10: Data transformation and non-parametric tests

What to do if t-test assumptions are substantially violated?

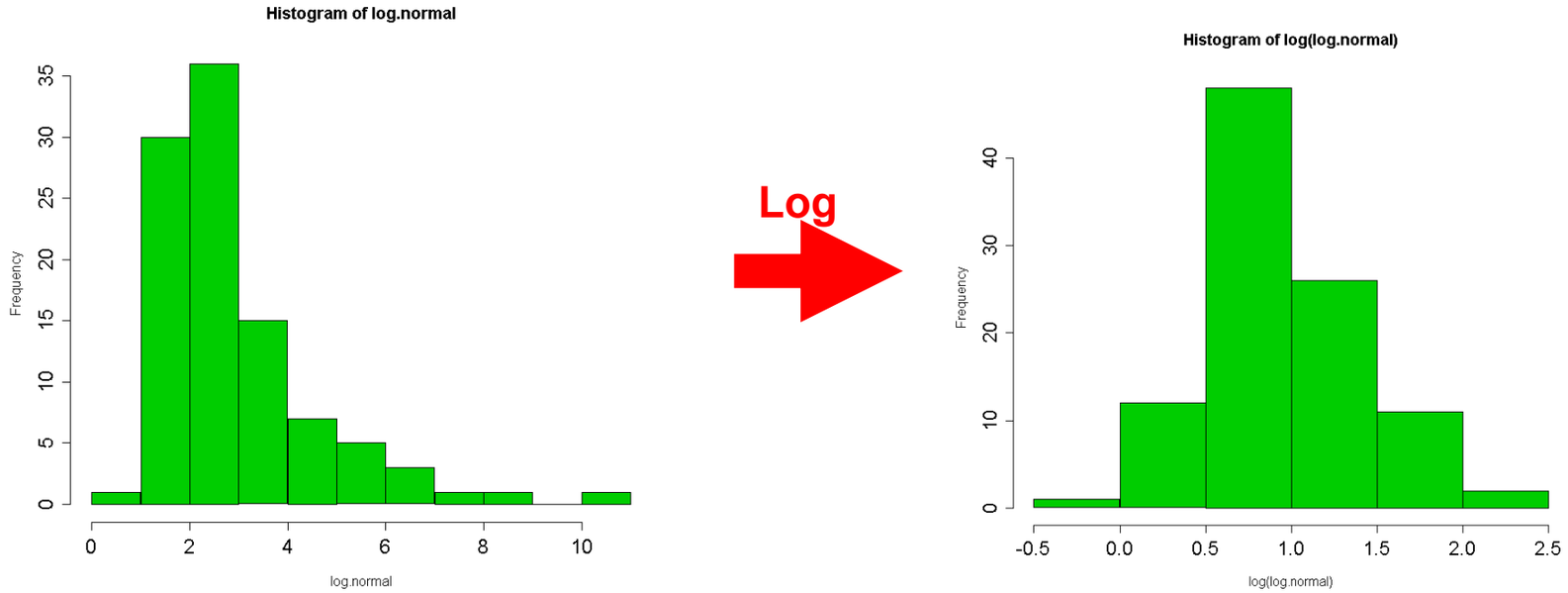
- Large difference in variances
  - Welch approximation usable only when the difference is low to moderate (and with rather high number of observations)
  - The data might follow the log-normal distribution → use transformation
  - Use a non-parametric test (but this might be tricky)
- Data do not come from a normal distribution
  - Check the log-normal possibility
  - Use non-parametric tests

The log-normal distribution

- $\log(X) \sim N(\mu, \sigma^2)$
- Positively skewed
- Defined for numbers > 0
- Very common situation in biological research
  - Masses, dimension of biological objects
  - Counts can be approximated by log-normal distribution

Data transformation using log - function

- Changes the scale from additive to multiplicative
  - geometric instead of arithmetic means; exp(mean(log-data) = geometric mean
  - H_0: The ratio between geometric means is 1.0
  - Results say how many times the mean is larger (e.g. 1.2 times = by 20%)
- If suitable, improves both normality and homogeneity of variances
- Test results do not depend on the type of logarithm used (just consistency is needed)



Some more tricky types of data

- Ordinal data
- e.g. behavioral experiments
  - Measures of reaction of an animal on an impulse
- Data do not follow the normal distribution
- Transformation provides no help
- **Non-parametric tests**
  - Do not test null hypotheses on parameters of the distributions

Various non-parametric analogues of t-tests

- Permutation tests
  - Based on the principle of repeated random re-assignment of data to groups and calculating the t
  - P-value corresponds to number of observations for which t is higher than that calculated based on the original data/total number of permutations

Number of permutations where $|t_{permut}| \geq |t_{data}|$ → $\frac{x+1}{n+1}$

Total number of permutations where $|t_{permut}| \geq |t_{data}|$ →

Non-parametric tests based on order

- Mann-Whitney test
  - Analogue of a two-sample t-test
  - Original values replaced by their order in the whole dataset
  - These are then used for the calculation of the U statistic
    - P-value based on direct comparison to theoretical U distribution
    - Or approximation to normalized normal distribution (Z) – usually applied if ties are present
- Wilcoxon test
  - Analogue of a paired t-test
  - P-value based also mostly on normal (Z) approximation (if ties are present)
- Kruskal-Wallis test
  - Analogue of ANOVA
  - Dunn test for multiple comparisons
- Spearman correlation coefficient
  - Order-based non-parametric correlation coefficient

Non-parametric tests have also some assumptions

- Identical (though not normal) distributions from which the samples come
  - If we state the null hypothesis about the shift (i.e. difference of means)
- Homogeneity of variances, quite similar to t-test/ANOVA
- Same size of intervals for data on the ordinal scale