Probabilities

- Joint probability. Independent events: chance of both occurring is multiplication of probabilities of both events (e.g. 2 heads is 0.5*0.5 = 0.25)
    - (e.g. probability of a card that is hearts and J, Q or K is 13/52 * 12/52 = 0.25 * 0.24 = 3/52 = 1/17 = 0.058)
- Mutually exlusive events: chance of any of them occurring is addition of their probabilities (e.g. rolling 2 or 3 is 0.1666+0.166 = 1/3)
- Not mutually exclusive events: addition of their probabilities minus the probability that their combination occurs (e.g. probability of a card that is hearts or J,Q or K is 13/52 + 12/52 – 3/52 = 22/52 = 11/26)
- Conditional probability: dependent on former occurrence (e.g. a bag with 2 red and 2 blue balls. 1 red ball is drawn (prior). Probability of drawing a red ball now (posterior) is 33% (posterior) instead of 50% at the beginning (prior))
- Bayes' rule (inverse probability) relates posterior to prior probability: P(A|B) = (P(B|A)*P(A))/P(B)
    - P(A|B) = posterior probability
    - P(B), P(A) = prior probabilities
    - In above example: P(blue|red) = (P(red|blue) * P(blue)) / P(red)
    - = (2/3 * 1/2) / 1/2 = 2/3

Descriptive Statistics

focus on the sample itself, no probabilities or confidence in regards to whole population

- Includes measures of central tendency (mean, mode etc.) and dispersion (variance, box plot etc.)
    - Variance = (Sumof(x-u)^2) / n
        - Where x is the data points and u is the mean
    - Standard deviation is square root of variance
    - Boxplot: middle line tells us the median, box starts with 1st quantile and ends at 3rd quantile
- Univariate analysis (1 variable) vs.
- Bivariate and Multivariate analysis: relation between variables
    - Correlation
    - Covariance
    - Scatter matrix (pd.scatter_matrix())

Inferential Statistics

statistics based on a sample, proposing about whole population. Concerned with: point estimate, interval estimate, hypothesis verification or clustering and classification

- Sampling – drawing sample distributions from the distribution of the whole population
- Overview: p-value, Chi-Square, z-score: Distribution ||| T-test, F-test, R-Squared, ANOVA: Mean

Distribution

- Bernoulli /binomial distribution: distribution of events categorized into 2 descrete outcomes
- **Z-score**: verifiying or rejecting null hypothesis of normal distribution

    - Test statistic: (sample mean – population mean) / (standard deviation / square root of sample size)
    - If result larger than z-score -> reject null hypothesis and verify alternative hypothesis
    - Alternative hypothesis
        - > than mean right-tail test
        - < than means left-tail test
        - ≠ than means two-tail test
    - Z table tells us the confidence interval for a random point being within a certain number of standard deviations away from the mean with a certain % confidence
    - **p-value based on Z-table / Z-score**
        - p-value is compared to significance level α to verify or falsify the null hypothesis (normal **distribution**) against an alternative hypothesis
            - If $p < \alpha$, null hypothesis is rejected, otherwise verified
            - SciPy: normaltest() function for p-value, ttest_ind() to compare means of distributions
- **Chi-Squared** test: tests independence (=null hypothesis) of two categorical variables
    - Sum of the square of the difference between observed and expected value, divided by expected value
    - Goodness of fit tests whether 2 samples are drawn from identical distributions (Kolmogorov-Smirnov test) or whether a dataset follows the hypothesis of a specific **distribution** (i.e. normal distribution, Pearson's chi-squared test)

Mean & Variance (Python library: statsmodels.stats

- R-squared: tells us, how much of the variance of the dependent variable can be explained by the variance of the independent variable(s)
- **T-test** tests the null hypothesis of what value the **mean** of normalized distribution is
    - Two-sample test compares means of 2 populations, checks for significant difference
    - **ANOVA** does the same for the means of more than 2 populations
        - Ex.: testing the effect of different treatments on patients
    - **F-test** tests null hypothesis that the means of normally distributed populations with the same standard deviation are equal
        - Defined as explained variance divided by unexplained variance or
        - Between-group variability divided by within-group variability
- Degrees of freedom, usually n-1: number of data points that are free to vary without hurting a statistical condition (e.g. given a specific mean and a dataset with 10 datapoints, the first 9 data points can vary in their values, thus degree of freedom is 9)

A/B Testing

- Python library: statsmodels.stats (also for F, Z and Chi Square)
- Change aversion (don't want anything new) and novelty effect (want something new)
- Structure
    - Define Marketing Metric (like DAU or CTR)
        - Sensitivity (metric really changes when things happen that you care about) and robustness (metric does not change when things happen that you don't care about)

- Define significance level (alpha, 0.95) for p-value (Type I error, false positive) and the statistical power (1-Beta, Type II error, false negative, mostly at least 80%)
- Calculate required sample size
    - based on Alpha, 1-Beta and unit of diversion (e.g. userid, pageviews)
- Schedule the tests, run the tests and analyze results (look into different subgroups, segments)
- Check (i.e. novelty effect, change aversion, seasonality, correlation/causation, business sense)