Biostatistics

Statistical tests for data analysis

When to use which?

Null Hypothesis and Testing

- A null hypothesis, proposes that no significant difference exists in a set of given observations. For the purpose of these tests in general
- **Null**: Given two sample means are equal
- **Alternate**: Given two sample means are not equal
- For rejecting a null hypothesis, a test statistic is calculated.
- This test-statistic is then compared with a critical value and if it is found to be greater than the critical value the hypothesis is rejected

Critical Value

- A critical value is a point (or points) on the scale of the test statistic beyond which we reject the null hypothesis
- It is derived from the level of significance α of the test.
- Critical value can tell us, what is the probability of two sample means belonging to the same **distribution**.
- Higher, the critical value means lower the probability of two samples belonging to same distribution.
- The general critical value for a two-tailed test is 1.96, which is based on the fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean.

Relationship between p-value, critical value and test statistic

- A null hypothesis, proposes that no significant difference exists in a set of given observations. For the purpose of these tests in general
- Null: Given two sample means are equal
- Alternate: Given two sample means are not equal
- For rejecting a null hypothesis, a test statistic is calculated. This test-statistic is then compared with a critical value and if it is found to be greater than the critical value the hypothesis is rejected

Z-test

- In a z-test, the sample is assumed to be normally distributed.
- A z-score is calculated with population parameters such as “population mean” and “population standard deviation” and is used to validate a hypothesis that the sample drawn belongs to the same population.
- Null: Sample mean is same as the population mean
- Alternate: Sample mean is not same as the population mean

Z-test

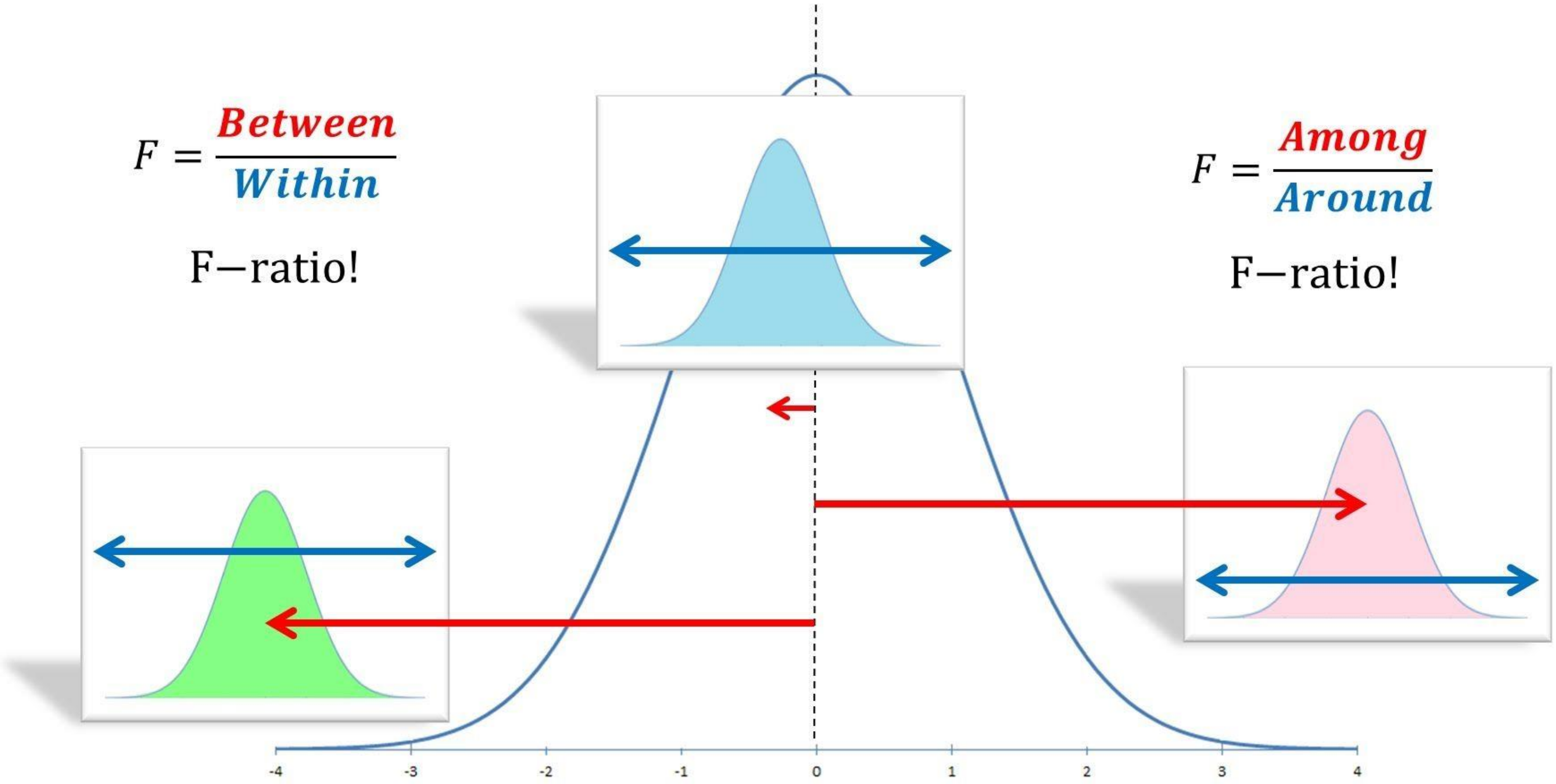
- The statistics used for this hypothesis testing is called z-statistic, the score for which is calculated as
- **$z = (x — \mu) / (\sigma / \sqrt{n})$,** where
- x= sample mean
- μ = population mean
- $\sigma / \sqrt{n}$ = population standard deviation
- If the test statistic is lower than the critical value, accept the hypothesis or else reject the hypothesis

Comparison of ANOVA and t test

- The *t*-test is a method that determines whether ***two*** populations are statistically different from each other
- Paired and unpaired t test
- ANOVA determines whether ***three or more*** populations are statistically different from each other.
- Both of them look at the difference in means and the spread of the distributions (i.e., variance) across groups
- however, the ways that they determine the statistical significance are different.

ANOVA: Analysis of Variance is a *variability ratio*

$$Variance\ Between + Variance\ Within = Total\ Variance$$



$$F = \frac{Between}{Within}$$

F−ratio!

$$F = \frac{Among}{Around}$$

F−ratio!

https://i.pinimg.com/originals/eb/94/f9/eb94f9bae12be2d6617549bd22e7d216.jpg

Comparison of ANOVA and t test

- These tests are performed when
- 1) the samples are independent of each other and
- 2) have (approximately) normal distributions or when the sample number is high (e.g., > 30 per group).
- More samples are better, but the tests can be performed with as little as 3 samples per condition.

Comparison of ANOVA and t test

- The *t*-test and ANOVA produce a test statistic value ("t" or "F", respectively), which is converted into a "**p-value.**"
- A p-value is the probability that the null hypothesis – that both (or all) populations are the same – is true.
- In other words, a lower p-value reflects a value that is more significantly different across populations.
- Biomarkers with significant differences between sample populations have p-values ≤ 0.05.

- 1. **One-way ANOVA**: It is used to compare the difference between the three or more samples/groups of a single independent variable.
- 2. **MANOVA**: MANOVA allows us to test the effect of one or more independent variable on two or more dependent variables. In addition, MANOVA can also detect the difference in co-relation between dependent variables given the groups of independent variables.
- The hypothesis being tested in ANOVA is
- Null: All pairs of samples are same i.e. all sample means are equal
- Alternate: At least one pair of samples is significantly different

- The statistics used to measure the significance, in this case, is called F-statistics. The F value is calculated using the formula
- **F= ((SSE1 — SSE2)/m)/ SSE2/n-k**, where
- SSE = residual sum of squares
- m = number of restrictions
- k = number of independent variables
- There are multiple tools available such as SPSS, R packages, Excel etc. to carry out ANOVA on a given sample.

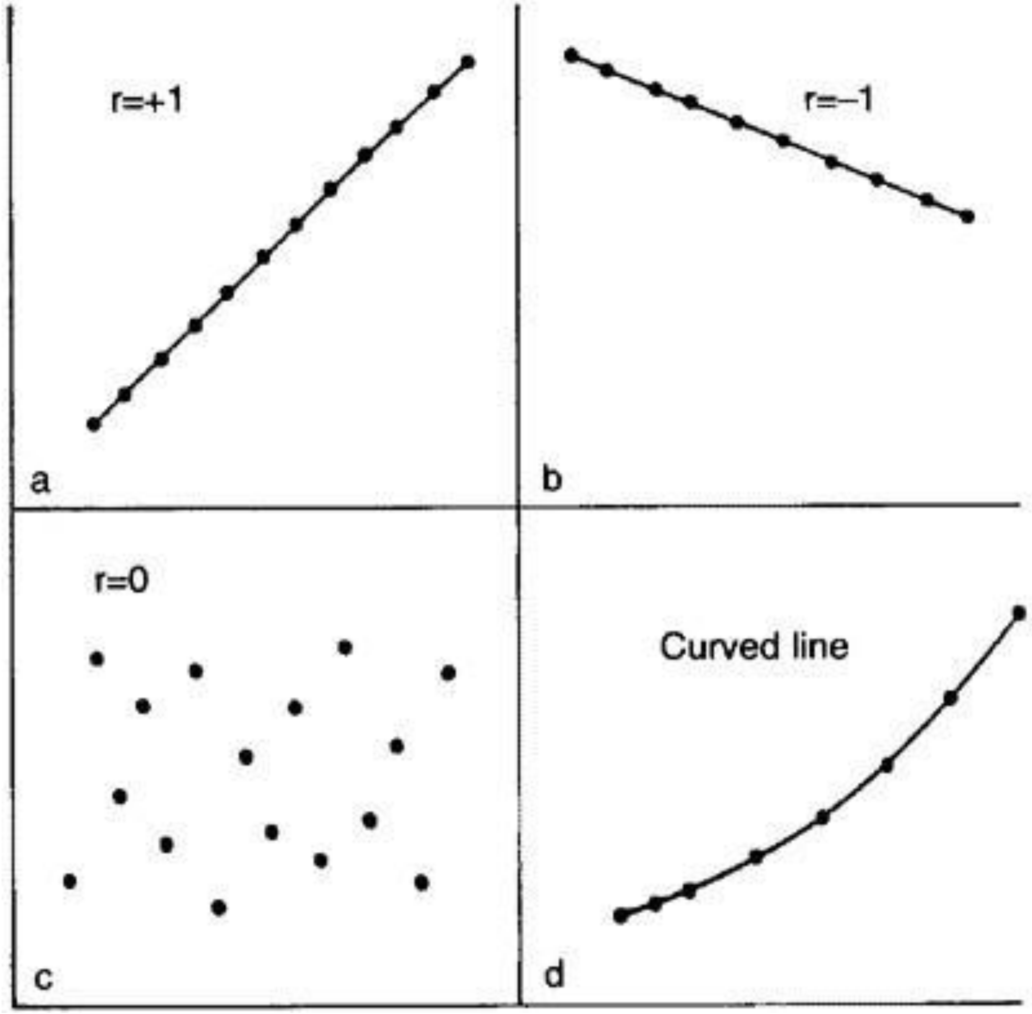
Simple Correlation

- Sometimes we wish to know if there is a relationship between two variables.
- A simple correlation measures the relationship between two variables.
- The variables have equal status and are not considered independent variables or dependent variables.
- Pearson*‘s r* measures a linear relationship between two continuous variables.
- Other types of relationships with other types of variables exist

Simple Correlation

- A sample research question for a simple correlation is,
- “*What is the relationship between height and arm span*?”
- A sample answer is, “There is a relationship between height and arm span, $r(34)=.87$, $p<.05$.”
- A canonical correlation measures the relationship between sets of multiple variables (a multivariate statistic)

Simple Correlation

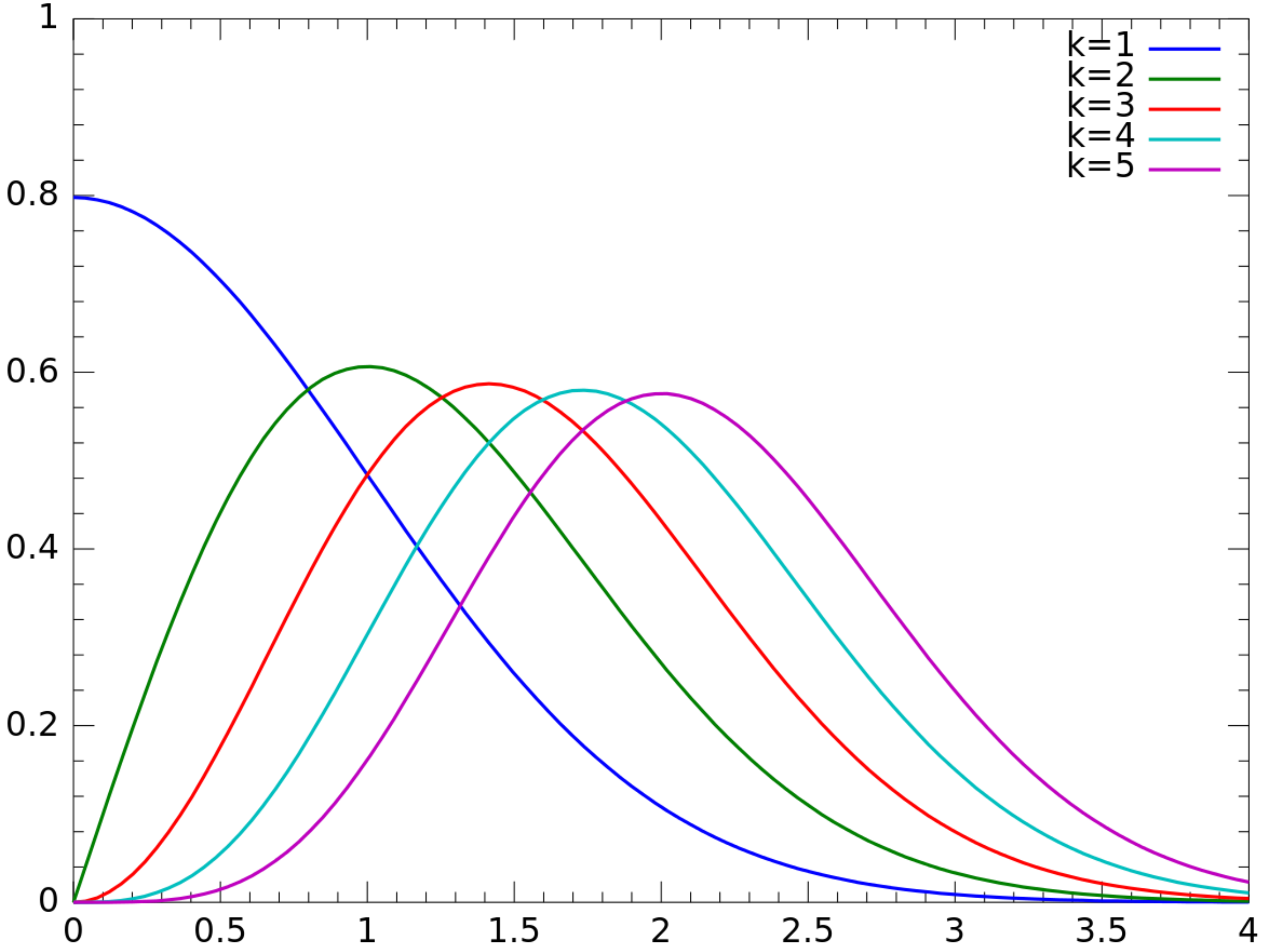


https://www.bmj.com/sites/default/files/imagecache/highwire_fragment_medium/attachments/resources/2011/08/figure111.jpg

Chi-Square Test

- ✓ Chi-square test is used to compare categorical variables. There are two type of chi-square test
- ✓ 1. Goodness of fit test, which determines if a sample matches the population.
- ✓ 2. A chi-square fit test for two independent variables is used to compare two variables in a contingency table to check if the data fits.
- ✓ a. A small chi-square value means that data fits
- ✓ b. A high chi-square value means that data doesn’t fit.

Chi Distribution



https://upload.wikimedia.org/wikipedia/commons/thumb/3/35/Chi_distribution_PDF.svg/1200px-Chi_distribution_PDF.svg.png