MAT 1372 Stat w/ Prob classwk 26 Fall 2013

**7.6 THE CHI-SQUARE DISTRIBUTION AND STATISTIC**

**(DISTRIBUTION OF THE SAMPLE VARIANCE OF A NORMAL POPULATION)**

**Degrees of freedom (d.o.f.)** is size of the sampling minus number of estimators used in calculation (such as sample mean).

For example, for the sample mean with 3 observations, d.o.f.= 3.

For the sample variance, we make use of sample mean, so d.o.f.= 3−1 = 2.

Given a family  of independent standard random normal variables, we define the **chi-squared** random variable to be .



For n=1, is . Note that density curve has shape of exponential decay.

**Thm:** For n=1, 

**Proof:** Recall that E[Z]=0 and V[Z]=1. Hence



As n increases, the Central Limit Theorem kicks in and the shape of the distribution becomes closer to that of a normal distribution, with mean = n (we are summing and not dividing by n):

**Thm:** 

**Proof:** .

**Fact**(which leave unproven): For n>0, the peak for is n-1.

For example, the peak for n=3 is at 2.

Let’s start with, family of normal random variables with mean μ and SD σ. We standardize , and substitute into the definition of, to get



with n degrees of freedom.

If we replace μ (fixed) with the estimator (the sample mean), we also get a distribution, but with 1 less degree of freedom.

**Thm:** with n−1 degrees of freedom.

The associated **statistic** is



Note:is closely related to the sample variance:



**Exercise**: check that

**1.** The following data sets come from normal populations whose standard

deviation σ is specified. For each data set, determine and the degrees of freedom (d.o.f.) for the associated distribution.

1. 104, 110, 100, 98, 106; σ = 4

**(b)** 1.2, 1.6, 2.0, 1.5, 1.3, 1.8; σ = 0.5

**(c)** 12.4, 14.0, 16.0; σ = 2.4

The data for all 3 parts and the calculations for (a) have been put into the accompanying Excel file.

We illustrate the process with the data set from part (a):

1. Find sample mean:

=103.6

1. Sum squares of deviations from mean and divide by variance: 
2. Degrees of freedom is number of data points minus one:

d.o.f.=5−1=4.

**13.2 CHI-SQUARED GOODNESS-OF-FIT TESTS**

**4.** A random sample of 100 student absences yielded the following data

on the days of the week on which the absences occurred:

Day M T W T F

Frequency 27 19 13 15 26

Test the hypothesis that an absence is equally likely to occur on any

of the five days. What are your conclusions?

If the absences are equally likely, then they should be more or less equally spread out over the week, but clearly they are clustered at the ends. What we must do is measure just how far they are from our hypothesis, so we calculate the deviations squared from the expected and take a sum weighted by the inverse of the expected (essentially we are scaling). This sum is our Test Statistic for use in the chi-squared test, with the degrees of freedom set to one less than the number of parts in our distribution: df = 5 – 1 = 4.

Here are the steps to complete a goodness of fit test.

1. Find the number of observed values by summing.

*Result is 100 for our problem.*

1. Multiply the number of observed by the expected percentages to get the number expected for each outcome.

*Result is 20 for each of our possible outcomes.*

1. For each outcome, square the difference between the observed and the expected and divide by the expected.

*Results vary from .05 to 2.45*

1. Sum the results of 3 to get the “test statistic” TS.

*Result is 8.*

1. To find the pvalue, use TS as input into “chidist”, with the degrees of freedom equal to the possible number of outcomes minus 1.

*Result is .0916, which is not enough evidence to reject the null hypotheis (that the variation observed is due to randomness).*

Note that the pvalue can be found directly using the command **CHITEST** with the observed (actual) and expected values as inputs.

For your convenience, I have posted the theory below. If you do nothing else, make sure you look for where the TS is defined. Note that the “N’s” are the Observed and the “e’s” are the expected.

****

**15.** The following gives the age breakdown, by percentages, of unmarried

women having children in 1986:



A recent random sample of 1000 births to unmarried women indicated

that 42 of the mothers were age 14 or younger, 403 were between

15 and 19 years old, 315 were between 20 and 24 years old, 150 were

between 25 and 29 years old, and 90 were 30 years or older. Do these

data prove that today’s percentages differ from those in 1986?