|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Week 12 | Extra Space  - analysis of variance (One-Way ANOVA)   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | DF | Sum of squares | Mean square | F ratio | P-value | | Group | K-1 (K is # groups) | SSGroup | SSGroup/ DFGroup | MSGroup / MSE |  | | Error | N-K | SSE | SSE / DFE |  |  | | Total | N-1 = DFG + DFE | SSTotal |  |  |  |   - conclusion: “the sample wasn’t randomly selected, so we aren’t positive on what the actual pop is”  - One-Way ANOVA F-test answers: Does at least one group mean differ from the rest?  - interpret logistic curve graph (ex): the pred prob of being accepted incr. as aptitude incr. for men + women. Women have slightly high pred prob of being accepted than men for each indiv aptitude score  - >0.1-little to no,>0.05-weak, >0.025-moderate, >0.001-strong,<0.001-overwhelming | Apr 13  - quantitative vars notation: 𝜇­1­ (mean var of interest in pop1), n1 (sample size of pop1), x̄ (mean of var of interest in sample 1), s1 (stand.dev. of var of interest in sample 1)  - parameter of interest: 𝜇­1 - 𝜇­2, statistic of interest: x̄1 - x̄2  - conditions (ind 2-sample data): (1) rand. sample for both groups, (2) normality cond. in BOTH groups [normal pops OR large sample size ~30], (3) independent groups  - compare means for each group to overall mean  - ANOVA: obs coms in groups – (1) all obs in group have same pred value, (2) see this in residual plot as vertical stripes of dots, (3) one stripe for each group, (4) residual vs predicted value plot still good thing  - null model: all groups have same population mean  - full model: all groups potentially have different means- not an experiment, so no random assignment. We can only make cause and effect statements when we have a study with random assignment. |
| Week 13 | Apr 18  - 𝜇 = pop grand mean, ȳ = sample grand mean, 𝜇­­k = group pop means, ȳk = group sample means, 𝛼­k = pop effects, 𝛼­k^ = estimated effects (𝛼­k^ = ȳk - ȳ), 𝜖 = pop errors (pop residuals), 𝜖^ = estimated errors (sample resid) (𝜖^ = yi - ȳk)  - pop/theoreti model: 𝑌 = 𝜇 + 𝛼­k + 𝜖 where 𝜖~N(0, σ)  - hypothesis test: H0: 𝜇­1­ = 𝜇­­2 = … 𝜇­k. Ha: at least one equality doesn’t hold  - if null hypothesis is true (no group effect), model becomes: 𝑌 = 𝜇 + 𝜖  - conditions: (1) samples are rand selected indep from the K groups [errors are indep], (2) sampled groups have distributions that are approx. normal or each sample size is ≥ 30 [residuals ~ normal], (3) pop variances are equal [res plot or smax/smin≤2] (std dev) | Apr 20    - pred odds of being accepted for males, for females:  - males: 0.6364 / 0.3636 = 1.75  - females: 0.3333 / 0.6667 = 0.5  - predicted prob of being accepted for males, for females:  - males: 0.6364, females: 0.3333 |
| Week 14 | Apr 25  - marginal proportions: one var by itself  - odds: p̂1 / (1- p̂1) = prob success / prob failure  - ex) odds: success is 0.25 times as likely as failure  - odds < 1, failure is more likely than a success  - how much more likely a failure is: 1/odds of success (1/0.25=4. A failure is 4 times as likely as a success).  - if one event doesn’t affect another, P(A if B)=P(A)  - Mosaic plot: graph of conditional probabilities  - Chi-square χ­2 test: hyp test to evaluate independenc  - H0: the 2 categorical vars are independent  - H0: the conditional probabilities are the same for all levels of the 2nd categorical variable  - H0: all conditional probs equal the marginal prob  - Ha: are dep/are not all the same/at least one differs  - conditions: all expected counts ≥ 5, random sample  - expected count = (row tot \* column tot) / grand tot  - logit regres: use x vars to pred probs for a cat y var  - OLS is a special case of the generalized linear model  - can’t use LSR b/c assumptions will be violated and give preds not between 0-1 when predicting prob of success | Apr 27  - P = sample proportion of successes  - Q = 1-P = sample proportion of failures  - 𝜋­­i = prob of being a success in population for person i  - p̂i = predicted prob of being a success for person i  - logit transformation: logit = ln(𝜋­­i / 1- 𝜋­­i)  - at what X does prob=0.5: X0.5 = -B0/B1  - logit regression: slope describes the relationship between X and the probability of Y  - LR Pop Model: (LR est samp eq is same, just w/ p̂i)  Log odds: ln(𝜋­­i / 1- 𝜋­­i) = B0 + B1X1+B2X2…  Odds: (𝜋­­i / 1- 𝜋­­i) = eB0 + B1X1+B2X2…  Prob: 𝜋­­i = 1 / (1+e-(B0 + B1X1 + B2X2….)) =prob of success |
| Week 15 | May 2  - actual pop (ex): all school districts in 2011 year in MN who were willing to participate in the study  - sample (ex): the 334 school districts in the 2011 school year in MN who participated in the study  - (ex)1-way ANOVA population model for avg teacher salary using poverty status (3 categories): y = 𝜇 + 𝛼­k + 𝜖 where 𝜖~N(0, σ), k=1,2,3  - (ex) ordered diff report: there is very strong evidence that pop mean poverty below 0.10 (Group 0) is diff than both btwn 0.10 + 0.15 (Group 1) and above 0.15 (Group 2) b/c the p-values are so small  - (ex)tukey confidence intervals: we are 95% confident that the pop mean poverty below Group 0 is \_\_ and \_\_ more than between Group 1. | May 4  - (ex) distribution of tot price of diamond: skewed right with a median of \_\_ and an IQR of \_\_. There appears to be multiple outliers for expensive diamonds.  - (ex) standard deviation of tot price of diamond: the avg difference between the total prices of diamonds to the mean price of [avg] is [stand dev]  - (ex) 95% confidence int: we are 95% confident the pop mean total price for a diamond is between \_\_ and \_\_.  - ex: “R2=0.12: 12% of the variability in happiness can be explained using a linear model including poor sleep quality. [after accounting/adjusting for the complexity of the model (adjusted R2 - penalty for adding new x)]”  - correlation: sqrt(r^2) |

- 4 conditions:

- Linearity: residual plot (proper LM and mean of residuals is 0)

- Independence: “it is unlikely one \_\_ will affect another.” (however, the \_\_ weren’t randomly selected raising concern about whether or not the sample is representative of the population of all perch)

- Normality: normal qq plot, residual normal quantile plot (residuals follow a normal distribution, no curvature)

- Equal variance: residual plot (no fan shapes)

- CBs are narrower than PBs b/c CBs are estimating a pop mean and the distr. of means has less varia. than indiv. like in a PI

- (ex) slope intpr for length (quantitative): when length increases by 1cm, we predict weight to increase \_\_ when width is held constant

- (ex) slope intpr for moderate (qualitative): the pred height for trees with moderate thorny cover is \_\_ higher than tree with no thorny cover when height and diameter were held constant

- (ex) confidence interv for slope: we are 95% confident that when the diameter increases by 1cm, the pop mean height increases between \_\_ aend \_\_ when the amount of thorny cover and height in 1990 are held constant

- to detect multicol:

- correlations btwn vars (|r| > 0.9), VIF > 5

- T-tests for all/nearly all indiv slopes not sig, but F-test for overall model sig (or R2 high)

- log-linear slope: when x increases by 1 unit, predict the median of Y to change by a factor of eb1

- linear-log slope: when x increases by p%, we predict Y to change by b­1ln(1+p/100)

- log-log slope: when x increases by p%, we predict the median of Y to change by a factor of (1+p/100)b1

|  |  |  |
| --- | --- | --- |
|  | x | Ln(x) |
| Y | Linear: Y^=b0+b1x | Lin-log: Y^=b0+b1ln(x) |
| Ln(y) | Log-lin:  Ln(y)^=b0+b1x | Log-log  Ln(y)^=b0+b1ln(x) |

- suppose you increase by 10%, p=0.10 (1.10b1)