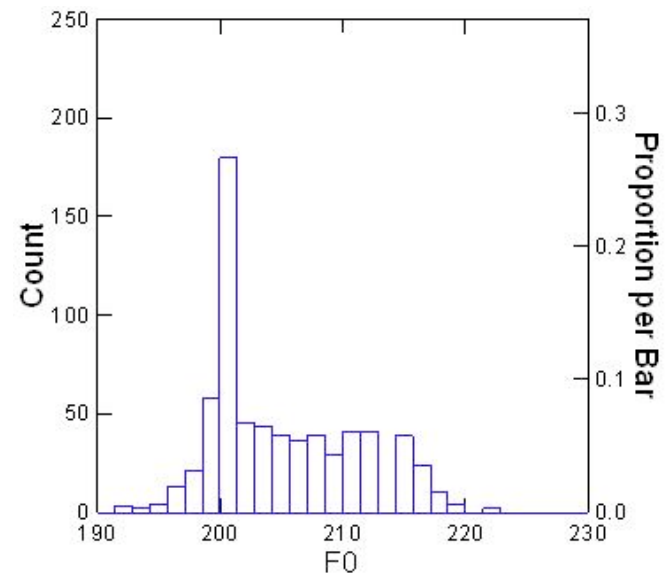
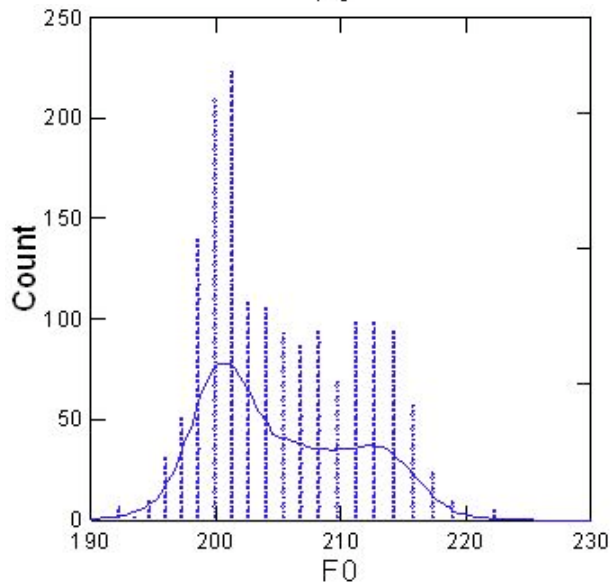
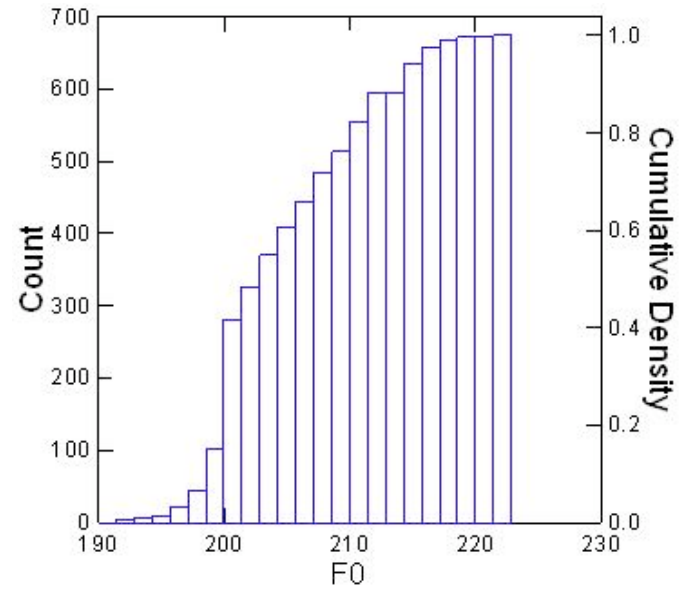
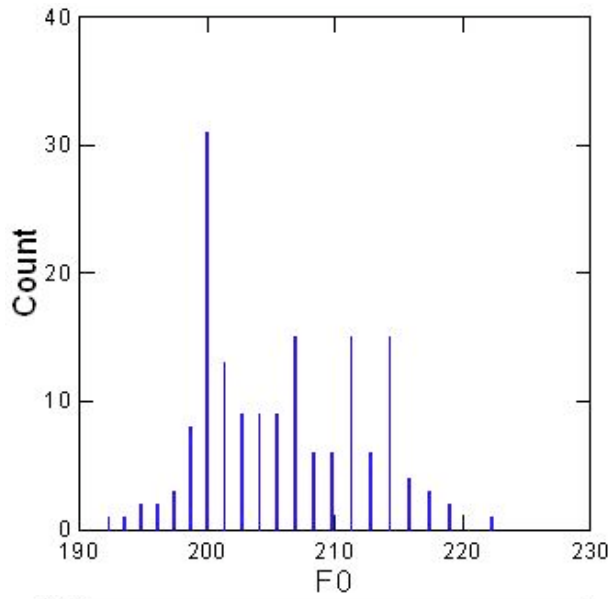
- Descriptive statistics
 - Inferential statistics
-

Descriptive Statistics

- “Data reduction:” Summarize data in compact form
 - Minimum
 - Maximum
 - Mean
 - Standard deviation
 - Range
 - Etc...
-

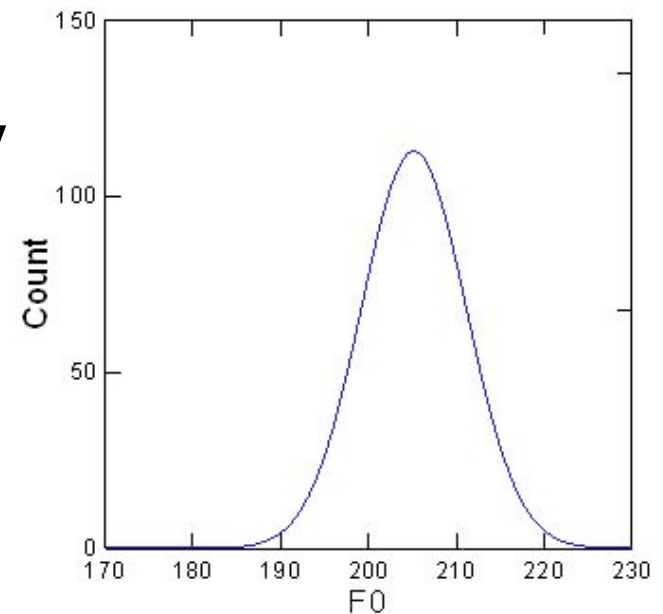
Frequency Distributions

- Description of data, versus theoretical distribution
 - Data can be plotted in various ways to show distribution
-



Theoretical Frequency Distributions

- There are lots, but we'll stick to one for now: the Normal Distribution
 - Described by a mean and a variance, about which more later
- The assumption of normality



III. Measures of Central Tendency

- Mean

- The average, equal to the sum of the observations divided by the number of observations ($\Sigma(x)/N$)

- Median

- The value that divides the frequency distribution in half

- Mode

- The value that occurs most often
 - There can be more than one—"multimodal" data.
-

▼ Descriptive Statistics

	F0
N of Cases	674
Minimum	192.308
Maximum	222.222
Range	29.915
Arithmetic Mean	205.154
Standard Error of Arithmetic Mean	0.229
Standard Deviation	5.950
Variance	35.400

Median = 204.08

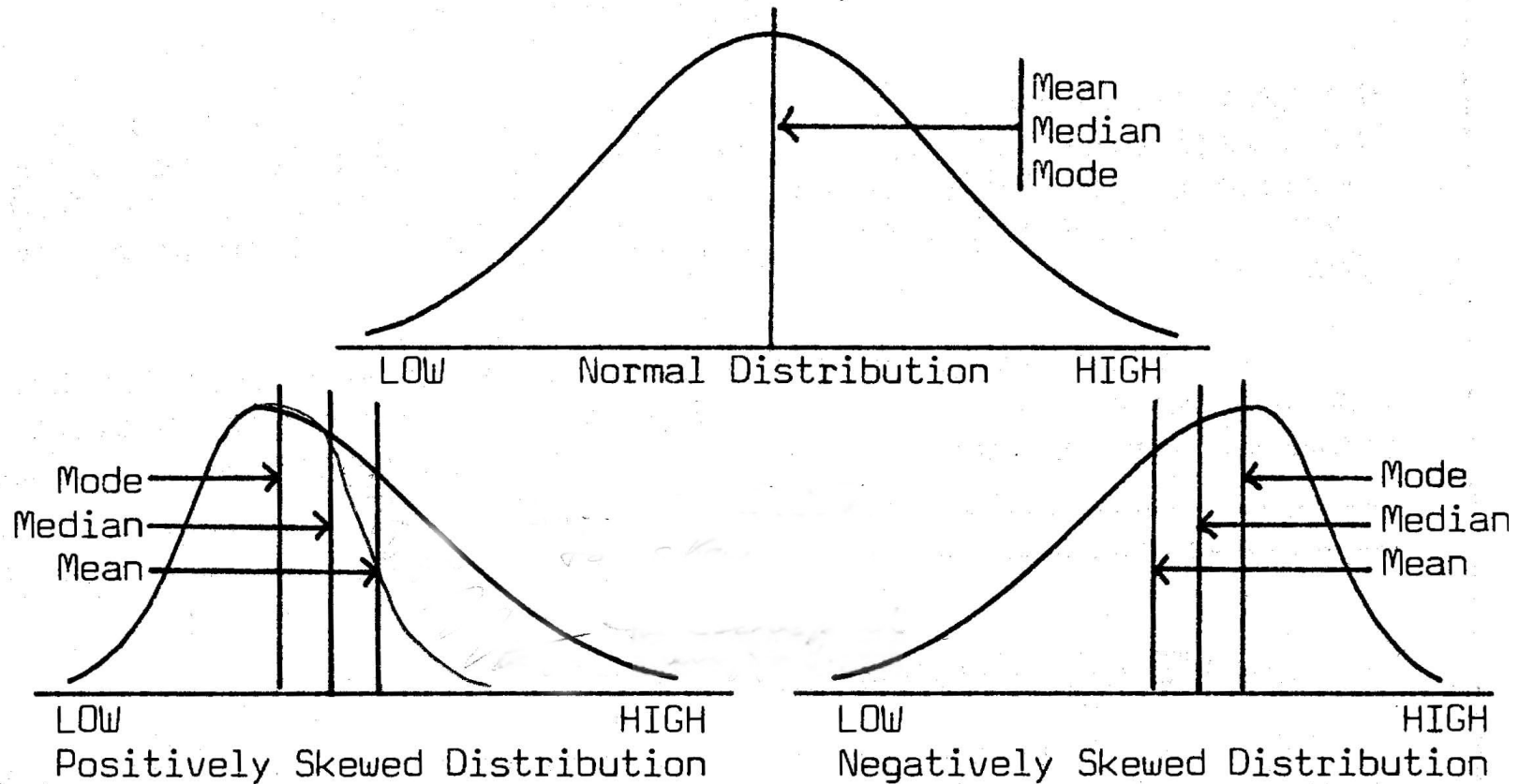
Mode = about 200.00

Which to Use?

- The mode is appropriate at any level of measurement.
 - The median is appropriate with ordinal, interval, or ratio data.
 - The mean is appropriate when data are measured at the interval or ratio level.
 - The relationship between measures depends on the frequency distribution.
 - When data are normally distributed, all values will be equal.
-

Mean, Median, and Mode

Figure 7: Relationship Between the Mean, Median, and Mode for Different Frequency Distributions



IV. Measures of Variability

- Range (largest score – smallest score)
 - Variance ($S^2 = \frac{\sum(x-M)^2}{N}$)
 - Standard deviation
 - Square root of the variance, so it's in the same units as the mean
 - In a normal distribution, 68.26% of scores fall within +/- 1 sd of the mean; 95.44% fall within +/- 2 sd of the mean.
 - Coefficient of variation = the standard deviation divided by the sample mean
-

Confidence Intervals

- Confidence intervals express the range in which the true value of a population parameter (as estimated by the population statistic) falls, with a high degree of confidence (usually 95% or 99%).
 - Example: For the F0 data in the previous slides, the mean = 205.15; the 95% CI = 204.70-205.60; the 99% CI = 204.56-205.75.
 - The range is narrow because N is large, so the estimate of the population mean is good.
-

V. Inferential Statistics: Logic

- Methods used to make inferences about the relationship between the dependent and independent variables in a population, based on a sample of observations from that population
-

Populations Versus Samples

- Experimenters normally use sample statistics as estimates of population parameters.
 - Population parameters are written with Greek letters; sample statistics with Latin letters.



Sampling Distributions

- Different samples drawn from a population will usually have different means.
 - In other words, sampling error causes sample statistics to deviate from population values.
 - Error is generally greater for smaller samples.
 - The distribution of sample means is called the *sampling distribution*.
 - The sampling distribution is approximately normal.
-

Standard Deviation Versus Standard Error

- The mean of the sampling distribution equals the population mean.
 - The standard deviation of the sampling distribution (also called the *standard error of the mean*) equals the population standard deviation / the square root of the sample size.
 - The standard error is an index of sampling error—an estimate of how much any sample can be expected to vary from the actual population value.
-

The Logic of Statistical Tests

- Hypothesis testing involves determining if differences in dependent variable measures are due to sampling error, or to a real relationship between independent and dependent measures.
 - Three basic steps:
 - Define the hypothesis
 - Select appropriate statistical test
 - Decide whether to accept or reject the hypothesis
-

Hypothesis Testing

- “If you have a hypothesis and I have another hypothesis, evidently one of them must be eliminated. The scientist seems to have no choice but to be either soft-headed or disputatious” (Platt, 1964, p. 350).
-

Accepting or Rejecting the Null Hypothesis

- The region of unlikely values is the level of significance, or α . Alpha (type I error) represents the likelihood of incorrectly rejecting the null hypothesis.
 - Type II error (β) is the probability of accepting the null hypothesis when it is actually false.
 - Beta is greatest when alpha is low, sample size is small, effects of independent variable are small, and/or sampling error is high.
-

Consequences of Statistical Decisions

Actual state of affairs

	Null hypothesis true	Null hypothesis false
Decision	Null hypothesis accepted Correct	Type II error
	Null hypothesis rejected Type I error	Correct ($1-\beta$ =power)

VI. Choosing a Statistical Test

- Choice of a statistical test depends on:
 - Level of measurement for the dependent and independent variables
 - Number of groups or dependent measures
 - The population parameter of interest (mean, variance, differences between means and/or variances)
-

Comparing Counts (Categorical Data): the χ^2 (Chi)-square test

- Single sample chi-square test: assesses the probability that the distribution of sample observations has been drawn from a hypothesized population distribution.
 - Example: Does self-control training improve classroom behavior? Teacher rates student behavior; outcome (the observed frequencies) compared to distribution of behavior ratings for entire school (the expected frequencies).
-

Chi-square test for two samples

- “Contingency table” analysis, used to test the probability that obtained sample frequencies equal those that would be found if there were no relationship between the independent and dependent variables.
-

Example (de Casso et al., 2008)

- Swallowing ability (success) in total laryngectomies who did or did not have radiotherapy
- Swallowing was significantly better in the group that did not receive radiotherapy ($p < .05$).

	Surgery only	Surgery + RT
Swallow		
Solid	77%	40%
Soft	8%	22%
Liquid	15%	38%

Comparing Group Means

- Choice of test can depend on number of groups
 - T-tests
 - Analysis of variance (ANOVA)
 - Calculating 95% or 99% confidence intervals
-

T-tests

- One sample t-test

- Compares sample value to a hypothesized exact value
 - Example: Students with hyperactivity receive self-control training and are assessed on a measure of behavior. The experimenter hypothesizes that their average score after training will be exactly equal to the average score for all students in the school.
-

T-tests

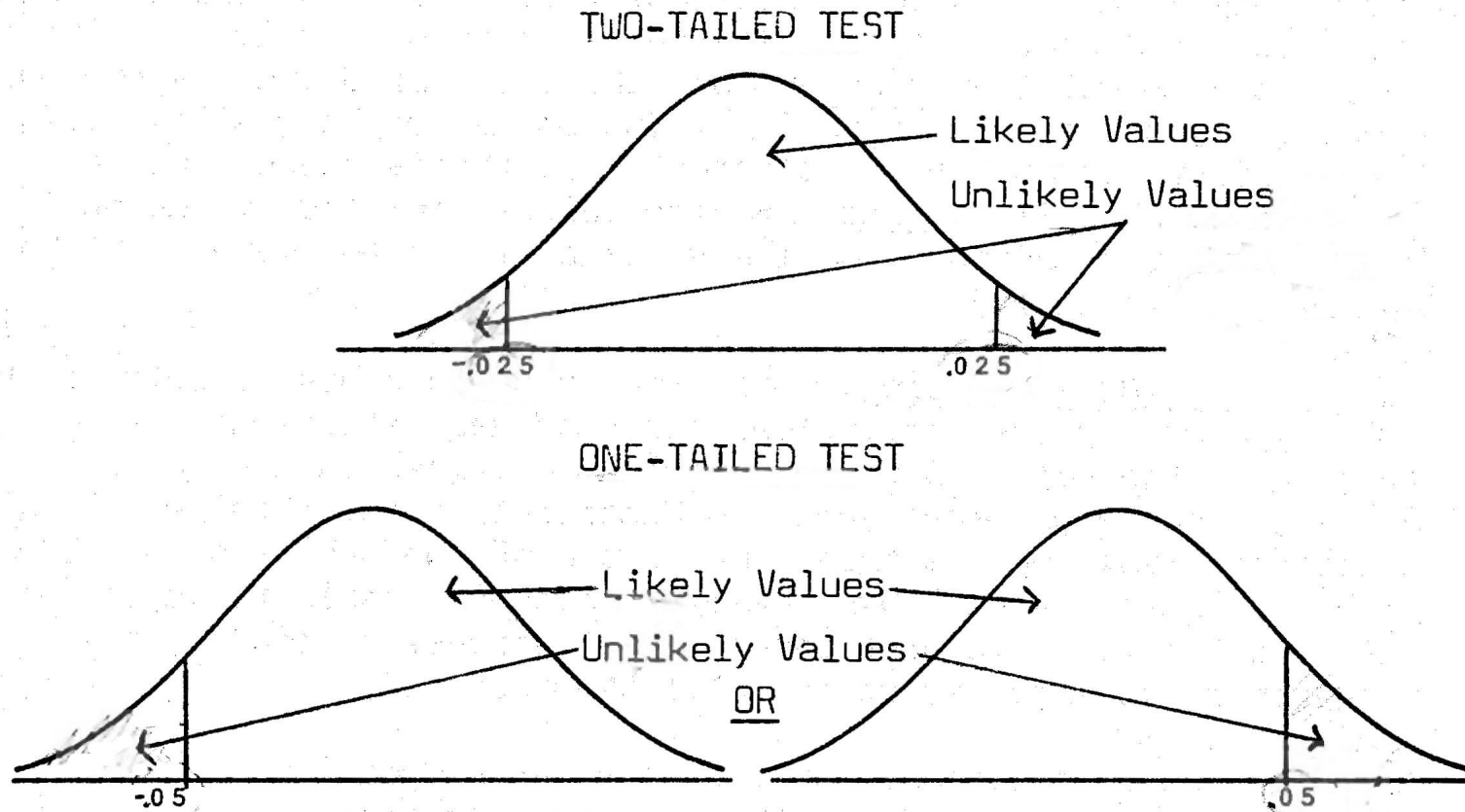
- Two independent sample t-test
 - Tests hypothesis that means of two independent groups are (not) equal.
 - Example: Two groups with high cholesterol participate. The first group receives a new drug; the second group receives a placebo. The experimenter hypothesizes that after 2 months LDL cholesterol levels will be lower for the first group than for the second group.
 - For two groups, t-tests and ANOVAs (F tests) are interchangeable.
-

T-tests

- Two matched samples t-test
 - Tests hypothesis that two sample means are equal when observations are correlated (e.g., pre-/post-test data; data from matched controls)
 - Example: 30 singers completed singing voice handicap index pre- and post-treatment. The mean score pre-treatment (42.49) was significantly less than the mean score post-treatment (27.5; $p < 0.01$) (Cohen et al., 2008).
-

One-tailed Versus Two-tailed Tests

Figure 11: Sampling Distribution -- Regions of Likely and Unlikely Values



Comparing Group Means: ANOVA

- One-way ANOVA: used when there are more than 2 levels of the (single) independent variable.
 - Tests the hypothesis that all sample means are equal.
 - Statistic = F
-

Example: One-way ANOVA (Al-Qahtahi, 2005)

- Fetuses hear recordings of voice, music, or control sounds. DV=measure of change in heart rate.
 - Mean changes:
 - Music: 4.68 (sd=10.58)
 - Voice: 3.54 (sd=9.99)
 - Control: -2.18 (sd=11.49)
 - The F test was significant at $p < 0.05$, indicating that at least one contrast is reliable.
-

Multi-way ANOVA

- Appropriate when there are two or more independent variables.
 - Assesses the probability that the population means (estimated by the various sample means) are equal.
-

Example: Two-way ANOVA (Zhuang et al., 2009)

- Speakers with polyps, nodules, or nothing
 - IVs: sex, diagnosis
 - DVs: Two measures of air flow (MFR, PTF; separate analyses)
 - Results:
 - Significant effects of sex and diagnosis
 - For PTF, post-hoc comparisons showed that only the difference between control and polyp subjects was significant.
 - For MFR, control differed from polyps, and polyps differed from nodules, but nodules did not differ from control.
-

Post-hoc Tests

- A significant F statistic means only that one of the group means is reliably different from the others.
 - Post-hoc tests identify which specific contrasts are significant (which groups differ reliably), and which do not, normally via a series of pairwise comparisons.
-

Post-hoc Tests

- Probabilities in post-hoc tests are cumulative: e.g., three comparisons at 0.05 level produce a cumulative probability of type I error of 0.15.
 - So: The probability of each test must equal $\alpha /$ the number of comparisons to preserve the overall significance level (Bonferroni correction).
 - Example: If there are 3 groups being compared at the 0.05 level (for a total of 3 comparisons), each must be significant at $p = 0.017$. If there are 4 groups (6 comparisons), $p = .0083$ for each.
-

Post-hoc Tests

- Many kinds of post-hoc tests exist for comparing all possible pairs of means: Scheffé (most stringent α control), Tukey's HSD, Neuman-Keuls (least stringent α control) are most common.
-

Example: Post-hoc Tests

- The significant F test in Al-Qahtani (2005) indicated that at least one contrast was significant.
 - Newman-Keuls post-hoc comparisons showed that fetal heart rate responses to music and voice did not differ ($p=0.49$), but that both differed significantly from the control condition (music vs. control: $p<0.014$; voice vs. control: $p < 0.033$).
-

Repeated Measures Designs

- Appropriate for cases where variables are measured more than once for a case (pre-/post-treatment, e.g.), or where there is other reason to suspect that the measures are correlated.
 - Matched pair t-tests are a simple case.
 - Repeated measures ANOVA is a more powerful technique.
 - Warning: Unless you know what you're doing, post-hoc tests in repeated measures designs require advanced planning and a statistical consultant.
-

Example: Repeated Measures ANOVA (Baker et al., 2008)

- A sample of normal children each performed 4 speech tasks: sustained /a/ in isolation and in a phrase; a repeated sentence; and counting to 10. F0 was measured for each task.
 - IVs: Task (within subjects), gender, age, evaluating clinician
 - Result: A significant effect of task was observed [$F(3, 41) = 6.12, p < .01$], but no other main effects were significant.
-

Example: Repeated Measures ANOVA

- Post-hoc comparisons showed that F0 was significantly higher during the counting task than during the phrase or sentence tasks.
- No other significant effects

	Vowel	Phrase	Sentence	Counting
Mean	240.5	236.6	235.7	246.5
SD	2.96	2.58	2.97	2.76

Correlation

- Correlation measures the extent to which knowing the value of one variable lets you predict the value of another.
 - Parametric version = Pearson's r
 - Nonparametric version = Rank-order correlation = Spearman's ρ
-

Correlation: Example (Salamão & Sundberg, 2008)

- Thirteen choir singers produced vowels in both modal and falsetto registers.
 - Measures of vocal source function were collected, including closed quotient, H1-H2, MFDR, etc.
 - Listeners judged whether the vowels sounded modal or falsetto, and the number of “modal” votes was counted.
 - Question 1: What relationships hold among the acoustic measures?
-

Correlation: Example

TABLE I. Pearson correlations between flow glottogram parameters, observed for the 104 flow glottograms analyzed.

	Q_{cl}	$H1-H2$	NAQ	ac amp	MFDR
Q_{cl}	1	-0.810 ^a	-0.861 ^a	0.245 ^b	0.651 ^a
$H1-H2$	-0.810 ^a	1	0.878 ^a	-0.317 ^a	-0.666 ^a
NAQ	-0.861 ^a	0.878 ^a	1	-0.172	-0.614 ^a
AC amp	0.245 ^b	-0.317 ^a	-0.172	1	0.766 ^b
MFDR	0.651 ^a	-0.666 ^a	-0.614 ^a	0.766 ^a	1

^aCorrelation is significant at the 0.01 level (two tailed).

^bCorrelation is significant at the 0.05 level (two tailed).

Regression

- Regression is basically the same thing as correlation, except you get an equation that lets you predict the exact values, rather than just a measure of how good that prediction is.
 - Multiple regression: Prediction of the dependent variable values is based on more than one independent variable.
-

Regression: Example 1

- Which acoustic variables are most strongly associated with the perception of modal versus falsetto voice quality?

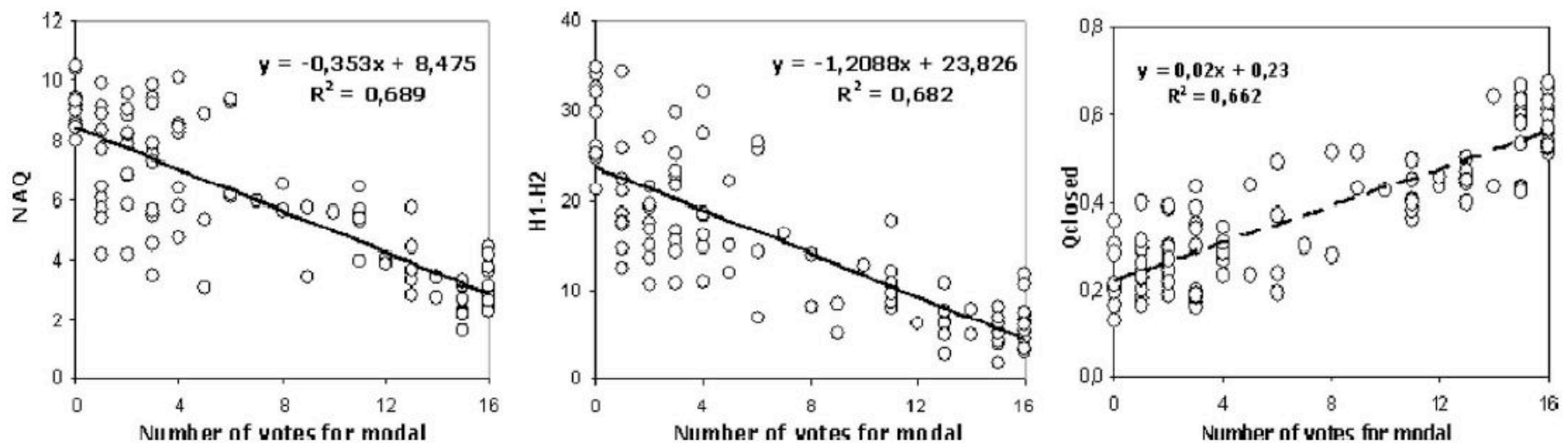


FIG. 3. [(a)–(c)] Q_{closed} , $H1-H2$, and NAQ as functions of the number of votes for modal register.

Regression: Example (Taylor et al., 2008)

- Question: How well do growl F0 and resonance frequencies predict the weight of a dog?
 - Answer: Together, they predict 66.3% of body weight [$F(2, 27) = 29.48, p < 0.01$]; BUT
 - Singly, resonance frequencies are a good predictor ($R^2=62.3\%$), but F0 is a bad predictor ($R^2 = 9\%$).
-

Correlation and Regression: Limitations

- Outliers
 - R^2 versus significance
 - The 'n' issue
 - Causation versus association
-

Global Average Temperature Vs. Number of Pirates

