# STAT 301 - OVERVIEW OF STATISTICAL PROCEDURES

# Selecting the inferential procedure: You should start with 3 questions

1. Does the research question need a confidence interval (estimation) or a test of significance (decision)?

2. Is the research question dealing with a quantitative response or a categorical response?

3. How many (independent) samples do I have: Am I comparing two groups or analyzing one group (including matched pairs)?

|  |  |  |
| --- | --- | --- |
| One sample  | Quantitative (Ch. 2) | Categorical (Ch. 1) |
| *Descriptive statistics* |
| Graphical summary | Histogram, boxplot, dotplot | Bar graph |
| Numerical summary | *n*, , *s*, *median, IQR* | *n* , X, =X/*n* |
| *Inference statistics* |
| Null hypothesis | H0:  = 00 = hypothesized population mean or process mean | H0:  = 0 0 = hypothesized population proportion or process probability |
| Simulation | * *RS:* Sample from hypothetical population or bootstrapping
* *MP:* Flip coin for each pair to determine sign of difference
 | * *RS:* Spinner (with probability 0) for each observational unit (large population or random process)
* *MP:* Flip coin to change order
 |
| Simulation applet | * Sampling from Finite Pop
* Matched Pairs Randomization
 | * One Proportion Inference
 |
| Exact probability model |  | Binomial*MP:* McNemar’s Test |
| Theory-based approach  | One sample *t* procedures | One sample *z* procedures |
|  Validity check | Normal population or *n* ≥ 30 | Wald ci: *n*>10, *n*(1-)>10 tos: *n*0 and *n*(1-0)> 10 |
| Test Statistic |  df = *n* – 1  | Also consider continuity correction |
| Confidence interval | for :   | for :  + *z*\*Plus Four: $\tilde{p}\pm 1.96 \sqrt{\frac{\tilde{p}\left(1-\tilde{p}\right)}{n+4}}$ (95%) |
| R | * *t.test( … paired = TRUE…)*
* *iscamonesamplet*
 | * *iscambinomtest*
* *iscamonepropztest*
 |
|  Applet | * TBI > One Mean
 | * One Proportion Inference applet
* TBI > One Proportion
 |
| Prediction interval | With normal population (*t*\* from *t* probability calculator applet) |  |

*Bootstrapping* is resampling with replacement from the observed data. This can be done in practice, with or without assuming the null is true (vs. making up populations to sample from to learn the behavior of the statistic).

|  |  |  |
| --- | --- | --- |
| Two independent samples or Randomized experiment | Comparing Two Means (Ch. 4) | Comparing Two Proportions (Ch. 3) |
| *Descriptive statistics* |
| Graphical summary | As above but on same scale | Segmented bar graph for each group (all bars 0-100%) or Mosaic plot |
| Numerical summary | , , *s*1, *s*2, *n*1, *n*2 |  -  or / (rel risk), ot  (odds ratio) |
| *Inferential statistics* |
| Null hypothesis | H0: 1 - 2 = 0 | H0: -2 = 0 or /2 = 1 or t = 1 |
| Simulation | *RA:* Index cards with responsevalues*RS:* two-sample bootstrapping (pooled for test, unpooled for CI) | *RA:* Index cards color-coded for success and failures*RS:* Independent binomial sampling with same probability of success |
| Simulation applet | Comparing Groups (Quant)Sampling from Two Populations | Analyzing Two-way TablesSampling from Two Processes |
| Exact probability model (rand. experiment) | All possible random assignments (Inv 4.1) | Fisher’s Exact Test (hypergeometric) |
| Theory-based approach | Two sample *t* procedure | Two sample *z* procedure |
|  Sample size check | Normal populationsor *n*1, *n*2 ≥ 20 | At least 5 successes and 5 failures in each sample |
| Test Statistic | (unpooled)approx df = min(*n*1 – 1, *n*2 – 1) |  = (total # of successes)/(*n*1+*n*2) |
| Confidence Interval | approx df=min(*n*1 – 1, *n*2 – 1) | Wilson adjustment for 95% CI (+1 in each cell) |
|  |
| R | *t.test( … var.equal=FALSE..)**iscamtwosamplet* | *fisher.test (two‐way table, nrow=2)**iscamtwopropztest* |
| TBI Applet | Two means | Two Proportions |

*RS* = random sampling; *RA* = random assignment; *MP* = matched pairs

**Note:** With skewed quantitative data, can also consider transformations or randomization tests involving

other statistics like medians.

**Statistical Investigation Process**

1. Formulate research question

2. Design data collection strategies

3. Collect and clean data

4. Exploratory data analysis

5. Statistical inference (see table for common inference procedures)

 Significance, Estimation

6. Draw Conclusions

Generalizability? Cause-and-effect?

7. Reformulate research question

# Methods of Analyses

|  |  |
| --- | --- |
| **Explanatory Variables** | **Response Variable (Variable of Interest)** |
| **1 Quantitative****(Normal errors)** | **1 Categorical****(Binary)** | **1 Categorical****(3+ Categories)** |
| **None** | One sample t (301)Paired t (301) | One sample *z* (301)*Chi-square goodness of fit* (302) | *Chi-square goodness of fit* (302) |
| **1 Quantitative** | *Simple linear regression* (302, 334) | Logistic Regression (334, 418) | Nominal logistic regression (418) |
| **1 Categorical****(Binary, 2 Groups)** | Two sample *t* (301)*One-way ANOVA* (302) | Two sample *z*, Fisher’s Exact Test (301)*Chi-square* *test* (302) | *Chi-square* *test* (302) |
| **1 Categorical****(3+ Categories/Groups)** | *One-way ANOVA* (302, 323) | *Chi-square* *test* (302) | *Chi-square* *test* (302) |
| **2+ Quantitative or Categorical variables** | *Multiple regression* (302, 323, 334) | Logistic Regression (334, 418) | Nominal logistic regression (418) |

**NOTE:**

* This is not an exhaustive list of methods; these are some of the methods you should have seen so far.
* There are some exceptions, but this provides some organization to the choice of method

**Future Courses:**

SAS Programming Language – Stat 330

R – Stat 331

Correlated observations (Dependent observations) – Stat 414

Correlated observations (Time Series data) – Stat 416

Time-to-event response/censored data (Survival analysis) – Stat 417

General Linear Model (Categorical data) – Stat 418

More than two quantitative response variables (Multivariate analysis) - Stat 419