Frequently Missed Concepts in Stats 2

How to read a Z table:

The Z table is set up with the Z value second decimal places on the top row and the whole number and first decimal part of them on left most column. So for the value 1.54 you go look at the first column and find the section that has 1.5 and draw a line going rightward. Then look at the top row and look up the section that has 0.04 and you draw a line going downward. Where your two lines intersect is the associated probability value, or P-value. The way to interpret this value is to view it as the probability of observing a Z value or less than the Z value.

You will often want the value of Z or greater (the tail value) so you will want to take the probability you found in the table (if the probability in the table is the opposite of what you want) and do 1-(Probability you found in the table) to find the tail probability. A little trick that is used often is exploiting the fact that the Z distribution is symmetric about 0. So if your interested in the P(Z>1.54), since the table is set up with P(Z<1.54), you can do P(Z>1.54)=1-P(Z<1.54) or you can observe that P(Z<-1.54)=P(Z>1.54) as the are both the same values at the opposite end of the distribution. Interestingly enough the Z table has P(Z<-1.54) so all you have to look up is P(Z<-1.54) and you'll get the same answer as if you did

1-P(Z<1.54).

How to read a t table:

The t value that you find in the table is entirely dependent upon how confident you wish to be with your observation and the degrees of freedom (df), which will be n-1 most of the stats II course (your teacher will tell you when it is different). To read the t table, first you find the df value that you have located in the far left column, and draw a line rightward. You then look up your a level, (or 1-a CI level, as they are the same value t value) and draw a line downward. Where the two lines meet, this is your t value of associated a level.

Knowing when to use Z or t tables:

To know when to use the Z or the t table you look at what kind of problem you are performing. If you are doing proportions, you always use the Z (when making inferences you should still check the $n*p \geq 15$ and $n*(1-p) \geq 15$) . When dealing with means you must check to see if $n \geq 30$. If $n \geq 30$, then use the Z table (as the t approximates to the Z when $n \geq 30$). But if $n < 30$, you must use the t table. But there is another Caveat; when using the t table, the data must come from a normal distribution or else no statistical inference can be done). For how to read the Z/t tables, refer to above.

Hypothesis testing:

Hypothesis testing is a statistical inference technique for a population parameter that tests to see how likely you are to observe a value that you found in an experiment given that you think (or hypothesized) that the true value of the parameter is a given value.

Setup:

- Read the question and decided what you would like to test against and also decide whether or not you will be working with proportions or means
- Decide what the null and alternative hypotheses are:

Null: $\mu=\mu_o/p=p_o$. This is your intuitive guess at what the parameter value is. You will be testing against this value

Alternative: $\mu/p > \text{or} < \text{or} \neq \mu_o/ p_o$. This is what you're seeing if there is enough evidence to decide this. You will be using X-bar/P-hat to compare to your null value to see if there is enough evidence.

- Perform the test

$Z=(X-\mu)/(s/\sqrt{n})$ for means or $t=(X-\mu)/(s/\sqrt{n})$ if $n<30$ and sample is from Normal distribution.

$Z=(\text{P-hat}-P)/ \sqrt{\frac{P\times(1-P)}{n}}$ for proportions

P-value:

Many times throughout stat II, you will be asked to find and interpret a p-value.

Finding P-values: To find a p-value, you must first calculate your test statistic (To compute the test statistic, refer to concepts from stat I that are integral for stats II). After finding the test statistic, you must look it up in the appropriate table (refer above on how to read the tables). The p-value is the value you find within the table that is associated with the test statistic. This value, whether it be the Z or the t is the p-value. If you are using the Z table, you will get an exact value, but if you are using the t table you may (and often do) end up with a range of possible p-values. This occurs when your test statistic is not exactly on the table but is between two associated p-values. In this case you just report the range. There is an exact value but you need to use a computer or the internet to find it exactly.

Interpreting P-values:

> P-value: The probability of observing a value as extreme or more extreme than the one we observed, GIVEN Ho IS TRUE.

What this means for you is that the p-value is the probability of observing the value you found (be it x-bar or p-hat) with Ho being true. So if the p-value is low, we are led to believe that Ho is untrue because we observed x-bar/p-hat and just hypothesized the Ho value. A p-value only has this interpretation; no more, no less. So don't be fooled.

An arguably childish saying that I agree is excellent for those who can't remember the definition of p-value for one reason for another is:

--P-value low, reject that Ho. It's always right and should be remembered just in case.--

Confidence Intervals: For the direct computation of confidence intervals refer to the concepts from stat I that are integral for stats II. Arguably more important though is the interpretation of the power of a confidence interval and what it actually means. The power of a confidence interval comes from the fact that when we say we are 95% (or 90% or 99% etc), it actually means that if we were to repeat more experiments of the same design then 95% of the intervals created will contain the true parameter value. Because of that fact, we can say that we are 95% confident that the true parameter value is contained in the interval (X,Y).

Some important notes about confidence intervals:

A confidence interval says nothing about the sample. It can predict noting about a new sample.

A confidence interval is not a measure of probability. We are not guaranteed that some percent of observations will have the value.

A confidence interval only gives a range of possible values for the population parameter. Not the sample mean (we know the sample mean already!

ANOVA: (or called Analysis of Variance)

ANOVA is a statistical test that examines data that is separated into groups and tries to detect a difference between the group means. An excellent use of ANOVA would be if you were an owner of a company and were considering airing different ads. You would then show each individual ad to a group of people and then they would rate how much they liked the product. To test if there is likely any difference in group means you would use ANOVA. After performing the test, you would know which ad to air. While the equations for ANOVA's various components are difficult, they can be found in your statistics book, which has good descriptions of the equations. The main problem students tend to have with ANOVA is the interpretation of each individual component.

SSR=SSTR=SSG: The sum of squares between the groups. It is a measure of how variable the group means are compared to the overall mean of all the groups. It is the squared deviation between each group mean and the overall mean. A higher SSE implies there is a large difference between means of some groups (at least 1) and the total group mean.

SSE=SSW: The sums of squares within each group. It is a measure of the variability within each group, or the square deviation each observation within a group is from the group mean. A high SSE implies that there is much variance within the group, or that there is a large range of data within the group.

SSTO: The total sums of squares. It is the totally variability of the data and is thus the squared deviation between each observation and the total mean.

SSE+SSR=SSTO

So you only need two of the above values to find the other

Degrees of Freedom

The degrees of freedom (df) are for all intents and purposes in ANOVA, a scaling measure for SSE and SSR. With each SSE and SSR there is an associated df:

With 1 way ANOVA which is what most of what you'll be doing is:

SSE: df_e=N-g where N is the total sample size and g is the amount of groups

SSR: df_g=g-1

SSTO: df_t=n-1

df_e+df_g=df_t

So you only need two of the above values to find the other

MSE: Mean square error. The scaled SSE that is used in ANOVA. It takes into account the number of observations in the sample. It is simply SSE/df_e

MSR/MSG: The scaled SSR/SSG that is used in ANOVA. It takes into account the number of groups that are included in the study. It is simply SSG/df_g

F* The ratio of The between group variability and the within group variability.

MSR/MSE ~F(.95,g-1,N-g). When you compute F* you compare it to F(.95,g-1,N-g). The value F(.95,g-1,N-g) represents the minimum F value for which you reject Ho. Any F* value greater than F(.95,g-1,N-g) would cause you to reject Ho (if you have any confusion, refer to the definition of p-value, listed above). To read the F table, the far left column is the numerator degrees of freedom (df_g). So you find your df_e on the far left column and draw a line horizontally from there. You then go to the top row and find your denominator degrees of freedom (df_e) and draw a line vertically. Where the lines meet is F(.95,g-1,N-g), your minimum rejection F value.

Regression:

Regression is another form of statistical inference in which a series of observations are collected, and then a model that best fits the data is created. This line (called the best fit line) is called the line of best fit and is the mathematically best line through the graphed data points that minimize the squared distance between the line and the data points. A regression line may be linear (which includes just a straight line, but also encompasses quadratic, and any polynomial terms) and non-quadratic, such as logistic. In stats II most of the regression that will be done is simple, as the math involved for any regression that is above simple (which is just a straight line)

The general form of simple linear regression is :

$\mu i=\alpha+\beta Xi + \varepsilon i$

Where:

μi: the mean value of the ith case

α: the intercept, or the mean value when Xi =0.

β: the slope of the line, or the increase in μi when the independent variable (X) is increased by 1 unit.

εi: The random error term that is associated with each observation. It is the amount each μi is off from what is expected of it from the regression equation. It comes from the fact that each observation is a random observation from a population so of course there is random "noise" as statisticians call it associated with each observation. εi~N(0,σ^2) which means the mean of all of the ε's are 0 and they have a variance of σ^2.

In practice, though we don't have the true regression formula, but instead have the observed formula, which is derived from the data.

The observed regression formula, which tries to predict the true regression components is:

Yi=a +b*Xi

Where Yi is the mean response of case i, a is the best fit intercept, b is the slope of the best fit line, and Xi is the value of the independent variable of case i.

To find b: r*(S_y/S_x) or r *(Standard deviation of Y's)/(Standard deviation of X's)

To find a: Y_bar – b*X_bar

R^2: (coefficient of determination) is the amount of variability in Y that is explained by X. In laymen's terms, R^2 is the amount of variability that you see in Y that is eliminated when X is included in the model (i.e. Yi =a for all X values) It takes values between 0 and +1.

r: (coefficient of correlation) does not have as strict a definition. Put simply, r is the amount of linear relation x and y have. It shows how well X predicts Y and Y predicts X. So if for all X values, if you increase X, Y increases, then they have a positive correlation. And vice versa for a negative correlation (X increases, Y decreases). If increasing X does not give any insight into whether or not Y will increase or decrease, then the correlation will be near zero.

Now try it yourself!

1. Given a Z value of

a)1.96

b)2.79

c)-3

d)-45939456.56763

Find the tail p-value

2. Given a t value of

a) 63.657 at df=1

b) 2.56 at df=10

c) 2.8 at df=20

d) 1.96 at df=1986858675483

Find the two tail p-values or range of p-values

3. A researcher is interested in the mean number of cute pug pictures the average American looks at in a given month. The commonly accepted mean number among academic circles is 75 pictures per month. This researcher doesn't believe this, however, and decided to go out and test this. He went out and asked 100 random people how many pug pictures they look at in a month. The mean of his 100 observations was 85 pug pics per month with a variance of 25.

a) Conduct a hypothesis test at a=.05 to see if the true mean pug picture viewing for Americans in a month is greater than 75. Check all assumptions, write down the null and alternative hypotheses, find the test statistic, report and interpret the p-value, and state the conclusion. Construct a 95% confidence interval for the true mean number of Pug pics looked at in a month.

b) Conduct the hypothesis again but instead of using n=100 for the test, use n=10000. What did you notice about the Test statistic and resulting p-value? What does that show you about the power of hypothesis testing as n increases?

Also

4. Chris walked up to Meredith and said that 50% of people would prefer rock god Jimmy Page to be president of the United States. Meredith, being the studious and inquisitive woman that she is decided to test this hypothesis. She traveled the country and asked 657 people whether they prefer Jimmy Page or Barrack Obama as president. She found that actually 49% of Americans want Jimmy Page as president (it would probably make the state of the union more rocktastical)

a) Conduct a hypothesis test at a=.01 to see if the true proportion of Americans that support Jimmy Page as president is 0.50. Check all assumptions, write down the null and alternative hypotheses, find the test statistic, report the p-value, and state the conclusion. Construct and interpret a 99% confidence interval for the true proportion of people that support Jimmy Page.

b) If Meredith had asked the first 657 people she met on campus at UF instead of traveling the country, would her results still be valid?

5. I am interested in testing whether or not people can taste the difference between milk that was produced organically and non-organically. I asked 16 random people and each of them was randomly assigned to either organic (group 1) or non-organic (group 2). Each person was then asked to rate the tastes on a scale of 1 to 10 (ten being the tastiest)The observations are as follows:

Group 1: 5, 6, 4, 7, 9, 10, 10, 7

Group 2: 4, 4, 6, 7, 5, 10, 3, 5

Find:

a)Y-bar for group 1 and 2 as well as the total mean

b)Si^2 for each group

c)SSE

d)SSG

e)SSTO

f)MSE and MSG

g)Find F*

h)state the conclusion

6. Data was collected about length of roots growth between 1 and 11 months. A simple linear regression model was fit:

Y={3, 5, 4, 6, 8, 13, 14, 16, 19, 22, 30}

X={1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}

r=.9641

Find each for the model:

i) a

ii) b

iii) The regression equation.

iv) Interpret the regression coefficients

Answers

1. (read how to read a Z table)

a).025

b).0026

c).0013

d)0.0000

2. (read how to read a t table)

a).01

b).02 and .05

c).01 and .02

d) .05

3. refer to hypothesis testing, with

Ho: μ=75

Ha: μ>75

X_bar=85

S=5

TS=20

p-value=0

At a=.05 we reject the null hypothesis and say that there is statistically significant evidence to say that μ>75. The 95% confidence interval for μ is (84.02,85.98). We are 95% confident that the true mean is in the interval (84.02,85.98). This supports the hypothesis test as 75 is not in the interval and thus not a probable value. We know this from that fact that, in repeated tests of the same set-up 95% of the intervals we create will contain the true parameter value. If we used n=10000

the TS=20000, which would give more evidence that the true mean value is above 75. This makes sense as when we get more data from the population, our guess at the parameter value will be closer and closer to the true value.

4. At a=.01 we fail to reject the null hypothesis and say that there is no statistically significant evidence to say that p<.50. The 99% confidence interval for μ is (.465,.554). We are 99% confident that the true proportion of people who support Jimmy Page as president is in the interval (.465,.554). This interval, however, contains .50, so we can't rule it out as a possible value. We know this from that fact that, in repeated tests of the same set-up 99% of the intervals we create will contain the true parameter value.

Also if Meredith had just went around UF, her sample would no longer be a random sample and her results would not be valid.

5.

Group 1 mean: 7.25

Group 2 mean: 5.5

Total mean: 6.375

n1/n2: 8

SSG: Since there are two groups, SSG only has 2 elements-

8*(7.25-6.375)^2 + 8*(5.5-6.375)^2=12.25

SSE=using the equation for variance of group i (1 and 2) = ∑(Xi-Xbar)^2. More on sample variance is found in important concepts from Stat I for Stat II.

S1=5.071

S2=4.857

Which yields SSE, which has two terms as well, to be-

(8-1)5.071+(8-1)*4.857=69.496

And SSTO=SSE+SSG=69.496+12.25=81.746

With dfG=2-1=1

And dfE=N-g=16-2=14

dfTotal=N-1=dfG+dfE=g-1+N-g=N-1=16-1=15 as expected

So MSG=SSG/dfG=1225/1=12.25

And MSE=SSE/dfE/69.496/14=4.964

So F*=MSG/MSE=12.25/4.964=2.46

To check to see if we reject, we check to see if F*>F(.05,1,14) where F(.05,1,14)=4.60.

Since F* is not >F(.05,1,14) we fail to reject Ho and say there is not enough evidence to say there is a difference in the mean taste rating of the organic and non-organic foods.

6.

a=-2.21818

b=2.49091

The regression equation is: -2.218+2.4909X

a does not have an interpretation, as obviously there not a negative length at time 0. The slope is interpreted as the amount of length increase with a 1 unit increase in the X (which is a 1 month increase)