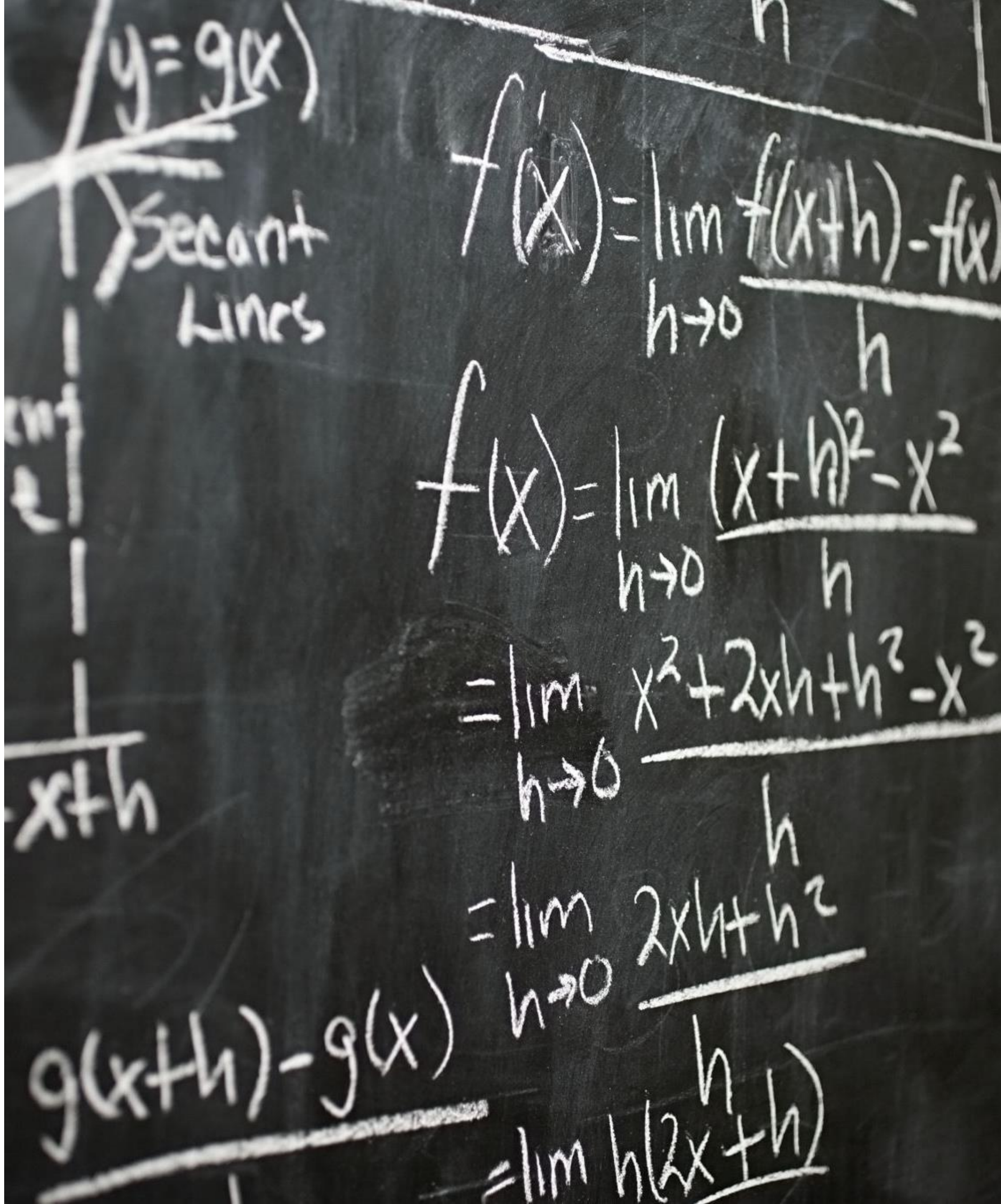
QUANTITATIVE METHODS **STATISTICAL TESTING**

Tomi Heimonen (he/him)

Computing and New Media Technologies
University of Wisconsin-Stevens Point



CONTENT

- Statistical hypothesis testing
- Statistical tests for different types of inference:
    - Comparing groups
    - Relationships
    - Associations

BIG PICTURE

- Hypothesis testing is a general framework for statistical inference.
- Within the framework, researchers need to select the appropriate statistical tests based on their hypotheses.
- Three general areas of test methods (not exhaustive)
  - Compare **differences** between two or more independent or paired groups
  - Examine the **linear relationship** between two or more variables
  - Measure **strength of association** between two categorical variables

HYPOTHESIS TESTING

- Hypothesis testing can be used to explore differences, relationships, associations or other dependencies between IVs and DVs.
  - Research question example: Was the average temperature in January 2024 higher than in January 2015?
- Hypothesis testing is a form of **confirmatory data analysis** – the goal is to test a hypothesis that applies to the relationship between two or more variables.
- When the data has been collected using randomized study design, hypothesis testing can also be used to **infer the cause** of the observed changes.

STATISTICAL HYPOTHESIS TESTING

- Statistical hypothesis testing is also called null hypothesis significance testing.
- The hypothesis we wish to confirm is proposed as an alternative (H_a) to the null hypothesis (H_o), which implies no relationship/association/difference (or effect) between variables – we generally want to be able to reject the null hypothesis.
  - *Reductio ad absurdum*: claim is assumed valid if counter-claim is improbable
- Example:
  - Null hypothesis: Type of detergent has no effect on the observed brightness of laundy.
  - Alternative hypothesis: New detergent formulation produces visibly brighter whites.

M. Gilchrist & P. Samuels, "Statistical hypothesis testing", http://www.statstutor.ac.uk/resources/uploaded/statisticalhypothesistesting2.pdf

SIGNIFICANCE LEVEL

- The effect is deemed to be statistically significant if the probability of observing an effect **if the null hypothesis is true** is under a certain significance level.
- The significance level is expressed as the **p-value**.
- Formally, p-value is the probability, under the null hypothesis, of obtaining a test statistic value equal or more extreme than what was observed.
- If the p-value is low enough, we reject the null hypothesis and accept the alternative hypothesis.
  - In practice, p-values of 0.05 and 0.01 are commonly used.

TESTING PROCEDURE

- Define the null hypothesis and alternative hypothesis.
- Decide the directionality – one-tailed or two-tailed hypothesis.
- Decide the desired significance level.
- Collect measurements.
- Calculate the test statistic and p-value.
- Decide if null hypothesis should be rejected or not.

STATISTICS USED IN HYPOTHESIS TESTING

- Statistical hypothesis testing generally examines some form of summary construct, such as **counts** (nominal variables), **proportions**, or **measures of central tendency** such as median (ordinal variables) or mean (continuous variables).
  - Null hypothesis (H_0): The average GPA of students attending an online program (M_{online}) and the average GPA of students attending a face-to-face program (M_{F2}) is the same, $M_{online} – M_{F2F} = 0$.
  - Alternative hypothesis (H_a): There is a difference in the average GPA between the programs.
- When can we generalize the findings of the test to the population?
  - When observations have been randomly sampled from the population and the sample size is sufficiently large (> ~30 observations).
  - When your dissertation supervisor and/or journal reviewers agree with your interpretation ☺

HYPOTHESIS DIRECTIONALITY

- Wen we construct our null hypothesis, the alternative hypothesis can take one of two forms:
  - A **one-tailed** prediction typically means that the null hypothesis posits an order between the groups – for example, that classroom teaching is more effective than online teaching
  - A **two-tailed** prediction does not posit a direction – it states that there is a significant effect, but it could be either positive or negative (average classroom GPA > online GPA OR online GPA > classroom GPA)
- Typically, a two-tailed prediction is preferred, unless a theory or the study procedure suggest specific direction (they almost never do).

POWER AND ERRORS

- **Type I error** occurs when we reject the null hypothesis when it holds (false positive).
  - The threshold of rejecting the null hypothesis, α, or the significance level that the p-value should not exceed.
- The probability of not rejecting the null hypothesis when it is false, β, is called **a Type II error** (false negative).
- The **power** of a test is its probability of correctly rejecting a false null hypothesis (1 - β).
- Lowering the required significance level of the test leads to increasing the probability of Type II error and lowering the power of the test.
  - Other factors affecting power include direction of test, standard deviation, and sample size.

Source: M. Gilchrist & P. Samuels, "Statistical hypothesis testing", http://www.statstutor.ac.uk/resources/uploaded/statisticalhypothesistesting2.pdf

INTERPRETING SIGNIFICANT TEST RESULTS

- Generally, if a significant difference is found, we have **some evidence** that the independent variable has an effect on the dependent variable beyond random chance.
- Statistical significance does not, however, imply practical significance!
- Practical significance can be inferred from the **effect size**.
  - For example, what is the difference between the sample means?
  - Does the difference have a meaningful real world impact?

INTERPRETING NON-SIGNIFICANT RESULTS

- Failure to reject the null hypothesis does not mean we must accept it!
  - It is impossible to prove the negative, since we do not know the exact true value of the population parameters.
- If the p-value is low but not under the desired significance level, it is still an indication of there being some effect, albeit inconclusive.
- Important: p-value does not express anything about how plausible a hypothesis is in reality!
  - The p-value guarantees, if calculated correctly, that Type I error rate of the test is at most α.
  - Example: if we reject a null hypothesis that the moon is not made out of cheese does not mean that the moon is indeed made out of cheese.

STATISTICAL TEST ASSUMPTIONS

- Most statistical tests make assumptions regarding the variable types and/or the distribution and variability of the data.
- Generally, the assumptions regarding variable types can be met through study design:
  - How independent and dependent variables are assigned.
  - How dependent variable values are measured (data type).
  - How participants are recruited, or how measurements are taken – each sample has an equal probability of being chosen.
- Assumptions regarding the distribution and variability of the observed values must generally be checked before the test is run.
  - Many statistical software, like SPSS, have the option to run the checks as a part of the test or separately.

COMPARING GROUPS

APPROACHES TO COMPARING GROUPS

- Things to consider:
  - Are we comparing data from the same set of users or across different users?
  - How many samples are we comparing: two samples or more than two samples?
- In statistical testing parlance, **independent samples** means that samples come from different groups of users.
- **Paired samples** (or matched pairs) means that measurements are collected from the same individual for both samples – also called a **repeated measures** approach.
- Generally, the appropriate statistical test for comparing groups is the **t-test** – it has variants for both scenarios.
- Process:
  - Calculate **t-value**, size of difference between samples relative to variation in the sample data.
  - The greater the value of **t**, the larger the evidence against null hypothesis.

INDEPENDENT SAMPLES T-TEST ASSUMPTIONS

1. Variability of data in each group is approximately equal (**homogeneity of variance**).
2. The data is approximately normally distributed.
3. The observations in the samples are independent: observations in one group do not depend on observations in the other group or on each other.

- Small violations of homogeneity of variance and normality are generally ok.
- The test is robust to departures from normality for large sample sizes ($n > 30$).

Patrick Runkel, "What are T Values and P Values in Statistics?", http://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-are-t-values-and-p-values-in-statistics

PAIRED SAMPLES T-TEST ASSUMPTIONS

1. Distribution of the **differences between groups** is normally distributed.
2. Pairs of observations should be independent of each other.

- In practice, we compute a one sample t-test on the mean difference between the paired measurements, with a null hypothesis that the mean difference equals 0.

David M. Lane, "Testing a Single Mean", http://onlinestatbook.com/2/tests_of_means/single_mean.html

COMPARING MORE THAN TWO GROUPS

- The initial step is to perform an **omnibus test** to determine if there is a significant effect of the independent variable across the groups.
- For most test designs, a **single factor analysis of variance** (ANOVA) is a sufficient approach.
  - Factor = independent variable
  - ANOVA can be carried out for both independent samples (one-way ANOVA) and paired samples (called repeated measures ANOVA).
- Assumptions are mostly the same as the t-test variants.
  - Repeated measures ANOVA also assumes **sphericity** – variances of all combinations of related groups should be equal.
- If the omnibus test is significant, calculate pairwise comparisons of group means to identify where the differences are.

PROBLEM WITH MULTIPLE PAIRWISE COMPARISONS

- Each statistical test has a small probability for a Type I error.
- Probability of making a Type I error is multiplied when comparing multiple groups.
  - With three tests with $\alpha = 0.05$, the probability of making at least one Type 1 error is ~ 15%.
- Solution?
  - Set a lower significance level for each individual comparison to protect against Type I error (e.g., $0.05\ / k$), where k is the number of comparisons – also called **Bonferroni correction**.
  - This is a very conservative approach and inflates Type II error.
- In practice, statistical software packages provide more sensitive techniques to correct for multiple comparisons.

WHAT TO DO IF THE ASSUMPTIONS FOR A STATISTICAL TEST ARE NOT MET?

- Each t-test and ANOVA variant generally has a **non-parametric** alternative that does not assume a specific distribution for the data.
- Other situations when a **non-parametric test** is a better alternative:
  - Median is a better measure of central tendency than mean.
  - Sample size is very small.
  - Measured data is ordinal or non-continous.
  - There are outliers in the data (data points that are more than 3 standard deviations away from the mean).

Jim Frost, "Choosing Between a Nonparametric Test and a Parametric Test", http://blog.minitab.com/blog/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test

NON-PARAMETRIC TESTS

- Note: non-parametric tests do have assumptions, too!
  - For example, distributions should be symmetrical or have similar shapes.
  - Always check the assumptions of the test!
- Problem: non-parametric tests have less **statistical power** than parametric tests.
  - We may end up missing significant results when they exist.

Parametric test	Non-parametric test	Characteristics
1-sample t-test	Sign test	• Test on median of signed differences to hypothesized value • "Distribution-free"
2-sample t-test (independent samples)	Mann-Whitney U	• Test on difference of medians • Assumes similar distribution shape
2-sample t-test (paired samples)	Wilcoxon signed-rank test	• Test on median of paired differences • Assumes symmetric distribution and at least interval scale data
One-way ANOVA	Kruskal-Wallis test	• Test on the equality of medians, two or more groups • Assumes similar distribution shape

Eston Martz, "Data Not Normal? Try Letting It Be, with a Nonparametric Hypothesis Test", http://blog.minitab.com/blog/understanding-statistics/data-not-normal-try-letting-it-be-with-a-nonparametric-hypothesis-test

RELATIONSHIPS AND ASSOCIATIONS

ASSOCIATION BETWEEN TWO CONTINUOUS VARIABLES: CORRELATION

- **Correlation** measures the strength and directionality of the association between two variables.
- Pearson correlation coefficient (known as *r*) is a measure of **linear correlation** between two **continuous variables**.
  - The value of *r* indicates how far observed values are from line of best fit – value of 1.0 indicates perfect fit.
  - Common interpretation*: r > 0.6 is strong, > 0.4 moderate and > 0.2 weak correlation.
- Assumptions:
  - Observations should be independent from each other.
  - Relationship between variables should be linear (scatterplot should approximately resemble a straight line).
  - Homoscedasticity (homogeneity of variance) between observed and fitted values.
  - Variable values should be approximately normally distributed.
- If the relationship is **monotonic** but not linear, Spearman's rank correlation may be used.

ASSOCIATION BETWEEN TWO CONTINUNOUS VARIABLES: LINEAR REGRESSION

- Correlation analysis provides information about the strength and direction of association – linear regression **estimates the parameters of the equation that is used to predict** the values of one variable (Y) based on another (X).
- Formula: $Y = \beta_0 + \beta_1 X + \epsilon$
    - Y = DV value
    - X = IV value
    - β_0 and β_0 are the intercept and slope of the equation, respectively
    - ϵ is the error term (captures the difference between observed and predicted values of Y)
- Relevant statistic is R^2, which represents the proportion of variance in the DV explained by the IV – or how well the regression line approximates the real data points.
- Correlation is more appropriate to use when trying to characterize the relationship between variable – linear regression is more appropriate when looking to predict or explain the behavior of one variable (DV) based on manipulation of the other (IV).

ASSOCIATION BETWEEN TWO CATEGORICAL VARIABLES

- The **chi-square test for independence** can be used to determine if there a significant association between categorical variables.
  - The null hypothesis is that there is no association.
- Example: Does class attendance affect performance?
  - Categories: Attends class, Skips class
  - Measured data: count of students who Pass and Fail, for each category
- Assumptions:
  - There are two nominal variables (categories).
  - Data in table cells should be counts.
  - Categories of variables must be mutually exclusive.
  - One participant can contribute to one and only one cell.
  - Sample size should exceed # of cells multiplied by 5 (e.g., 4 cells → N > 20).

CHI-SQUARE TEST FOR INDEPENDENCE

- Process:
  - The data are put into a 2×2 contingency table that summarizes the frequencies.
  - Calculate the expected frequences for each cell in the table.
  - Compute the chi-square (χ^2) statistic and look up the p-value based on degrees of freedom and χ^2.
- Assumptions:
  - Observations should be independent.
  - The data are categorical.
  - Data in table cells should be counts.
  - Categories of variables must be mutually exclusive.
  - The expected frequency in each sell should be at least 5.
- An alternative for small sample sizes is known as Fisher's Exact Test.

BUT MY RESEARCH QUESTION OR SETTING IS DIFFERENT – WHAT DO I DO?

- My general advice would be to consult a more experienced researcher before data collection to verify that your experimental setting is reasonable (ask me how I know).
  - Caveat emptor – some folks may have their "pet" statistical methods they use for everything, which can lead to a Maslow's hammer situation.
- There are textbooks and tools available that can support the process of identifying the appropriate test or tests – but you do need to have general understanding of the testing procedures.
  - Example: Social Science Statistics test wizard