Quantitative Methods and Machine Learning

Inferential Statistics

Mathias Müller

Recap: descriptive statistics

- What is statistics? What are statistical questions?
- Descriptive statistics vs. inferential statistics
- Combinatorics
- Variables, scales
- Categorical vs. numerical

Recap: descriptive statistics

- Central tendencies, “averages”: mean, median and mode
- Measures of “spread”: range, IQR, variance, standard deviation
- impact of outliers

Distributions

- What is a distribution?
- Discrete vs. continuous
- Probability mass / density functions
- Area under a density curve
- Distributions with useful properties: (standard) normal, binomial, Bernoulli (…)
- Skewedness, tails to the left and right

Normal distribution

- “Gaussian”, “bell curve”
- Likely if data generating process is random
- Standard normal distribution: zero mean and unit variance
- Area under the curve, probability in an interval and distance from mean
- 68-95-99.7 rule
- Incredibly handy! We want every problem to be a normal distribution!

Sampling from a population

- Reasonable samples
- Sources of bias: e.g. undercoverage, nonresponses
- Random vs. selective sampling
- Uncertainty about population introduced by sampling

Population vs. sample terminology / notation

- Number of data points, mean, variance, standard deviation

Population vs. sample properties

- Population descriptives likely unknown
- need: way to estimate population parameters from sample
- Sample mean,
- (unbiased) sample variance, degrees of freedom

Point estimation

- Estimators for unknown model (population) parameters
- “educated guesses”

Z scores

- Distance from mean expressed as number of standard deviations
- Z table
- When sample size is too small: t statistics

Central limit theorem

- If samples of size n are drawn an infinite number of times, the mean of the sampling distribution of the sample mean will equal the mean of the population
- Sampling distribution will approach normal distribution as the number of trials grows
- When to assume approximately normal distributions

Confidence intervals

- interval in sampling distribution of the sample mean to indicate probability area for a certain margin of error
- Single sample vs. resampling

Correlation

- Correlation vs causation
- Relationship between variables: strength, linearity, direction
- Correlation coefficient
- Preview: linear regression, minimize sum of squared residuals

Hypothesis testing

- Null hypothesis, alternative
- Type 1 error
- Significance level, p value