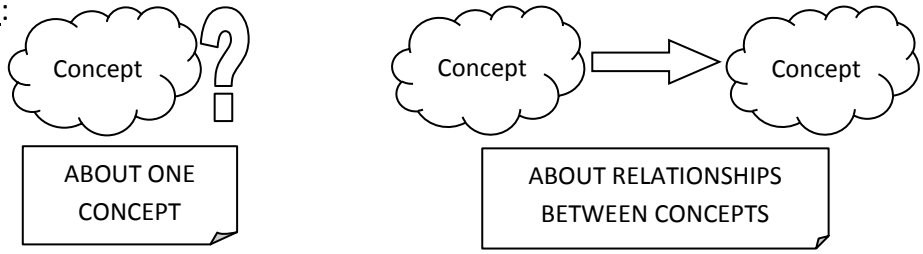
Turning a **research question** into a **statistical question**.

ORIGINAL QUESTION:



TYPE OF QUESTION:

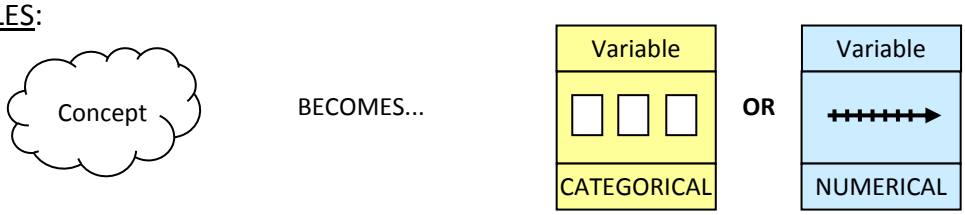
DESCRIBE – what's going on?

DECIDE – yes or no?

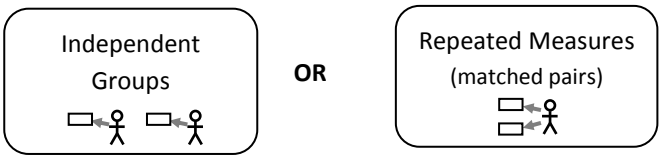
ESTIMATE – what's this number?

PREDICT/EXPLAIN – what's the formula?

TYPES OF VARIABLES:



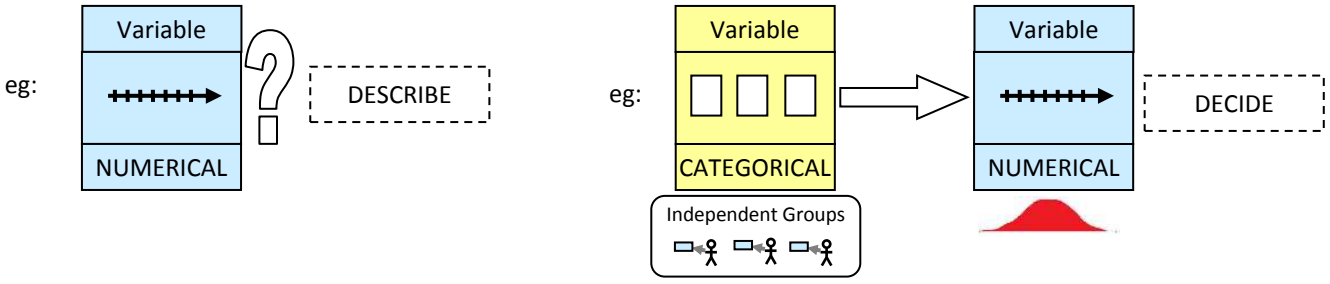
WHAT EXPLANATORY CATEGORICAL VARIABLES DEFINE:



DISTRIBUTION OF OUTCOME NUMERICAL VARIABLE:

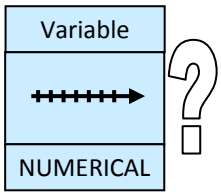
Note: This probably doesn't matter if you have a lot of data.

STATISTICAL QUESTION:

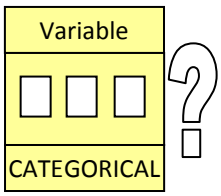


Note: In the list below, the outcome variables are usually assumed to be normal.

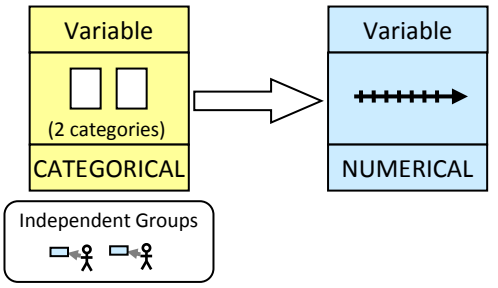
Statistical methods for statistical questions



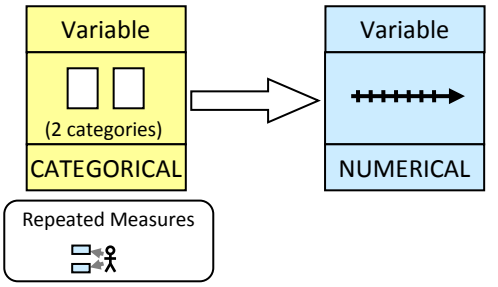
DESCRIBE: Numbers: Mean & standard deviation (median & IQR)
Graphs: Histogram / Boxplot.
DECIDE: "Is the mean equal to #?" – one sample t-test.
"Is the median equal to #?" – sign test.
ESTIMATE: "What is the mean?" – confidence interval for a mean.



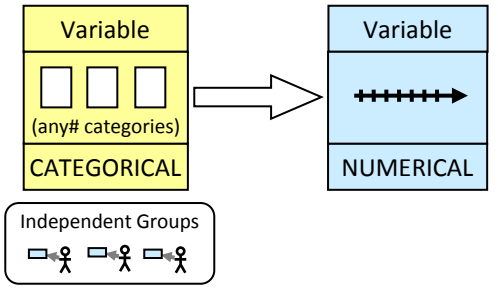
DESCRIBE: Numbers: Table of percentages or proportions.
Graphs: Bar graph showing percentages.
DECIDE: "Is this percentage equal to #?" – z-test for a single proportion.
"Are percentages distributed according to #, #, #?" – chi-squared test for goodness of fit.
ESTIMATE: "What is this percentage?" – confidence interval for a proportion.



DESCRIBE: Numbers: Means & standard deviations for each group (medians & IQRs for each category).
Graphs: Histograms on same scale / side-by-side boxplots.
DECIDE: "Are the means equal?" – unpaired t-test (Mann-Whitney U-test or Wilcoxon rank-sum test).
ESTIMATE: "What is the difference between the means?" – confidence interval for the difference in means.

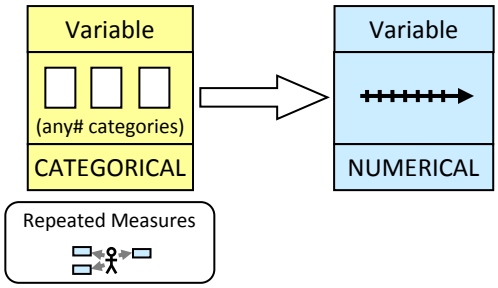


DESCRIBE: Numbers: Mean & standard deviation of differences between measurements.
Graphs: Histogram of the differences between measurements.
DECIDE: "Is there a mean difference?" – paired t-test (Wilcoxon signed ranks test).
ESTIMATE: "What is the mean difference?" – confidence interval for the mean difference.



DESCRIBE: Numbers: Mean & standard deviation of each group.
Graphs: Histograms/boxplots on the same scale. Bar graph showing mean of each group.
DECIDE: "Are the means equal?" – one-way analysis of variance ANOVA with post-hoc t-tests (Kruskal-Wallis test).
ESTIMATE: "What are the differences between means?" – confidence intervals for each difference in means.

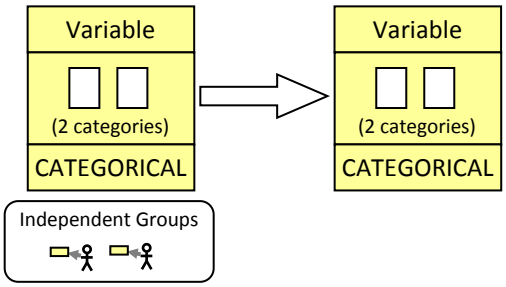
Statistical methods for statistical questions



DESCRIBE: Graphs: Line graph for each subject showing changing value of variable.

DECIDE: "On average, does the value change for each person across categories?" – repeated measures ANOVA with post-hoc paired t-tests or mixed effects regression.

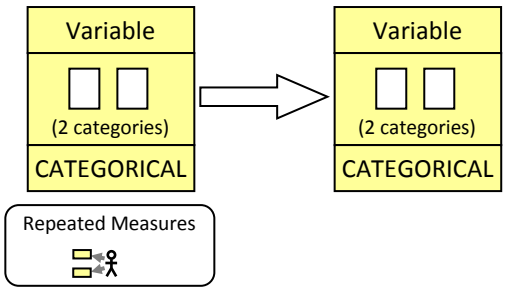
ESTIMATE: "What are the mean differences between categories?" – confidence intervals for mean differences.



DESCRIBE: Numbers: Two-way table of counts or %s. Odds ratios.
Graphs: Histogram for each explanatory category.

DECIDE: "Is the outcome just as likely for both explanatory categories?", "Are the two variables associated?" – chi-squared test for independence (or Fisher's exact test).

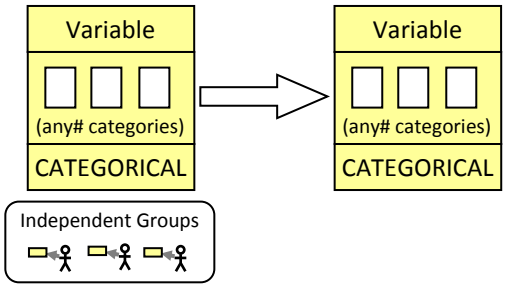
ESTIMATE: "How much more likely is the outcome in this category?"– confidence interval for difference in proportions, confidence interval for odds ratio.



DESCRIBE: Numbers: Two-way table of counts or %s.
Graphs: Bar graph for each explanatory category.

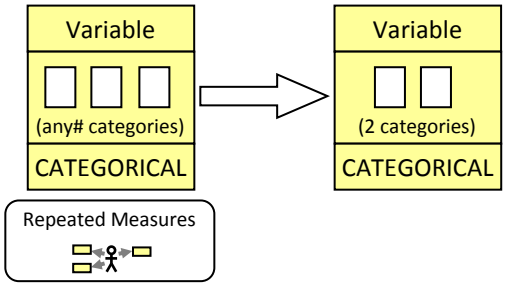
DECIDE: "Is the outcome just as likely for both explanatory categories?" – McNemar's test.

ESTIMATE: "How much more likely is the outcome in one category compared to the other?"– confidence interval for difference in proportions.



DESCRIBE: Numbers: Two-way table of counts or %.
Graphs: Histogram for each explanatory category.

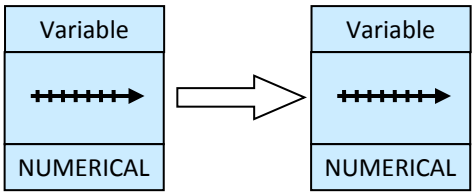
DECIDE: "Do the percentages in the outcome change across the explanatory categories?", "Are the two variables associated?" – chi-squared test for independence.



DESCRIBE: Numbers: Two-way table of counts or %.
Graphs: Histogram for each explanatory category.

DECIDE: "Do the percentages in the outcome change across the explanatory categories?", "Are the two variables associated?" – Cochrane's Q-test.

Statistical methods for statistical questions



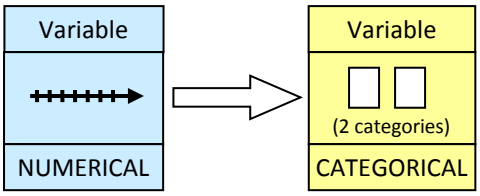
DESCRIBE: Numbers: Correlation coefficient (R)
Graphs: Scatterplot.

DECIDE: "Does a relationship exist?" – linear regression: t-test on coefficient.

ESTIMATE: "How much does the output variable change when the explanatory variable changes?" – linear regression: confidence interval for slope.

PREDICT: "How can you calculate the output knowing the explanatory variable?" – linear regression formula: $y = \beta_0 + \beta_1 x$.

NOTE: May need to do a nonlinear regression if the scatterplot indicates a curved sort of relationship.

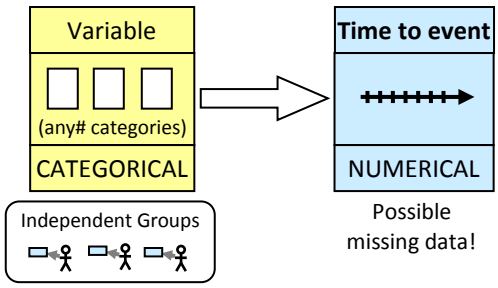


DESCRIBE: Numbers: Mean & standard deviation for each category of the outcome.
Graphs: Histograms/boxplots on the same scale.

DECIDE: "Does the numerical variable have an effect on the chances of the outcome?" – unpaired t-test using the outcome to define the two groups.

ESTIMATE: "How much does a change in the numerical variable affect the chances of the outcome?" – logistic regression: confidence interval for odds ratio.

PREDICT: "How can you calculate the chances of the outcome knowing the value of the explanatory variable?" – logistic regression formula: $\log(\text{odds of } y) = \beta_0 + \beta_1 x$.

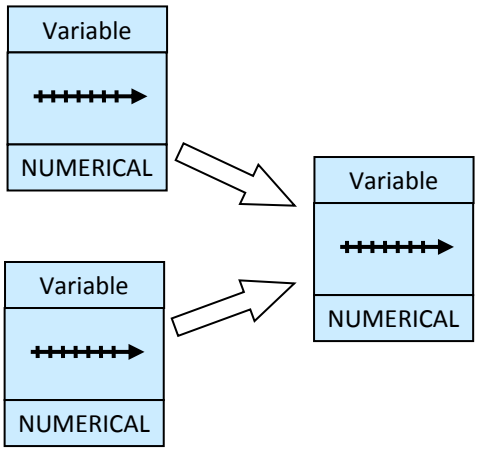


DESCRIBE: Numbers: Proportion reaching event at certain time (eg 5-year survival), median times to reach event.
Graphs: Kaplan-Meier curve showing survival percentages.

DECIDE: "Is the time to reach the event the same in all groups?" – survival analysis: log-rank test.

ESTIMATE: "What is the difference in proportions reaching the end point at this particular time?" – confidence interval for the difference in proportions.
"How much more at risk of the event is this group than this group?" – Cox regression: confidence interval for relative hazard.

Statistical methods for statistical questions



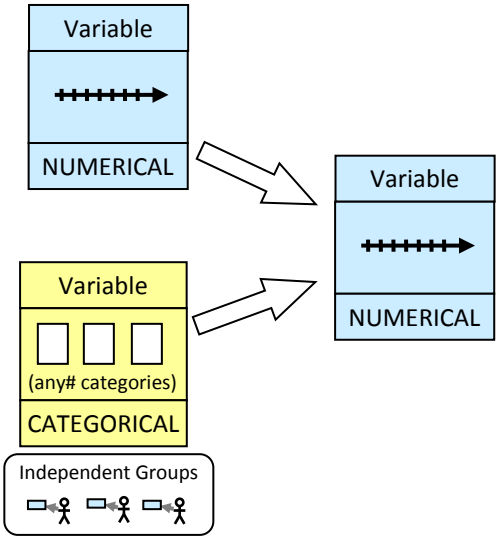
DESCRIBE: Graphs: Scatterplot for each explanatory variable with the outcome variable.
Numbers: multiple linear regression: R^2 value

DECIDE: "Does a relationship exist with any of the variables at all?" – multiple linear regression: F-test.
"Does a relationship exist with *this* varable, taking into account the others?" – multiple linear regression: t-test on one coefficient.

ESTIMATE: "How much does the output variable change when *this* explanatory variable changes?" – multiple linear regression: confidence interval for one slope.

PREDICT: "How can you calculate the output knowing the explanatory variables?" – multiple linear regression formula: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

NOTE: This can be done for many explanatory variables.



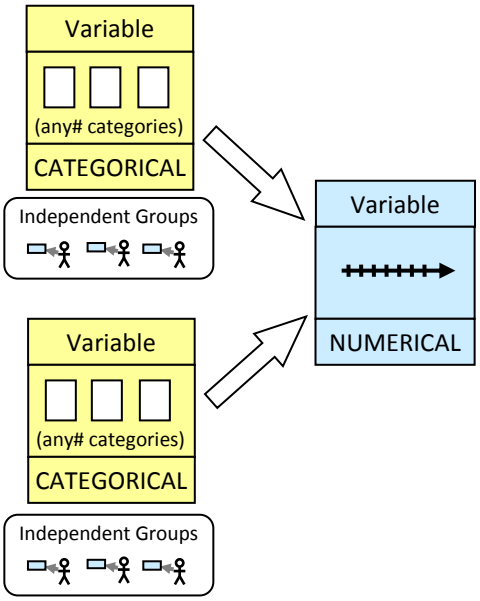
DESCRIBE: Graphs: Scatterplot of both numerical variables for each category.
Numbers: multiple regression: R^2 value

DECIDE: See above for multiple regression.

ESTIMATE: See above for multiple regression.

PREDICT: See above for multiple regression.

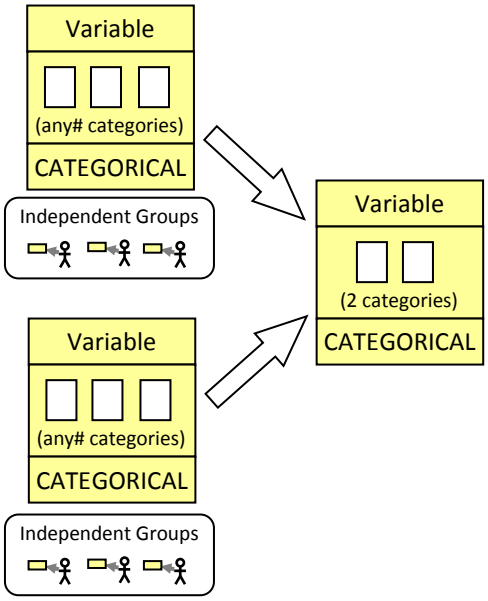
NOTE: This can be done for many explanatory variables of both types.



DESCRIBE: Graphs: Histogram for each combination of explanatory categories. Line graph showing mean of each group.

DECIDE: "Does a relationship exist with any of the variables at all?" – *two-way* ANOVA: F-test.
"Does a relationship exist with *this* varable, taking into account the others?" – two-way ANOVA: F-test for one effect.
Note: both can also answered with multiple regression (see above).

PREDICT: "How can you calculate the output knowing the explanatory variables?" – multiple linear regression formula: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.



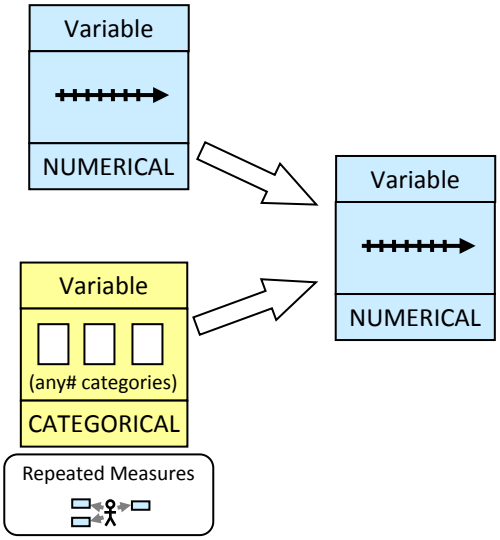
DESCRIBE: Graphs: Histogram for each combination of explanatory categories.

DECIDE: "Does a relationship exist with any of the variables at all?" – multiple logistic regression: chi-squared test for covariates.
"Does a relationship exist with *this* varable, taking into account the others?" – multiple logistic regression: Wald test.

ESTIMATE: "How much does the chance of the outcome change when *this* explanatory variable changes?" – multiple logistic regression: confidence interval for odds ratio.

PREDICT: "How can you calculate the chances of the outcome knowing the explanatory variables?" – multiple logistic regression formula: $\log(\text{odds of } y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

NOTE: This can be done with many explanatory variables – even if some of them are numerical.



DESCRIBE: Numbers: multiple linear regression: R^2 value

DECIDE: "Does a relationship exist with any of the variables at all?" – mixed effects regression: F-test.
"Does a relationship exist with *this* varable, taking into account the others?" – mixed effects linear regression: t-test on one coefficient.

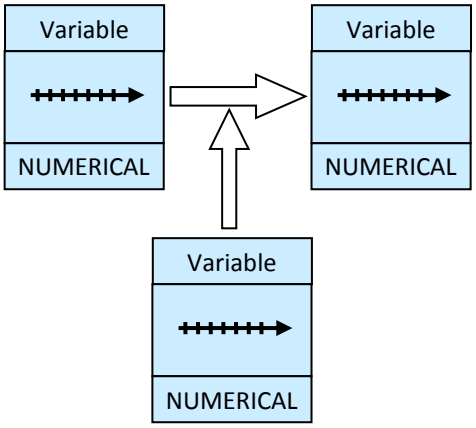
ESTIMATE: "How much does the output variable change when *this* explanatory variable changes?" – mixed effects regression: confidence interval for one coefficient.

PREDICT: "How can you calculate the output knowing the explanatory variables?" – mixed effects regression formula.

NOTE: "mixed effects" may also be called "random effects".

NOTE: This can be done for many explanatory variables, of both types, and with a mixture of repeated-measures and independent-groups

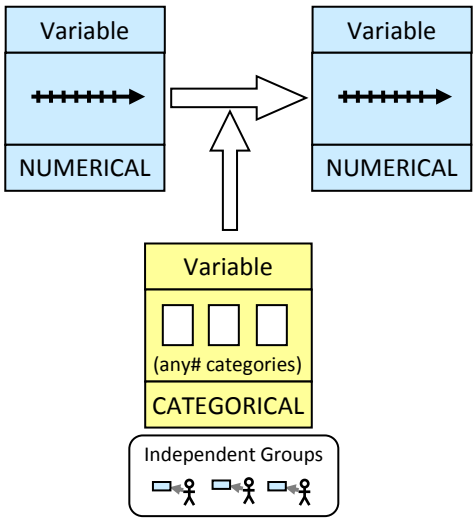
---



DECIDE: "Does one variable change the way the other affects the outcome?"– multiple linear regression: t-test on the interaction effect.

ESTIMATE: "How much does the second variable change the effect of the first on the outcome?"– multiple linear regression: confidence interval for the interaction effect.

PREDICT: "How can you calculate the output knowing the explanatory variables?" – multiple linear regression formula: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$.



DESCRIBE: Graphs: Scatterplot for each category, showing line of best fit in each case.

DECIDE: "Does one variable change the way the other affects the outcome?"– Analysis of Covariance (ANCOVA) / multiple linear regression: t-test on the interaction effect.

ESTIMATE: "How much does the second variable change the effect of the first on the outcome?"– multiple linear regression: confidence interval for the interaction effect.

PREDICT: "How can you calculate the output knowing the explanatory variables?" – multiple linear regression formula: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$.

NOTE: This can be done for many explanatory variables of both types. ANCOVA refers specifically to the case where the interaction variable is categorical.

NOTE: There are many other methods dealing with more specific and difficult questions including (but definitely not limited to):

- "Does this variable affect the *variance* of the outcome?"
  → F-test for two variances
- "Do these variables affect this categorical outcome (which has several categories)?"
  → Multinomial regression
- "Does the data come from a normal distribution?"
  → Investigate normal quantile-quantile plot; Shapiro-Wilk test
- "To what degree do these two measuring systems agree?"
  → Intraclass correlation coefficient
- "What is the best cut-off for this measurement in order to say someone needs medical attention?"
  → ROC analysis
- "Do all these measurements vary together so that they could be considered as measuring some smaller number of underlying concepts?"
  → Factor analysis / Principal Component Analysis
- "Can the subjects be grouped into a few similar groups based on the similarity in their measurements?"
  → Cluster analysis
- and so on ...