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 or a test of significance?
2. Is the question dealing with a mean (quantitative response) or a proportion (categorical response)?
3. How many (independent) populations do I have: Am I comparing two groups or analyzing one group?

One sample	**Quantitative (Ch. 2)**	**Categorical (Ch. 1)**
Graphical summary	Histogram, boxplot, dotplot	Bar graph
Numerical summary	*n*, $\bar{x}$, *s*, *median, IQR*	$\hat{p}$, *n*
Null hypothesis	H_0: $\mu = \mu_0$ μ_0 = hypothesized mean (or population median)	H_0: $\pi = \pi_0$ π_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 π_0) for each observational unit (large population or 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* procedure	One sample *z* procedure Wald adjustment for 95% CI
Validity check	Normal population or $n \geq 30$	Wald ci: $n\hat{p} \geq 10$, $n(1-\hat{p}) \geq 10$ tos: $n\pi_0$ and $n(1-\pi_0) \geq 10$ Adjusted Wald: $n \geq 5$
Test Statistic	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ df=$n-1$	$z = \frac{\hat{p}-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$
Confidence interval	for μ: $\bar{x} \pm (t^*_{n-1})(s/\sqrt{n})$	for π: $\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$ Adjusted: $\tilde{p} \pm 1.96\sqrt{\tilde{p}(1-\tilde{p})/(n+4)}$
JMP	• *Journal File* • *Analyze > Distribution (Test Mean, Confidence Interval)* • *Analyze > Matched Pairs*	• *Journal file (Hypothesis Test One Proportion, Confidence Interval One Proportion)* • *Analyze > Distribution**
R	• *t.test(… paired = TRUE…)*	• *iscambinomtest* • *iscamonepropztest*
Applet	• TBI > One Mean	• One Proportion Inference applet • TBI > One Proportion
Prediction interval	$\bar{x} \pm (t^*_{n-1})s\sqrt{1+1/n}$ With normal population (*t** from *t* distribution in applet) *JMP: (Prediction Interval)*	

*With one categorical variable, Analyze > Distribution assumes a binomial p-value for one-sided and actually the theory-based p-value for two-sided.

Two independent samples or Randomized expt	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%)
Numerical summary	$\bar{x}_1$, $\bar{x}_2$, s_1, s_2, n_1, n_2	$\hat{p}_1$ - $\hat{p}_2$ or $\hat{p}_1 / \hat{p}_2$ (rel risk), $\hat{\tau}$ (odds ratio)
	Inferential Statistics	
Null hypothesis	H_0: $\mu_1 - \mu_2 = 0$	H_0: $\pi_1 - \pi_2 = 0$ or $\pi_1 / \pi_2 = 1$
Simulation	*RA:* Index cards with response values *RS:* bootstrapping	*RA:* Index cards color-coded for success and failures *RS:* Independent binomial sampling with same probability of success
Simulation applet	Comparing Groups (Quant)	Analyzing Two-way Tables
Exact probability model	All possible random assignments (Inv 4.1)	Fisher's Exact Test (hypergeometric)
Theory-based approach	Two sample *t* procedure	Two sample *z* procedure Wilson adjustment for 95% CI
Sample size check	normal populations or $n_1, n_2 \geq 20$	At least 5 successes and 5 failures in each sample
Test Statistic	$t = \dfrac{\bar{x}_1 - \bar{x}_2 - hypoth\,diff}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ (unpooled) approx df = min($n_1 - 1$, $n_2 - 1$)	$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ $\hat{p}$ = (total # of successes)/(n_1+n_2)
Confidence Interval	$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ approx df=min($n_1 - 1$, $n_2 - 1$)	$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ $\exp[\ln(\hat{p}_1 / \hat{p}_2) \pm z^* \sqrt{\dfrac{1}{a} - \dfrac{1}{a+c} + \dfrac{1}{b} - \dfrac{1}{b+d}}]$
JMP	*Analyze > Fit Y by X (Test Mean, Confidence Interval)* *Journal File (Two Means)*	*Analyze > Fit Y by X (Test Mean, Confidence Interval)* *Journal File (Two proportions)*
R	*t.test(… var.equal=FALSE..)* *iscamtwosamplet*	*fisher.test (two-way table, nrow=2)* *iscamtwopropztest*
TBI Applet	Two means	Two Proportions

Note: With skewed quantitative data, can also consider transformations or randomization tests involving other statistics like medians.

RS = random sampling
RA = random assignment
MP = matched pairs
Bootstrapping is resampling with replacement from the observed data. This can be done in practice, with or without assuming the null is true (vs. our explorations where we made up populations to sample from to learn the behavior of the statistic).

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
    - Generalizability?
    - Cause-and-effect?
6. 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, 324)	Logistic Regression (324, 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	*Multiple regression* (302, 324)	Logistic Regression (324)	Nominal logistic regression (418)
2+ Categorical	*Multi-way ANOVA* (302, 323)	Logistic Regression (324)	Nominal logistic regression (418)
Both Categorical and Quantitative variables	*Multiple regression/ ANCOVA* (302, 323)	Logistic Regression (324)	Nominal logistic regression (418)

NOTE:

- This is <u>not</u> 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:

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