2.1 Selecting statistical approaches

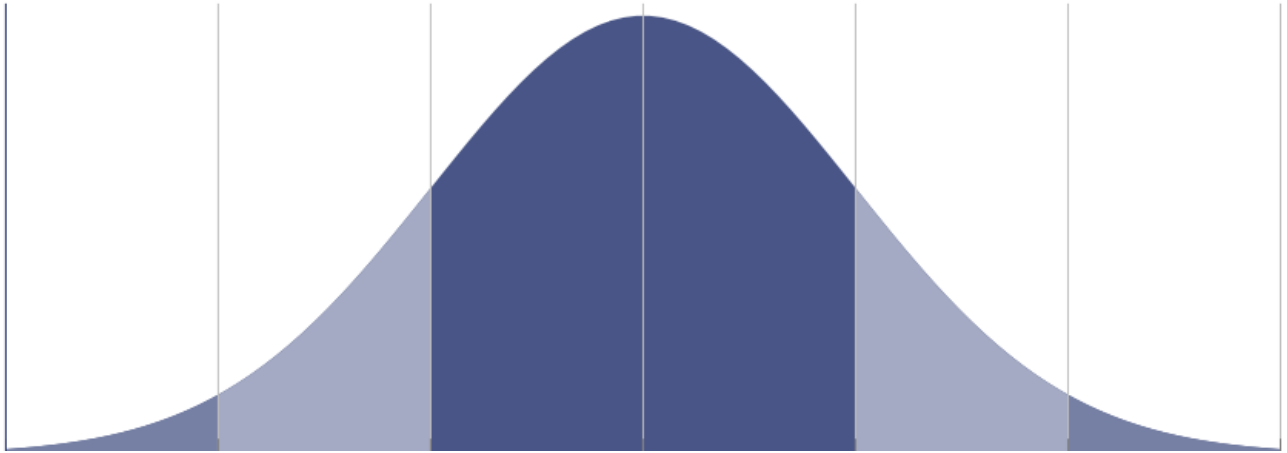
Overview

1. **Normal distribution**
2. Hypothesis testing
3. Summary statistical tests

Recap last week

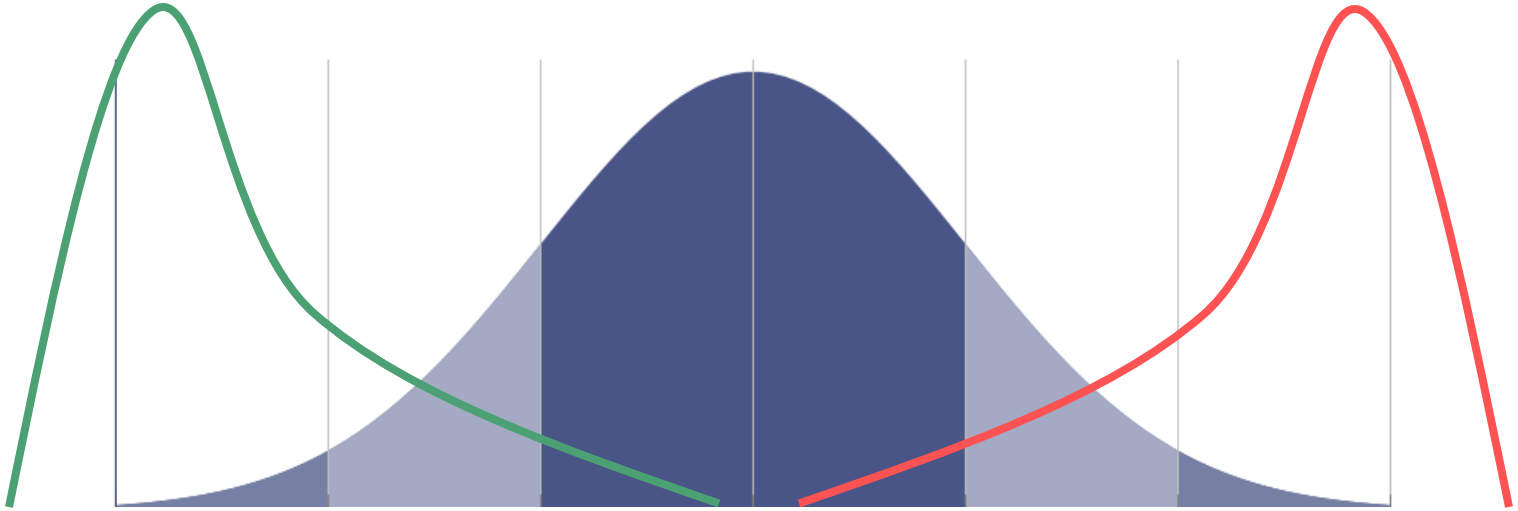
Mode	Most common value	Nominal (categorical), ordinal, interval/ratio
Median	(most) middle value	Ordinal, interval/ratio
Mean	Sum all values divided by amount of observations	Interval/ratio

Distribution



Normal distribution
Mode = median = mean

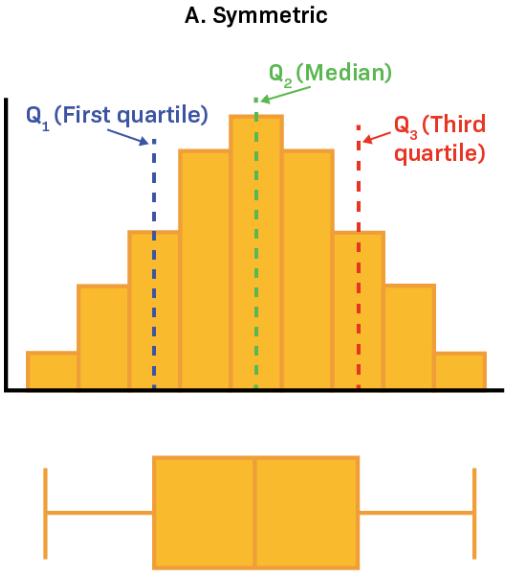
Distribution: skewness



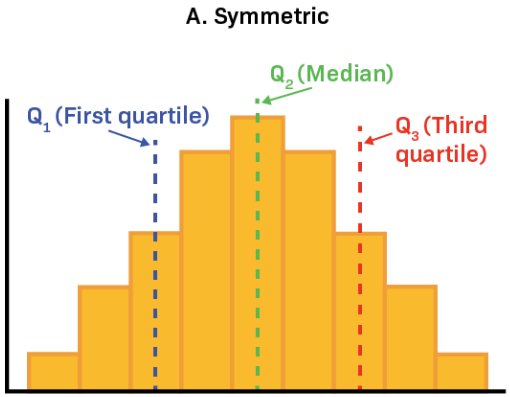
Positive (right) skew
Median < mean
E.g. wages, capital

Negative (left) skew
Median > mean
E.g. age of death

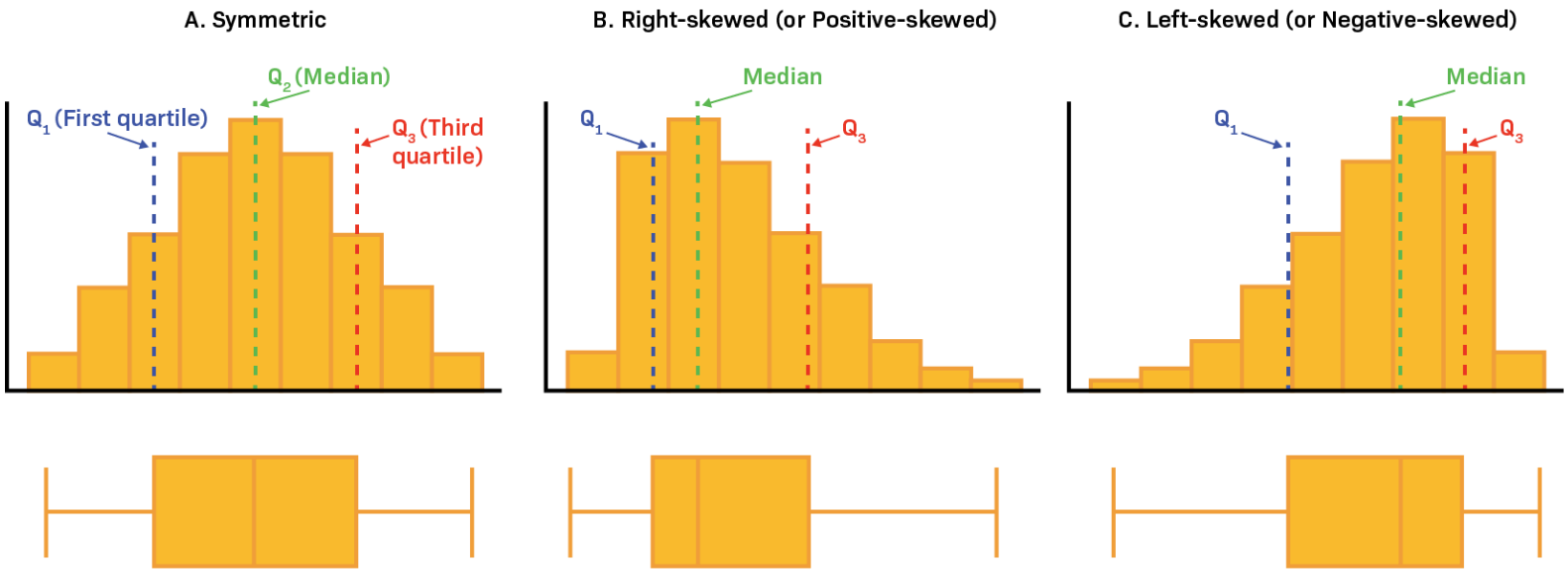
Distribution: boxplot



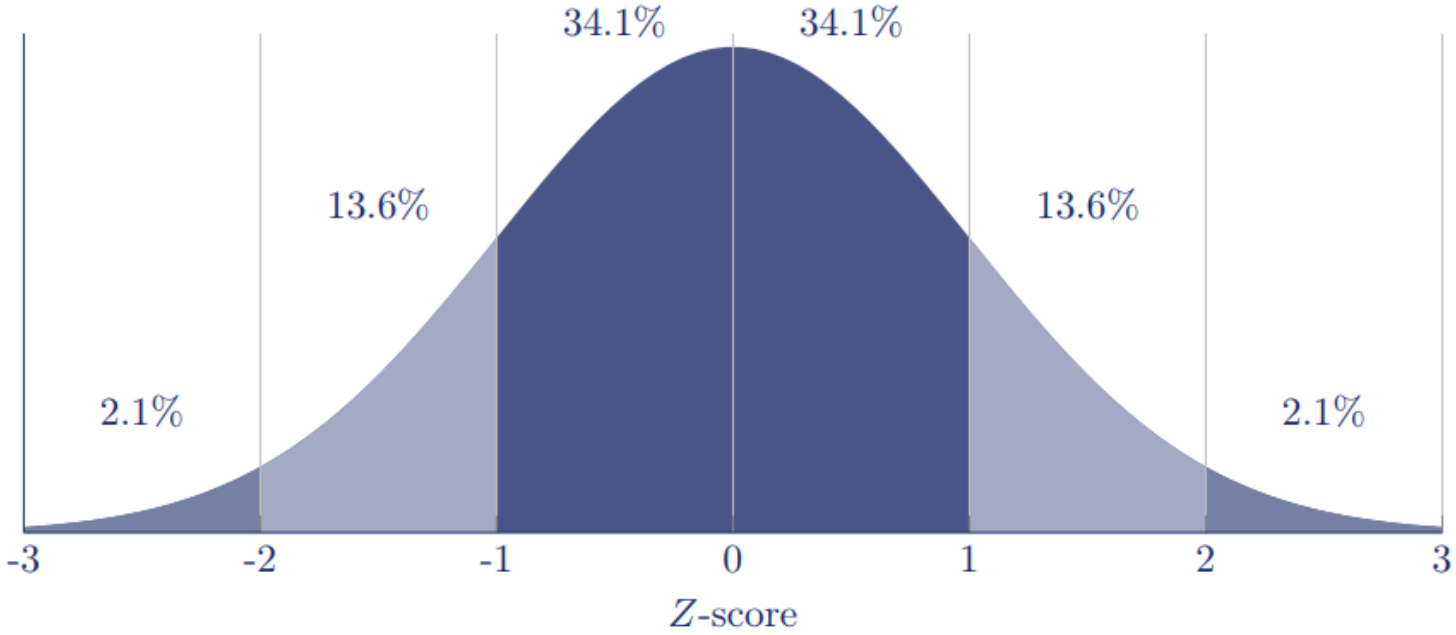
Distribution: boxplot



Distribution: boxplot



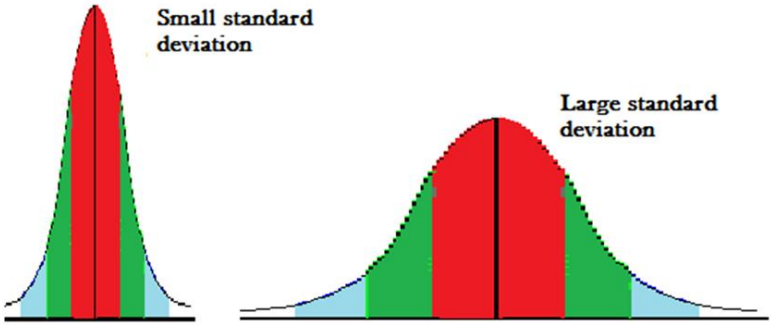
Normal distribution: z-score & standard deviation



About 68% of values 1 standard deviation from mean

About 95% of values 2 standard deviations from mean

Standard deviation



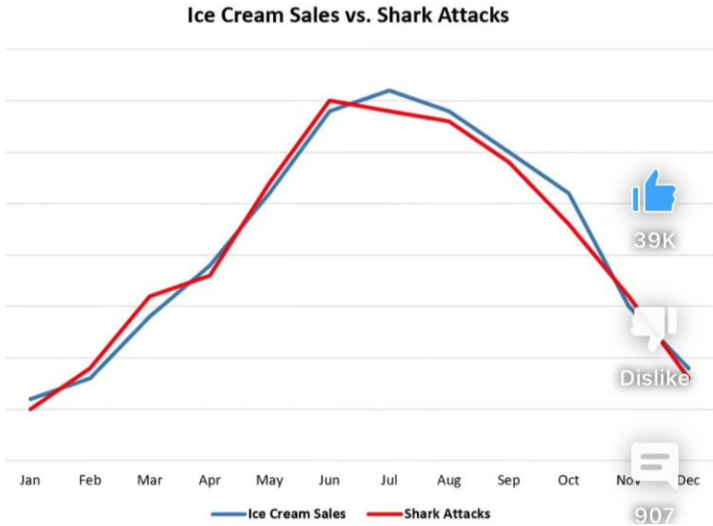
- Population σ or samples *s*
- in same units as the distribution (i.e. the unit of the research material or the normalised version)
- Problem: how do we compare different units?
    - Normalise by dividing through mean: coefficient of variation

Overview

1. Normal distribution
2. **Hypothesis testing**
3. Summary statistical tests

Hypothesis testing?

- Is there statistical association between two (or more) variables?
    - If yes ➔ How strong is that association? Is it significant or not?
- First level of analysis for variables
    - Often only one for nominal variables
    - (ordinal & ratio/interval: [rank] correlation and regression analysis)
- Association ≠ causation!!



Hypothesis testing!

- Null hypothesis (H_o)
  - No association between variables x and y
  - Distribution is random
  - What you want to reject (strawman)
- Alternative hypothesis (H_1)
  - Association between variables x and y
  - Distribution is not random, there is a patterns e.g. normal
  - What you actually want or research

➔ H_o and H_1 must exclude one another: if H_o is true, then H_1 must be false

➔ Together they must cover all possible cases

Hypothesis testing: example

- H_0: "On average, Lisa and Mike produce sentences that have the same length."
- H_1 : "There is a difference in length between the sentences that Lisa and Mike produce on average."
- Non-directional ⇔ directional hypothesis
  - H_0 : "On average, Mike produces *longer* sentences than Lisa."
  - H_1 : "On average, Mike does *not* produce longer sentences than Lisa."
- = Two-tailed ⇔ one-tailed test

Hypothesis testing: workflow

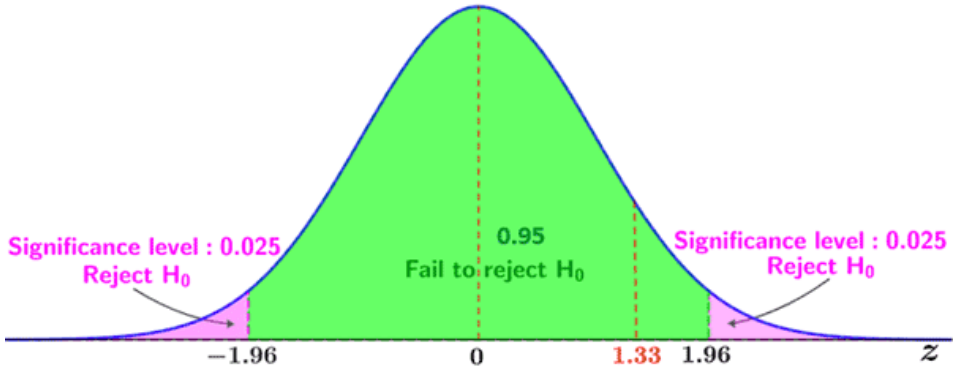
Test statistic

- Chi Square, (rank) correlation, regression…

P-value

- < 0,05 ➔ significant association

Reject/accept H_0/H_1



Test statistics

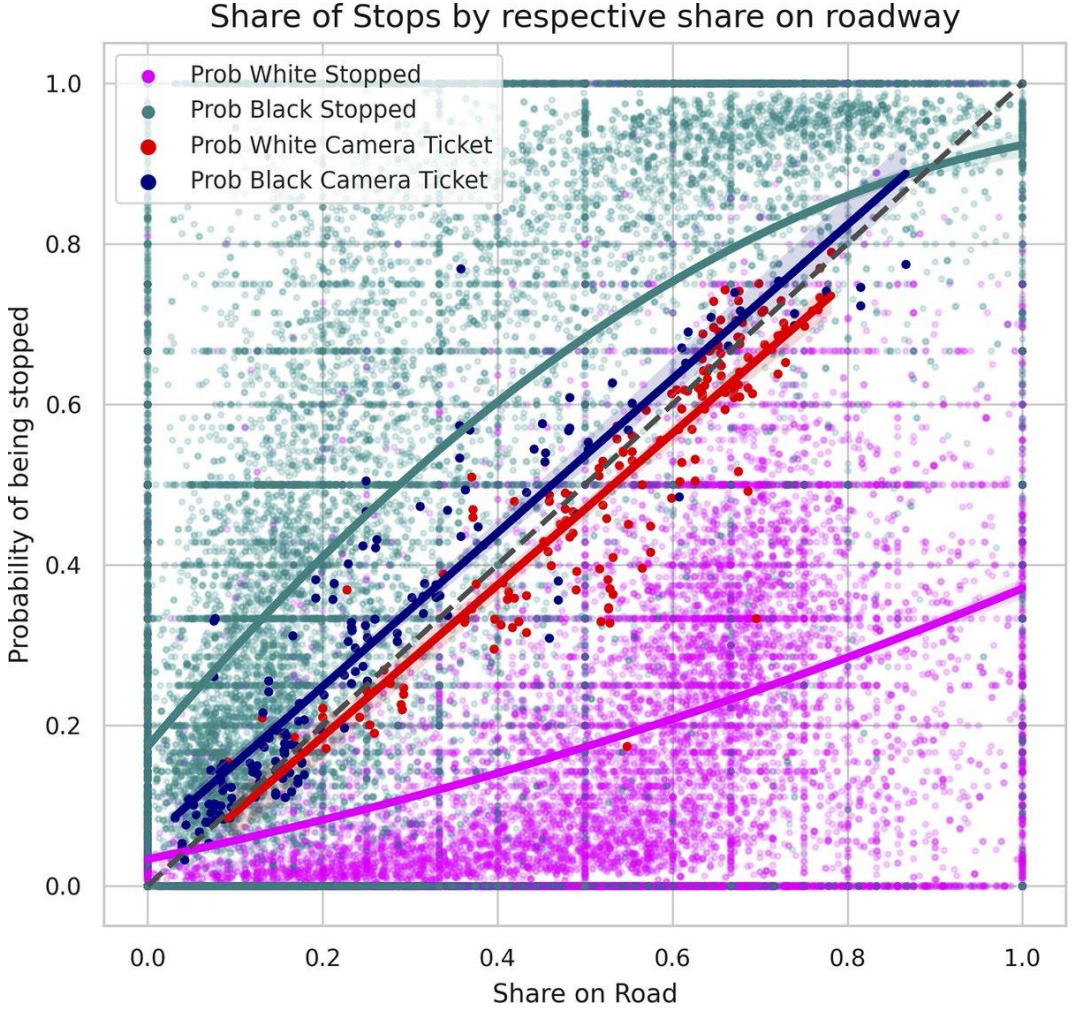
- Brilliant
  - Accept H_o ➔ values are in a known distribution (e.g. normal) ➔ we know everything (or at the very least the mean and standard deviation)
- What if we can't accept H_o ?
  - Fun (i.e. research) begins

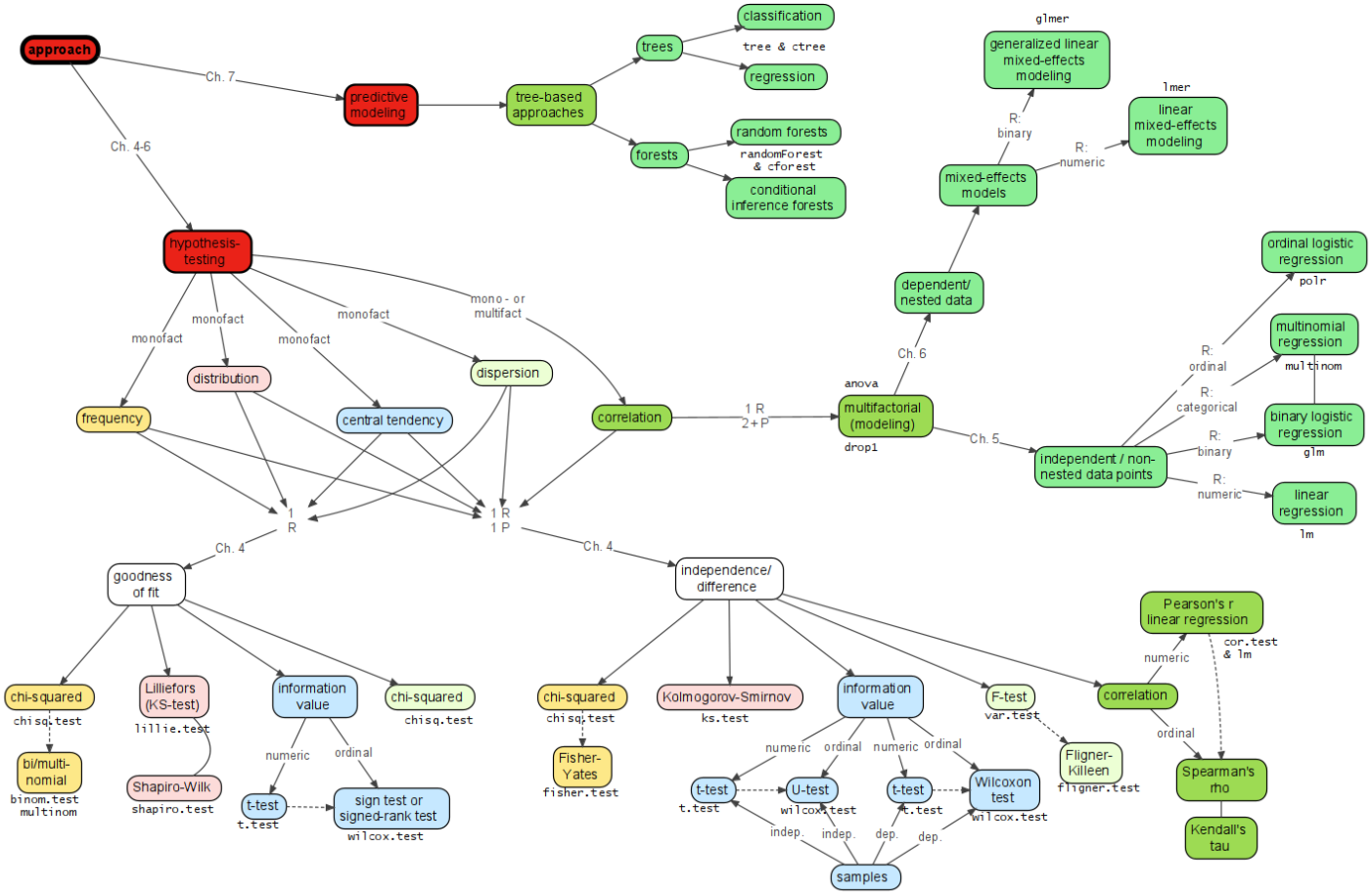
Ethical statistics

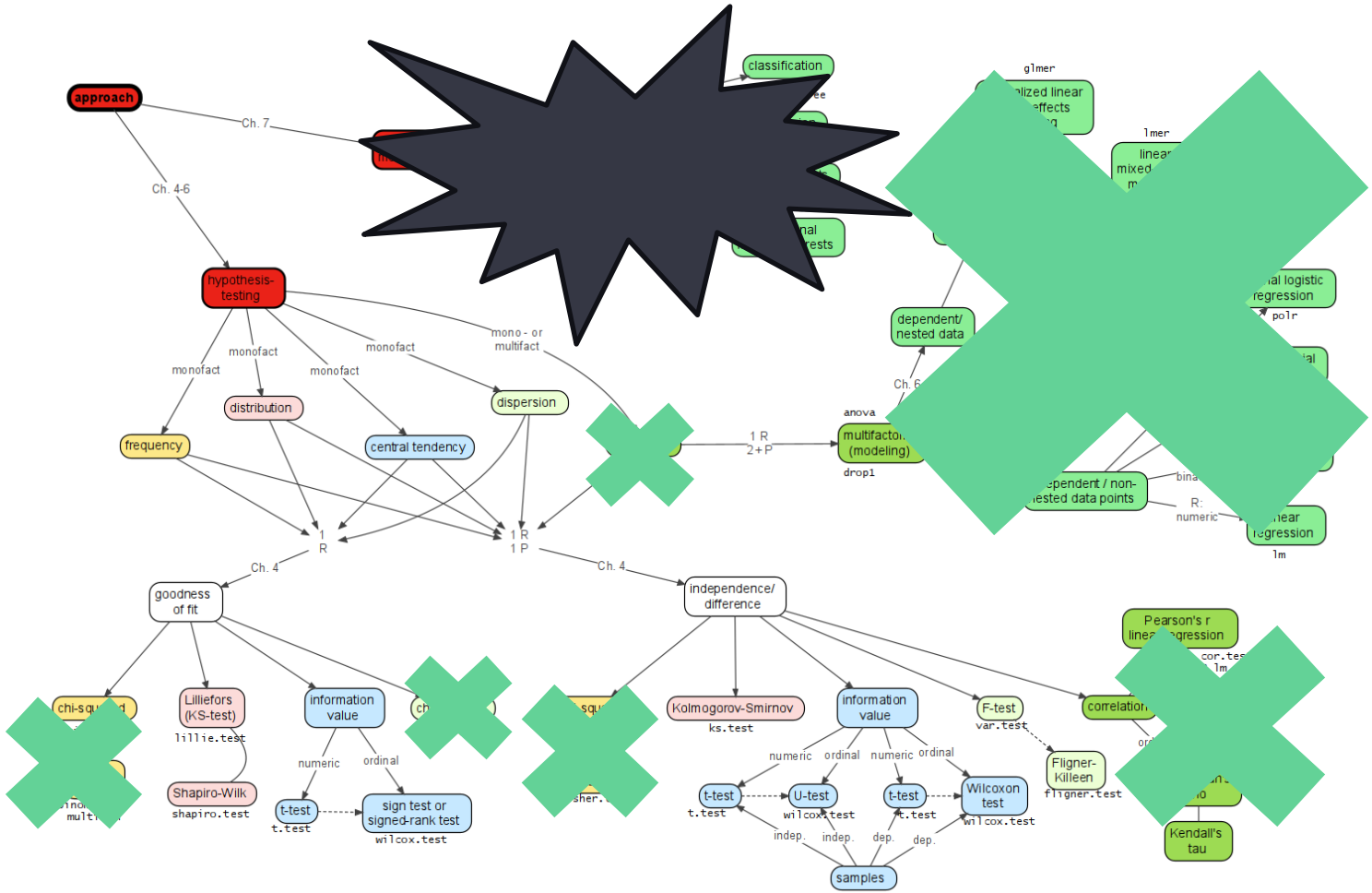
- Cherry picking
  - Only showing result that support your hypothesis
- HARKing
  - *Hypothesizing After Results are Known*
- P-hacking
  - Manipulating results to get a significant result

Overview

1. Normal distribution
2. Hypothesis testing
3. **Summary statistical tests**


Share of Stops by respective share on roadway
Prob White Stopped
Prob Black Stopped
Prob White Camera Ticket
Prob Black Camera Ticket
Probability of being stopped
Share on Road
0.0
0.2
0.4
0.6
0.8
1.0


approach
Ch. 7
predictive modeling
tree-based approaches
trees
classification
tree & ctree
regression
forests
random forests
randomForest & cforest
conditional inference forests
Ch. 4-6
hypothesis-testing
monofact
monofact
monofact
monofact
mono- or multifact
distribution
dispersion
frequency
central tendency
correlation
1 R
2+P
anova
multifactorial (modeling)
drop1
Ch. 6
dependent/ nested data
mixed-effects models
R: binary
glmer
generalized linear mixed-effects modeling
R: numeric
lmer
linear mixed-effects modeling
Ch. 5
independent / non-nested data points
R: ordinal
polr
ordinal logistic regression
R: categorical
multinom
multinomial regression
R: binary
glm
binary logistic regression
R: numeric
lm
linear regression
1 R
1 P
Ch. 4
goodness of fit
chi-squared
chisq.test
bi/multi-nomial
binom.test
multinom
Lilliefors (KS-test)
lillie.test
Shapiro-Wilk
shapiro.test
information value
numeric
ordinal
t-test
t.test
sign test or signed-rank test
wilcox.test
chi-squared
chisq.test
Ch. 4
independence/ difference
chi-squared
chisq.test
Fisher-Yates
fisher.test
Kolmogorov-Smirnov
ks.test
information value
numeric
ordinal
t-test
t.test
U-test
wilcox.test
indep.
samples
numeric
ordinal
t-test
t.test
Wilcoxon test
wilcox.test
dep.
F-test
var.test
Fligner-Killeen
fligner.test
correlation
Pearson's r linear regression
cor.test & lm
numeric
ordinal
Spearman's rho
Kendall's tau


approach
Ch. 7
Ch. 4-6
classification
hypothesis-testing
monofact
monofact
monofact
monofact
mono - or multifact
distribution
dispersion
frequency
central tendency
1 R
2+P
anova
drop1
dependent/ nested data
Ch. 6
glmer
lmer
polr
lm
numeric
1 R
1 P
Ch. 4
goodness of fit
independence/ difference
Lilliefors (KS-test)
lillie.test
Shapiro-Wilk
shapiro.test
information value
numeric
ordinal
t-test
t.test
sign test or signed-rank test
wilcox.test
Kolmogorov-Smirnov
ks.test
information value
numeric
ordinal
numeric
ordinal
t-test
t.test
U-test
wilcox.test
t-test
t.test
Wilcoxon test
wilcox.test
indep.
indep.
dep.
dep.
samples
F-test
var.test
Fligner-Killeen
fligner.test
Pearson's r
cor.test
Kendall's tau

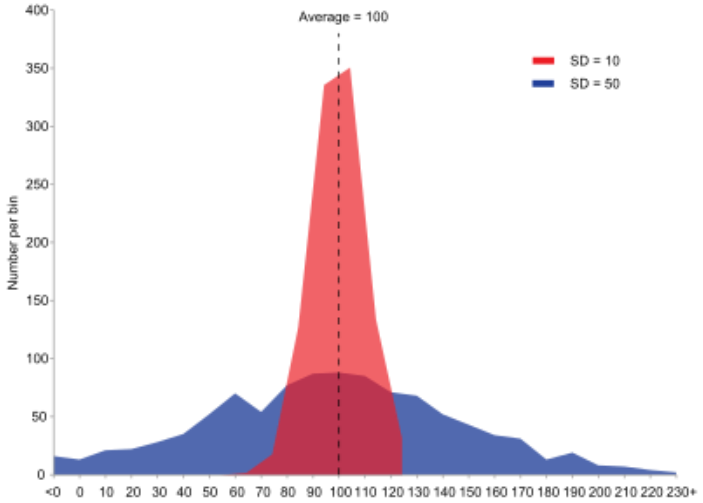
Navigating the flowchart step 1: variables

1. Which kind of variables do you have?
    - Categorical (AKA nominal), with levels
    - Ratio-scaled (doubles)
    - Counts/interval (integers)
    - Ordinal

Navigating the flowchart step 2: amount of variables

2. How many variables do you have?
    - Univariate (just one)
    - Bivariate (exactly two)
        - Bivariate: are your variables **paired** or **unpaired**? (See later slide)
    - (Multivariate: more than two)

Navigating the flowchart step 3: distribution



3. Which distributions (and variance) do your variables have?

- Often, variables need to have a normal distribution (in fact, we'll also cover how you can test for **normality**)
- Often, if you want to compare two variables, tests will require that they have a similar degree of variance
  - Variance = how far a set of numbers is spread out from their average value
  - Usually measured in SD because variance is squared

Navigating the flowchart step 4 & 5: size and question

4. What's the size of your data?
    - Many tests have a critical lower boundary for *n* (often 30)
5. What's your question?
    - Means? distributions? variances?
    - Directional or non-directional?

Navigating the flowchart: terminology

- Monofactorial
  - Involving or controlled by a single factor e.g. time
  - ⇔ multifactorial
- Parametric tests
  - Based on a fixed set of parameters e.g. normal distribution
  - ⇔ non-parametric tests
- Paired vs. unpaired samples (dependent /independent)
  - Paired samples have the same size because one (or more) value “pairs” them e.g. taxes of B. Roccoli in 1602 and 1632

Navigating the flowchart: for all test statistics

1. Pick (at least) one suitable test, e.g. using the flowchart
2. Check the test's requirements (i.e. whether you're allowed to apply it)
    - Goodness-of-fit e.g. check normality with Shapiro-Wilk test for up to 5000 data points or Kolmogorov-Smirnov
    - Independence e.g. F-test for homogeneity of variances (one-tailed vs. two-tailed)
    - Difference e.g. T-test (or Wilcoxon) for difference in measures of central tendency
3. Apply it: get the statistic, p-value, and preferably some other things as well (e.g. confidence interval)
4. Report it in proper prose, explicitly adding all relevant numbers