Lecture 5 – Inference for Numerical Data

**One-sample t-test**

t test has a thicker tail than the z test

The test statistic for inference on a small sample (n < 30) mean is the T statistic with df = n - 1:

SE =

Confidence Intervals = point estimate ± ME, ME = t\* x SE

Using the t table the t\* value is at the intersection of the df row and two tail probability of .05

Summary

If σ is unknown and we do not have a reliable estimate, use the t-distribution with SE =

Conditions 1) Independence of observations 2) No extreme skew

**Paired t-test**

When two sets of observations have a special correspondence (not independent), they are said to be paired

To analyze paired data it is useful to look at the difference in outcomes of each pair of observations diff = x – y (difference in means)

Parameter of interest = average difference between population values µdiff

Point estimate = average difference between sampled values

No different than before we just use the one sample which is the difference between the observations

**Two sample t-test**

Mean difference µgroup1- µgroup2

SE =

Confidence Intervals = point estimate ± ME, ME = t\*df x SE

Conditions: 1) independence within groups (often verified by a random sample, if sampling without replacement, n<10% of population) and between groups 2) no extreme skew

**ANOVA (ANalysis Of VAriance)**

To compare means of 3+ groups, ANOVA is used to assess whether the mean of the outcome

variable is different for different levels of a categorical variable.

Use F statistic

H0 : The mean outcome is the same across all categories, µ1 = µ 2 = ….. = µk; where µ i represents the mean of the outcome for observations in category i.

HA : At least one mean is different than others.

Test Statistic: F =

Conditions: 1) observations should be independent within and between groups 2) Observations should be nearly normal 3) the variability across the groups should be about equal.

Large test statistics lead to small p-values.

Degrees of freedom: groups dfG = k – 1, where k is the number of groups, total dfT = n-1, where n is tht total sample size, error: dfE = dfT - dfG

Sum of squares between groups is SSG = SG1+SG2+…

Total sum of squares,

Sum of squares error, SSE = SST – SSG

Mean square error: MSG = SSG/df df=k-1, MSE=SSE/df df=n-1

F = MSG / MSE

Estimate any within-group standard deviation with , use the error degree of freedom n-k for t distributions

Difference in two means after a significant F-test

α\*=α/k

**Lecture 6 Inference for Categorical Data**

**Single Proportion**

Parameter of interest: proportion of all - p(a population proportion)

Point estimate: proportion of sampled – (a sample proportion)

Confidence Intervals = point estimate ± ME

Standard error of a sample proportion is SE =

Central limit theorem for proportions: Under certain conditions sample proportions will be nearly normally distributed with mean equal to the population mean, p, and standard error equals to SE below

)

Conditions: 1) independent observations 2) at least observed 10 successes and 10 failures (large sample)

Parameters: Population Parameter: p, Point Estimate p(hat)

Confidence Interval

CI vs HT: CI at least 10 observed Successes and failures, HT at least 10 expected successes and failures calc using null value

H0: p = p0 HA: p ≠ p0

Test statistic

**Two Proportions**

Parameter of interest: Difference between the proportions of all group1 and all group2: p1-p2

Point estimate: difference between the proportions of sampled1 and sampled2

Confidence Intervals = point estimate ( ± ME

Conditions 1) observations within groups are independent 2) observations between groups are independent 3) at least 10 observed successes and failures in each group

Confidence Interval:

For HT we don’t have a given H0 so we need to calculate a pooled estimate of a proportion

CI use observed sample (p1,p2) for HT use pooled estimate phat

Hypotheses: H0: p1 = p2 HA: p1 ≠ p2

Test statistic:

**Chi-Square Test of GOF (goodness of fit)**

H0: There is no inconsistency between the observed and expected counts

HA: There is an inconsistency between the observed and expected counts

Test Statistic , k=total number of cells, Larger Χ2 means stronger deviation from null

Expected cell counts Ei, Observed Cell Counts Oi

hypothesis testing

Degrees of freedom = k-1

Hypothesis Testing Conditions: 1) Independence, Each case within the table is independent of all other

cases in the table 2) Sample size, Each cell must have at least 5 expected cases 3) df>1, degrees of freedom must be greater than 1

**Chi-Square Test of Independence**

Hypothesis testing

Test Statistic where df = (R-1)x(C-1)

K is the number of cells, R is the number of rows, and C is the number of columns

p-value is the area under the X2 curve

**Lecture 7 Linear Regression**

Correlation: Quantifies the linear relationship between two numerical variables.

Regression: Models a numerical response variable using one (or more) numerical or categorical explanatory variables.

**Correlation – describes the strength of the linear association between two variables**

Y=response variable, x=explanatory variable

Correlation takes values between -1 negative association and 1 positive association, 0 no association

**Best fit regression line**

Residuals are the leftovers from the model fit Data = Fit + Residual

The residual of an observation (ei) is the difference between the observed value (yi) and its predicted value ,

Want small residuals. Option 1 minimize the sum of magnitudes: |e1|+|e2+…+|en|

Option 2 minimize the sum of squared residuals (least squares regression)

Least squares regression line , = predicted y, β0 = intercept, β1 = slope, x = explanatory variable

The slope of the regression

Intercept is where regression line intersects the y-axis

Conditions for the least squares regression line 1) linearity 2) Nearly normal residuals 3) constant variability 4) independence of observations

The strength of the fit of a linear model is most commonly evaluated using the Coefficient of Determination, R2

Categorical Explanatory Variables, must convert categories into numerical form usually 0/1

Outliers are points that lie away from the cloud of points, if they lie horizontally away from the center of the cloud they are called high leverage points, if they actually influence the slope of the regression line they are called influential points

Always use a t test for inference for regression

Df = n-2 or Df = n-k+1

Confidence Intervals = point estimateb1 ± ME, ME = t\* x SEb1

Hypothesis test

Null value usually 0 since we are looking for any relationship between explanatory and response variable, Use two tailed

**Lecture 8 Multiple Linear Regression**

, using ANOVA we can calculate the explained variability and total variability in y

Two predictor variables are said to be collinear when they are correlated, and this collinearity complicates model estimation.

, p is the number of predictors, n is the number of cases, high is better

Conditions: 1) residuals are nearly normal 2) residuals have constant variability 3) residuals are independent 4) each variable is linearly related to the outcome

Backward-elimination 1)start with full model 2)drop one variable at a time and record of each smaller model 3) pick the model with the highest increase in 4)repeat until none of the models yield an increase in

p-value approach: 1) start with the full model 2) drop the variable with the highest p-value and refit a smaller model 3) repeat until all variables left in the model are significant

Forward-selection

adj approach: 1) Start with regressions of response vs. each explanatory variable 2) Pick the model with the highest 3)Add the remaining variables one at a time to the existing model, and once again pick the model with the highest 4) Repeat until the addition of any of the remaining variables does not result in a higher

p-value approach: 1) Start with regressions of response vs. each explanatory variable 2) Pick the variable with the lowest significant p-value 3)Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value 4) Repeat until any of the remaining variables does not have a significant p-value In forward-selection the p-value approach isn’t any simpler, so there’s almost no incentive to use it.