commerce undergraduate society

COMM 291

FINAL EXAM REVIEW SESSION

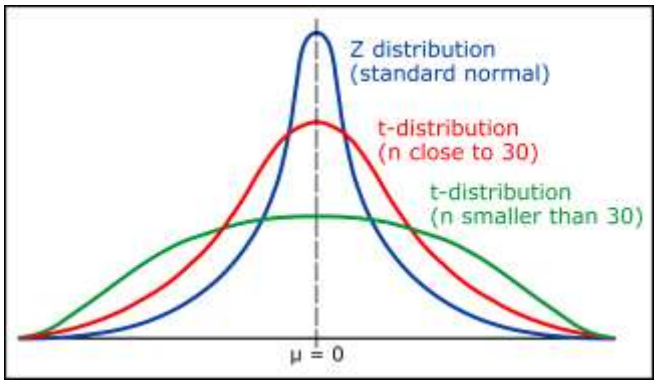
by Simon Roberts

Contents

Hypothesis Testing for Means

A sample with the mean $\bar{y}$ will have a sampling distribution with the same mean, but a standard deviation of $\frac{\sigma}{\sqrt{n}}$. The shape of the sampling distribution will be normal, if the sample size is large enough.

Student's t-distribution: Normal Model did not work for small sample sizes; the t-distribution models the sampling distribution for each sample size, deriving to degrees of freedom.



The t-distribution is still unimodal and symmetric, but the tails are stretched out much farther when n is small. Thus, it follows that using a t-distribution will give wider margins of error when constructing confidence intervals and will be less sensitive to rejecting the null hypothesis when testing. As n approaches infinity, the t-distribution's shape approaches that of the Normal model.

$$t = \frac{\bar{y} - \mu}{SE\ (\bar{y})}, SE\ (y) = \frac{s}{\sqrt{n}}$$

Confidence Interval:

$$\bar{y} \pm t^*_{n+1} * SE(y)$$

Paired Data refers to when two groups are not independent. Measurements are paired as before and after or are matched to each-other. The relationship to examine is the difference between the two paired measurements. For example:

- One pair of before and after data: change in student satisfaction survey responses for a particular student in 2016 and 2017.
- Matching relationships: Difference in rating of a book between husband and wife

The difference can be essentially analyzed using the same mechanics as a **one-sample t-test:**

Answer the following questions about a call center.

The cable television company measured last month that the average wait time for callers of the technical support line to speak to a customer service representative (CSR) was 40 minutes. The manager wants to determine whether the average wait time has decreased this month. To reduce costs, they have decided that if the wait time has decreased, the scheduling of CSRs will be adjusted. 100 people are randomly sampled, and they waited for a mean of 38 minutes, with a standard deviation of 5 minutes. Determine if there has been a statistically significant decrease in wait time.

Comparison of Two Means

Parameter of interest: The 'true' difference between two means. Does the difference between two sample means indicate a real difference in population means, or was it the result of a random fluctuation?

Essentially, we'll be using a very similar test!

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE}, \text{ where SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

This test only works for independent groups; if the groups are related, or not created randomly in relation to each-other (Sibling data, same observations over different time periods, etc.) For a conservative estimate he number of degrees of freedom is the smaller of the two sample sizes minus 1.

The Pooled T-Test has one additional assumption: the variances of the two samples are assumed to be equal. If the sample sizes are equal, the pooled variance is simply the average of the two groups' variances. Degrees of freedom are reduced. This approach to comparison of means problems is mathematically consistent with ANOVA and Regression but performs poorly when the condition is not met. Check out this exploration by jbstatistics for more information.
df = (n1 – 1) + (n2 – 1)

Answer the following questions about a bookstore.
The bookstore wanted to boost sales by inducing students to decide to purchase more books immediately during their visits to the store. The manager thought that by randomly distributing a coupon to give percentage discounts valid for only 30 minutes after the coupon was given out, students would be incentivized to buy the books right away, instead of leaving to shop around and compare pricing at other sources. On one day, the manager asked all students entering the store to swipe their student cards and randomly distributed coupons for 5% and 10% off, valid only with the student card swiped on entry. The results of the campaign are as follows:

	5% Off	10% Off
Mean Spending	$200	$215
Standard Deviation	$30	$30
Number of Students	25	25

State the null and alternative hypotheses and carry out a difference of sample means tests at the 5% significance level to make a decision about the null hypothesis

Comparison of Two Proportions

Confidence Interval:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* * SE(\hat{p}_1 - \hat{p}_2)$$

Standard Error:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Question:
Suppose the Acme Drug Company develops a new drug, designed to prevent colds. The company states that the drug is equally effective for men and women. To test this claim, they choose a simple random sample of 100 women and 200 men from a population of 100,000 volunteers.

At the end of the study, 38% of the women caught a cold; and 51% of the men caught a cold. Based on these findings, can we reject the company's claim that the drug is equally effective for men and women? Use a 0.05 level of significance.

Chi-Squared and Goodness of Fit

A Chi-Squared model allows us to examine the difference between observed and actual counts of events occurring. Each entry in the summation can be referred to as “The observed minus the expected, squared, divided by the expected.” The chi square value for the test as a whole is “The sum of the observed minus the expected, squared, divided by the expected.”

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

The "goodness-of-fit test" is a way of determining whether a set of categorical data came from a claimed discrete distribution or not. The null hypothesis is that they did and the alternate hypothesis is that they didn't. It answers the question: are the frequencies I observe for my categorical variable consistent with my theory? The goodness-of-fit test expands the one-proportion z-test.

The "test of homogeneity" is a way of determining whether two or more sub-groups of a population share the same distribution of a single categorical variable. For example, do people of different races have the same proportion of smokers to non-smokers, or do different education levels have different proportions of Democrats, Republicans, and Independent. The test of homogeneity expands on the two-proportion z-test.

The "test of independence" is a way of determining whether two categorical variables are associated with one another in the population, like race and smoking, or education level and political affiliation.

Note** Ensure the expected counts for each of the cells you are working with is at least 5. Although this is arbitrary this rule ensures that we're working with sufficiently sized data for the problem.

Question: Are gender and education level dependent at 5% level of significance? In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained? Which chi0

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

Linear Regression

Equations fit with straight lines can be given confidence intervals about the slope and intercept. The estimated slope from a regression based on a sample will be close to, but not equal to the parameter slope. There will be a sampling distribution of the slope, just like there is on proportions and means.

Prediction intervals are wider than confidence intervals for the slope of a regression line because confidence intervals provide a range of values for the mean/average observation while a prediction interval is for the actual range of value for a particular observation.

Multiple Regression and ANOVA (Analysis of Variation for Regression)

n = number of samples
k = number of predictor variables used

Two sources of variation between the actual y-value and the mean y-value of all the data points:

1) Difference in x-values: we expect points with different x-values to have different y-values these are predicted by the model
2) Other variables and random errors

Recall, variation is the same as the square of a standard deviation. The total variation in a multiple regression model is explained by the two above sources.

SST = SSM + SSE

ANOVA for Simple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Model	SSM	1	SSM	MSM/MSE
Error	SSE	n-2	SSE/n-2	
	SST	n-1		

R-Square = SSM/SST = 1 – (SSE/SST)

ANOVA for Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Model	SSM	k	MSM	MSM/MSE
Error	SSE	n-k-1	MSE	
	SST	n-1		

Adding more variables will never decrease r-squared, but it can lower the adjusted r-square value. According to the principle of parsimony, we want to reduce our model to one with the fewest predictors and the most predicting power.

Adjusted R-Square = (SSE/n-k-1)/(SST/(n-1))

The F-test to compare full and reduced models provides justification for whether a variable should be kept in a multiple regression model.

H0: Reduced model is adequate
HA: Reduced model is inadequate

Questions on Multiple Regression

The model below details a department store that keeps track of several metrics to determine causes of shoplifting.

OUTPUT 2. Multiple Regression (Four predictors)

Regression Statistics	
Multiple R	0.8106
R Square	0.6571
Adj. R Square	0.5428
Standard Error	302.0344
Observations	17

ANOVA

	Df	*SS*	*MS*	*F*	*Sig. F*
Regression	4	2097934	524483	5.749	0.0080
Residual	12	1094698	91225		
Total	16	3192632			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	4875.365	1543.275	3.159	0.0082
TEMP	6.263	5.963	1.050	0.3142
SALETRAN	0.239	0.073	3.275	0.0066
HOLIDAY	-190.259	202.545	-0.939	0.3661
EMPLOYEE	-27.322	11.974	-2.282	0.0415


CMP