



Session 7.2: Estimation Techniques

- 1 Estimation - Introduction
- 2 Parametric Test – T-test
- 3 Parametric Test – Analysis of Variance
- 4 Parametric Test – Regression – Least Squares
- 5 Parametric Test – Regression – Cat. Dep. Var
- 6 Non-Parametric Test - Introduction
- 7 Non-Parametric Test - Chi-Square
- 8 Non-Parametric Test – Fisher’s Exact test
- 9 Non-Parametric Test – Wilcoxon-Mann-Whitney test

- 1 Estimation- Introduction
- 2 Parametric Test – T-test
- 3 Parametric Test – Analysis of Variance
- 4 Parametric Test – Regression – Least Squares
- 5 Parametric Test – Regression – Cat. Dep. Var.
- 6 Non-Parametric Test - Introduction
- 7 Non-Parametric Test - Chi-Square
- 8 Non-Parametric Test – Fisher’s Exact test
- 9 Non-Parametric Test – Wilcoxon-Mann-Whitney test

Estimation - Introduction

- Types
 - Point and interval Estimation
 - Parametric and Non-parametric

- Properties of Estimators
 - Unbiasedness
 - Consistency
 - Relative Efficiency

Parametric Tests - Introduction

- **Parametric** analysis relies heavily on the data being normally distributed. This enables the researcher to estimate the underlying population parameter using a representative sample
- The normal distribution condition implies that parametric analysis can only be carried out on quantitative data as it is only quantitative data that can have normal distribution.

Parametric Tests - Introduction

When the normal distribution assumption is satisfied, parametric tests allow us to test the hypothesis about the population parameter. For example we use **t-test** or **anova** to assess hypotheses about μ

μ Is the symbol for population mean

- 1 Parametric Test - Introduction
- 2 Parametric Test – T-test
- 3 Parametric Test – Analysis of Variance
- 4 Parametric Test – Regression Least Squares
- 5 Parametric Test – Categorical Dep. Variables
- 6 Non-Parametric Test - Introduction
- 7 Non-Parametric Test - Chi-Square
- 8 Non-Parametric Test – Fisher’s Exact test
- 9 Non-Parametric Test – Wilcoxon-Mann-Whitney test

Parametric-One Sample T-test

- The one sample t-test is used to test if the sample mean is significantly different from a hypothesized value
- Assumption :
 - It works under the assumption that the data is normally distributed
 - The variable is numerical

Parametric-One Sample T-test

- For instance we may wish to test whether the mean weight of the heaviest load carried last week (**week**) is equal to 12
- We formulate the hypothesis as

$$H_0 : m = 12$$

$$H_1 : m \neq 12$$

- In your Do file type

```
ttest weight =12
```

Parametric-One Sample T-test

```
. ttest weight=12
```

```
One-sample t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
weight	965	24.71917	.3322426	10.32093	24.06717	25.37117

```
mean = mean(weight)                                t = 38.2828
```

```
Ho: mean = 12                                     degrees of freedom = 964
```

```
Ha: mean < 12           Ha: mean != 12           Ha: mean > 12
```

```
Pr(T < t) = 1.0000       Pr(|T| > |t|) = 0.0000       Pr(T > t) = 0.0000
```



Parametric-One Sample T-test

- The results give the number of observations, the mean, standard error and 95% confidence interval.
- Below the table are three alternate hypotheses statements; ***Ha : mean < 12***, ***Ha : mean! = 12*** and ***Ha : mean > 12***

Parametric-One Sample T-test

- **Mean** - This is the mean of the variable.
- **Std. Err.** - This is the estimated standard deviation of the sample mean.
- **Std. Dev.** - This is the standard deviation of the variable.
- **95% Confidence Interval** - These are the lower and upper bound of the confidence interval for the mean. It specifies a range of values which the population parameter (μ) will lie. In this case we are 95% confidence that μ lies between 24.06717 and 25.37117

Parametric-One Sample T-test

- H_0 - Is the null hypothesis that is being tested. In case of a one sample t-test the H_0 we evaluate that the mean is equal to the given number.
- **Degrees of freedom** - The degrees of freedom for the single sample t-test is simply the number of valid observations minus 1. We lose one degree of freedom because we have estimated the mean from the sample.

Parametric-One Sample T-test

- **$\Pr(T < t)$, $\Pr(T > t)$** - These are the one-tailed p-values evaluating the null against the alternatives that the mean is less than 12 (left test) and greater than 12 (right test). These probabilities are computed using the t distribution. If p-value is less than the pre-specified alpha level (usually .05 or .01), we will conclude that mean is significantly greater or less than the null hypothetical value.

Parametric-One Sample T-test

- $\Pr(|T| > |t|)$ - This is the two-tailed p-value evaluating the null against an alternative that the mean is not equal to 12. It is equal to the probability of observing a greater absolute value of t under the null hypothesis. If p-value is less than the pre-specified alpha level (usually .05 or .01) we will conclude that mean is statistically significantly different from the hypothesized value. For example $\Pr(|T| > |t|)=0.0000$ which is smaller than 0.05 hence we fail to accept the null hypothesis and conclude that mean weight is statistically different from 12.

? Parametric-Independent Group T-test

This is used to compare means of the same variable between different groups. For instance one may wish to test the hypothesis that the mean age of female children is not different from the mean age of male children. In this case we are assuming that age is normally distributed. The test also assumes that the variances of the two populations are the same

Parametric-Independent Group T-test

```
. ttest age, by(sex)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
2. F	528	13.31629	.1102168	2.532588	13.09977	13.53281
3. M	472	13.34534	.1146991	2.491902	13.11995	13.57072
combined	1000	13.33	.0794444	2.512252	13.1741	13.4859
diff		-.0290511	.1592156		-.3414868	.2833846

```

diff = mean(2. F) - mean(3. M)                                t = -0.1825
Ho: diff = 0                                                  degrees of freedom = 998

Ha: diff < 0          Ha: diff != 0          Ha: diff > 0
Pr(T < t) = 0.4276    Pr(|T| > |t|) = 0.8553    Pr(T > t) = 0.5724

```

Parametric-Independent Group T-test

- The t-statistic is less than 2 ($|T|=0.1825$). Hence we fail to reject the null hypothesis.
- The same conclusion can be reached using $H_a: \text{diff} \neq 0$ and its corresponding $\Pr(|T| > |t|)$ of **0.8553**. Since **0.8553** is greater than **0.05**, we fail to reject the null hypothesis and conclude that mean age of female children is not different from that of male children.
- The result also gives additional information on the individual means for the sex categories (male=13.34534 and female=13.31629), the difference (diff=-0.0291) and the confidence intervals of the various statistic calculated

- 1 Parametric Test - Introduction
- 2 Parametric Test – T-test
- 3 Parametric Test – Analysis of Variance
- 4 Parametric Test – Regression Least Squares
- 5 Parametric Test – Regression Cat. Dep. Var.
- 6 Non-Parametric Test - Introduction
- 7 Non-Parametric Test - Chi-Square
- 8 Non-Parametric Test – Fisher’s Exact test
- 9 Non-Parametric Test – Wilcoxon-Mann-Whitney test

Parametric-Analysis of Variance (ANOVA)

ANOVA is used when you have a **categorical independent** variable (with two or more categories) and a dependent variable which is interval and you wish to test for differences in the means of the dependent variable across the levels of the independent variable.

Parametric-Analysis of Variance (ANOVA)

- In this example we test whether household receipt of remittance affect **weight** carried by child
- In this instance the dependent variable is interval whilst the independent is nominal with more than two outcome.
- Command in Stata

```
anova weight remit
```

Parametric-Analysis of Variance (ANOVA)

```
anova weight remit
```

```
Number of obs =    947    R-squared    =  0.0360
Root MSE      = 10.1878    Adj R-squared =  0.0329
```

Source	Partial SS	df	MS	F	Prob > F
Model	3652.21653	3	1217.40551	11.73	0.0000
remit	3652.21653	3	1217.40551	11.73	0.0000
Residual	97875.4709	943	103.791592		
Total	101527.687	946	107.323137		

Parametric-Analysis of Variance (ANOVA)

- The mean weight is significantly different across the various remittance group. This conclusion is drawn because the $P > F$ is less than 0.05.
- However we are unable to tell where the differences are.
- To obtain this we try

oneway weight remit, t bon sid sch

- 1 Parametric Test - Introduction
- 2 Parametric Test – T-test
- 3 Parametric Test – Analysis of Variance
- 4 Parametric Test – Regression – Least Squares
- 5 Parametric Test – Regression Cat. Dep. Var.
- 6 Non-Parametric Test - Introduction
- 7 Non-Parametric Test - Chi-Square
- 8 Non-Parametric Test – Fisher’s Exact test
- 9 Non-Parametric Test – Wilcoxon-Mann-Whitney test

Parametric Test - Regression

- The previous examples dealt with the bivariate relationship between two variables. In this section we consider relationships involving two or more variables using **Regression Analysis**
- Regression analysis is used to produce an equation that will predict a dependent variable using one or more independent variable(s).

Regression

- In this course we consider two basic models of regression analysis;
 - Linear Regression models - Ordinary Least Squares (OLS)
 - Categorical Dependent Variable Models - Logit and Probit Models
- OLS is use based on moments (mean, standard deviation and shape) – convenient for dependent variables that are continuous in nature
- Categorical dependent variable models are based on Maximum Likelihood Estimations (probability distribution functions) – convenient for dependent variables that are discrete

Parametric Test - Regression – Least Squares

Regression analysis is used to produce an equation that will predict a dependent variable using one or more independent variable(s).

$$y_i = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + e_i$$

??

The main thrust is to minimize the error term

Regression – Least Squares

- where:
- y is the dependent variable
- α is the constant
- β s are the coefficients to be estimated
- X s are the explanatory (independent) variables in the model and
- ϵ is the error term

Regression – Linear Models

- Linear regression is used to estimate the effect of a change in the independent variable on the dependent variable
- Assumptions
 - Linear in parameters
 - Random sampling
 - No perfect collinearity
 - Zero conditional mean
 - Homoskedasticity
 - Normality

Calculating

- Slope of the regression line is

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$$

- Precision of the estimate is given by the standard error

$$s.e.(\hat{\beta}_1) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- Where s is the standard error of the estimate

Testing

- Hypotheses:
 - $H_0: \beta_1 = 0$ (no significant slope)
 - $H_1: \beta_1 \neq 0$
- Student's t-test

$$t = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)}$$

- Confidence interval

$$\hat{\beta}_1 \pm t_{n-2} * se(\hat{\beta}_1)$$

Regression – Linear Models

- The command for linear regression in Stata is

```
regress y x1 x2 x3 ...xn
```

- Where y is the dependent variable and x_1 , x_2 , x_3 ... x_n are the set of explanatory variables
- Suppose we want to investigate the effects of **age** and **sex** of child on **weight** carried. We can achieve this by running a linear regression

Regression – Linear Models

- Linear regression is appropriate in this situation because the dependent variable (**weight**) is continuous. The command is

```
regress weight age sex
```

Or

```
reg weight age sex
```

- Note Stata always takes the first variable as the dependent variable

Parametric- Regression – Linear Models

Source	SS	df	MS			
Model	6605.81708	2	3302.90854	Number of obs =	960	
Residual	95863.3069	957	100.170645	F(2, 957) =	32.97	
Total	102469.124	959	106.849973	Prob > F =	0.0000	
				R-squared =	0.0645	
				Adj R-squared =	0.0625	
				Root MSE =	10.009	

weight	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	1.012099	.1295235	7.81	0.000	.757916	1.266282
sex	1.445205	.647075	2.23	0.026	.1753549	2.715054
_cons	7.644213	2.381121	3.21	0.001	2.971393	12.31703

Parametric- Regression – Linear Models

- In the next example, we investigate rural access to roads as measured by the variable `access_index` on GDP of a country .
- `CPI` and `population` are used as controls in the model
- The results are presented in the slide below

Parametric- Regression – Linear Models

```
gdp_constant_2005 rural_access_index total_pop CPI , robust
```

```
regression
```

```
Number of obs =      50
F(   3,   50) =   14.5
Prob > F      =   0.000
R-squared     =   0.291
Root MSE    =   5.1e+
```

	Robust				
constant_2005	Coef.	Std. Err.	t	P> t	[95% Conf. In
access_index	4.85e+11	2.13e+11	2.28	0.027	5.69e+10 9
total_pop	1364.286	234.7563	5.81	0.000	892.7645 1
CPI	1.22e+10	6.62e+09	1.84	0.071	-1.09e+09 2
_cons	-1.27e+12	6.53e+11	-1.94	0.059	-2.58e+12 4

Parametric- Regression – Linear Models

- The output shows that rural access has a positive and significant effect on gross domestic product.
- We fail to accept the hypothesis that the coefficient of rural access has no effect on GDP
- Population and CPI are also significant in explaining GDP.
- The three variables explain a third of the variation in GDP

Parametric- Regression – Linear Models

- **Number of obs** is the total number of observations use in the regression
- **$F(2,957)$** : is the joint significance of the variables included in the model. **The H_0 is that the explanatory variables are jointly not significant**
- **$P > F$** : is the p-value of the F test. In this case we fail to accept the null hypothesis that **age** and **sex** are jointly not significant in explaining **weight**

Parametric- Regression – Linear Models

- ***R-squared*** : is the variance in the dependent variable (weight) that is explained by the independent variables in the model. Age and sex explains about 6.5 per cent of the variability in weight.
- ***Adj R-squared*** : is the R-square which has been adjusted by the number of explanatory variables.
- ***Root MSE*** : root mean squared error, is the sd of the regression. The closer to zero better the fit.

Regression – Linear Models

- In addition to the regression descriptive, the command produces the regression output in a seven column table
- The first and second columns show the variables and their respective coefficients. The coefficient of age is approximately 1.012 and that of sex is 1.45. The last row is the regression constant.

☐ Regression – Linear Models

- The third column is the standard error of the coefficients
- The t-statistic is presented in column 4. this tests the H_0 that the coefficient is not different from zero. To reject this we need a t-value that is greater than 1.96.

Parametric- Regression – Linear Models

- Column 5 is the two tailed p-value of the coefficients. This also test the same null as the t-value. At five per cent alpha value we fail to accept the null hypothesis if the p-value is less than 0.05.
- The last two columns are the 95% confidence intervals. For instance we are 95% confident that the coefficient of age will lie between .758 and 1.267

Parametric- Regression – Linear Models

- Interpretation of results:
 - Before we can interpret a coefficient we need to be convinced that it is significantly different from zero. The output gives so many avenues to do that. We can use the t-values, p-values and the confidence interval to do that. In both case we realize that the p-values are less than 0.05, hence we conclude that **age** and **sex** are both significant in explaining **weight**

Parametric- Regression – Linear Models

- As noted earlier on a the coefficients measure the effect of a unit change in the independent variable on the dependent variable. In the case of age, one can say that an additional year increases weight by approximately 1.012kg when all other factors are held constant.
- On the effect of sex on weight, the interpretation is quite different. The variable sex is categorical (dummy) with **female coded as 0** and **male as 1**

Parametric- Regression – Linear Models

- The coefficient is interpreted as how much more or less weight does male carry compared to females. The result shows that all else being the same males have 1.45 more weight than females.
- The coefficient is interpreted as how much more or less weight does male carry compared to females. The result shows that all else being the same males have 1.45 more weight than females.

- 1 Parametric Test - Introduction
- 2 Parametric Test – T-test
- 3 Parametric Test – Analysis of Variance
- 4 Parametric Test – Regression – Least Squares
- 5 Parametric Test – Categorical Dep. Variables
- 6 Non-Parametric Test - Introduction
- 7 Non-Parametric Test - Chi-Square
- 8 Non-Parametric Test – Fisher’s Exact test
- 9 Non-Parametric Test – Wilcoxon-Mann-Whitney test

Regression – Binary Response

- The previous example dealt with a continuous dependent variable. However, there are instances when the
- Researchers are sometimes confronted with a categorical variable (binary, ordinal or nominal).
- When the dependent variable is categorical, OLS can no longer give us unbiased and consistent estimates.
- In this course we concentrate on binary categorical dependent variable

Regression – Binary Response

- If we have a binary categorical dependent variable, we either use **logit** or **probit** models to estimate it.
- The difference between **logit** and **probit** is beyond the scope of this course. However, these models, end up with almost the same standardized effects on the independent variables

Regression – Binary Response

- We want to find out factors that influence a child being in school. The dependent variable is dichotomous (0 and 1). The independent variables are age and sex of child, whether father is alive and settlement type.

- Let's summarize the data

```
su sch age sex father_alive settype
```

Regression – Binary Response

- We want to find out factors that influence a child being in school. The dependent variable is dichotomous (0 and 1). The dependent variables are age and sex of child, whether father is alive and settlement type.

- Let's summarize the data

```
su sch age sex father_alive settype
```

Regression – Binary Response

```
. su sch sex age father_alive settype
```

Variable	Obs	Mean	Std. Dev.	Min	Max
sch	548	.7609489	.4268937	0	1
sex	1000	.472	.4994652	0	1
age	1005	13.3194	2.512612	8	18
father_alive	1005	50.61493	703.5951	0	9996
settype	1005	1.59602	1.072094	0	3



```
. logit sch age sex father_alive i.settype, robust
```

```
Iteration 0:   log pseudolikelihood = -299.68541
Iteration 1:   log pseudolikelihood = -236.17954
Iteration 2:   log pseudolikelihood = -230.23642
Iteration 3:   log pseudolikelihood = -230.19325
Iteration 4:   log pseudolikelihood = -230.19324
```

Logistic regression

```
Number of obs   =           546
Wald chi2(6)    =           85.53
Prob > chi2     =           0.0000
Pseudo R2      =           0.2319
```

Log pseudolikelihood = -230.19324

sch	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.456662	.0517692	-8.82	0.000	-.5581278	-.3551963
sex	1.382022	.2616743	5.28	0.000	.8691499	1.894894
father_alive	.0000122	.0001276	0.10	0.924	-.000238	.0002624
settype						
1	.7460935	.390477	1.91	0.056	-.0192275	1.511414
2	.9972762	.3369815	2.96	0.003	.3368046	1.657748
3	1.265924	.3744027	3.38	0.001	.5321079	1.99974
_cons	6.242029	.741975	8.41	0.000	4.787785	7.696273

Regression – Binary Response

- The iteration log is an indication of how quickly the model converges
- The Wald chi-square of 85.53 with p-value of 0.000 tells us our model is significantly fit.
- The command also produces a table of coefficients, std errors, t-value etc. as in the case of the linear model.

Regression – Binary Response

- To obtain the odds ratio one can issue either of these commands

```
logit sch age sex father_alive settype, or
```

OR

```
logistic sch age sex father_alive settype
```

- We use the `logistic` command to get the odds ratio

- 1 Parametric Test - Introduction
- 2 Parametric Test – T-test
- 3 Parametric Test – Analysis of Variance
- 4 Parametric Test – Regression – Least Squares
- 5 Parametric Test – Categorical Dep. Variables
- 6 Non-Parametric Test - Introduction
- 7 Non-Parametric Test - Chi-Square
- 8 Non-Parametric Test – Fisher’s Exact test
- 9 Non-Parametric Test – Wilcoxon-Mann-Whitney test

NONPARAMETRIC TESTS - Introduction

- **Review of Parametric tests**
 - Parametric tests are designed to test specific population parameters **t-test** and **anova** are used to test hypotheses about μ population parameters. *E.g. Testing that the number of accidents in particular year is different from or equal to a hypothesized value*
 - these tests require (sometimes very strong) assumptions about the population distribution and other
 - They sometimes require numerical scores for each sample unit. They usually require data from an interval or ratio scale
 - Homogeneity of variance

NONPARAMETRIC TESTS - Introduction

- Real world situation confronts researchers with situations that do not conform to the assumption of parametric tests. *For example we wish to test the association between type of job and mode of transport among people living in a particular region*
- Because these assumption are violated it may not be appropriate to use parametric tests. In such situation we fall on **Non-parametric tests**
- Nature of Non-parametric tests
 - use sample data to test hypothesis about the proportions
 - use sample data to test hypothesis about relationships
 - do not test hypothesis in terms of specific parameters
 - not reliant on many assumption if any

- 1 Parametric Test - Introduction
- 2 Parametric Test – T-test
- 3 Parametric Test – Analysis of Variance
- 4 Parametric Test - Regression
- 5 Parametric Test – Categorical Dep. Variables
- 6 Non-Parametric Test - Introduction
- 7 Non-Parametric Test - Chi-Square
- 8 Non-Parametric Test – Fisher’s Exact test
- 9 Non-Parametric Test – Wilcoxon-Mann-Whitney test

CHI-SQUARE GOODNESS OF FIT TEST

This allows us to test if the observed proportion of a categorical variable in our sample is different from a hypothesized proportion. The test determines how well the obtained sample proportions fit the population proportions specified by the null hypothesis.

CHI-SQUARE GOODNESS OF FIT TEST

- *Example, suppose we believe that the general population consist of 2% no religion, 25% Muslim, 55% Christian, 18% traditional and we wish to test if the **RELIGN** variable in our data set follows this distribution*
- the stata command for the test is ***csgof*** which can be installed by typing *findit csgof* in Stata's command window
-

CHI-SQUARE GOODNESS OF FIT TEST

- After a successful installation of **csgof**
- **csgof RELIGN, expperc(2 25 55 18)** *make sure the order of the hypothesized proportions are presented in the same order as in the sample*
- the null for the test is that "**the composition of our sample does not differ significantly from the hypothesized values**"

CHI-SQUARE GOODNESS OF FIT TEST

```
. csgof RELIGN, expperc(25 2 55 18)
```

RELIGN	expperc	expfreq	obsfreq
no religio	25	250.25	12
Muslim	2	20.02	112
Christian	55	550.55	872
traditiona	18	180.18	5

```
chisq(3) is 1007.42, p = 0
```



CHI-SQUARE GOODNESS OF FIT TEST

- The command produces a table of 4 columns with an additional information on the chi and probability values
- The second column is our hypothesized proportions, the third are the expected frequencies based on these values and the last is the set of observed frequencies in the sample. For example we hypothesized that 25% are Muslims, hence the expected frequency is calculated as $\frac{25}{100} \cdot 100$

CHI-SQUARE GOODNESS OF FIT TEST

- Since $p < 0.001$, we fail to accept H_0 and conclude that the composition of our sample is statistically different from the hypothesized values.

CHI SQUARE OF INDEPENDENCE

- The chi-square statistic may be used to test whether or not there is a relationship between two categorical variables.
- The chi-square statistic is given as:

$$c^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

CHI SQUARE OF INDEPENDENCE

- Assumptions

- **Independence of observation:** that is each observed frequency is generated by one observation. A chi-square test will be inappropriate if a person could produce responses that can be classified in more than one category or contributes more than one frequency count to a single category
- **Size of Expected Frequencies:** a chi-square test will not be appropriate if the expected frequency of any cell is less than 5

CHI SQUARE OF INDEPENDENCE

- Suppose we want to test if the sex of a child can be used to predict whether the child was absent from school last week or not. In other words we are testing whether absenteeism last week depends on child's sex.
-
- Two variables are needed, one that captures child sex (**sex**) and other that captures absenteeism last week (**lwabsent**). These are two categorical variables so we can use Chi-square to test this hypothesis.

CHI SQUARE OF INDEPENDENCE

- We can conduct this test by simply issuing the command

```
tab lwabasent sex, chi
```

- H_0 : Absenteeism is not dependent on child sex

CHI SQUARE OF INDEPENDENCE

```
. tab lwabsent sex, chi
```

selected child attended sch last week?	sex Sex of selected child		Total
	F	M	
Yes	420	420	840
No	30	22	52
Total	450	442	892

Pearson chi2 (1) = 1.1591 Pr = 0.282



CHI SQUARE OF INDEPENDENCE

- Chi-square is appropriate because none of the cells have frequency of less than 5
- These results indicate that there is no statistically significant relationship between absenteeism and child sex (chi-square with one degree of freedom = 1.1591 , $p = 0.282$).

CHI SQUARE OF INDEPENDENCE

Let's look at one more example of the chi-square test

- This time we want to test if there is an association between sex and the use of bike. Our two variables are sex of child (**sex**) and how often the child uses bike (**bike**)
- We again specify the command as

```
tab bike sex, chi
```

CHI SQUARE OF INDEPENDENCE

```
. tab bike sex, chi
```

Does the child use bike and how often	sex Sex of selected child		Total
	F	M	
use every day	10	47	57
use 4-6 times / week	32	50	82
use 1-3 times / week	82	103	185
use less than once a never use	167	196	363
	236	75	311
Total	527	471	998

Pearson chi2(4) = 113.2309 Pr = 0.000



CHI SQUARE OF INDEPENDENCE

- In this example, we fail to accept the H_0 that Frequency of bike usage is independent of child's sex (chi-square with 4 degree of freedom = 113.2309, $p = 0.000$). In other words, knowing a child's sex can help us to predict the frequency of bicycle usage.
- Now what test can we conduct if we have at one cell having a frequency of less than 5?

- 1 Parametric Test - Introduction
- 2 Parametric Test – T-test
- 3 Parametric Test – Analysis of Variance
- 4 Parametric Test - Regression
- 5 Parametric Test – Categorical Dep. Variables
- 6 Non-Parametric Test - Introduction
- 7 Non-Parametric Test - Chi-Square
- 8 Non-Parametric Test – Fisher’s Exact test
- 9 Non-Parametric Test – Wilcoxon-Mann-Whitney test

FISHER'S EXACT TEST

- Fisher's Exact test is used when the frequency assumption makes it inappropriate to use chi-square test. In other words when you have at least a cell with less than 5 observations.
- Note: Fisher's exact test has no such assumption and can be used irrespective of how small the expected frequency is.

FISHER'S EXACT TEST

- In the example below we have cells with observed frequencies of less than 5 (e.g. 0, 1 and 2). Hence we use Fisher's exact test to test whether birthord (**birthord**) determines the frequency of bus usage (**bus**)
- The command for Fisher's exact test in Stata is
`tab bus birthord, exact`

FISHER'S EXACT TEST

bus	birthord					Total
	1	2	3	4	5	
use 1-3 times / week	0	0	0	1	0	1
use less than once a	0	2	0	1	0	3
never use	121	418	134	295	20	988
Total	121	420	134	297	20	992

Fisher's exact = 0.932

FISHER'S EXACT TEST

- Unlike the chi-square test, Fisher's exact does not produce any test statistic but only the p-value of the test.
- In our example above the p-value is 0.932, hence we fail to reject the null hypothesis that birth order and frequency of bus usage are independent of each other.

FISHER'S EXACT TEST

- We use Fisher's exact in the next example to test whether the time it takes to school (**time_sch**) influences the frequency of bus usage (**bus**).
- The command for Fisher's exact test in Stata is

```
tab bus time_sch, exact
```

FISHER'S EXACT TEST

```
. tab bus time_sch, exact
```

```
Enumerating sample-space combinations:
```

```
stage 4: enumerations = 1
```

```
stage 3: enumerations = 0
```

```
stage 2: enumerations = 0
```

```
stage 1: enumerations = 0
```

bus	time_sch				Total
	15 mins o	16-45 min	46 mins t	1hr 31min	
use less than once a	2	1	0	0	3
never use	404	257	163	21	845
Total	406	258	163	21	848

```
Fisher's exact = 1.000
```

FISHER'S EXACT TEST

- In the example above the p-value is 1.000, hence we fail to reject the null hypothesis that time to school and frequency of bus usage are independent of each other.

- 1 Parametric Test - Introduction
- 2 Parametric Test – T-test
- 3 Parametric Test – Analysis of Variance
- 4 Parametric Test - Regression
- 5 Parametric Test – Categorical Dep. Variables
- 6 Non-Parametric Test - Introduction
- 7 Non-Parametric Test - Chi-Square
- 8 Non-Parametric Test – Fisher’s Exact test
- 9 Non-Parametric Test – Wilcoxon-Mann-Whitney test

Wilcoxon-Mann-Whitney test

This is the non-parametric version of the independent samples t-test and. It is used when we do not assume that dependent variable is normally distributed. One only assumes that the dependent variable is at least ordinal. In the next example we wish to test the null hypothesis that the mean age of male children is not statistically different from that of female children

Wilcoxon-Mann-Whitney test

- We assume that age is not normally distributed
- the variables of interest in the sample are age of child (**AGE**) and sex of child (**sex**)
- Using the same data we conduct this test as

```
ranksum AGE, by(sex)
```

Wilcoxon-Mann-Whitney test

```
. ranksum AGE, by(sex)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test
```

sex	obs	rank sum	expected
F	528	263380.5	264264
M	472	237119.5	236236
combined	1000	500500	500500

```

unadjusted variance      20788768
adjustment for ties     -274183.38
-----
adjusted variance       20514585

Ho: AGE (sex==F) = AGE (sex==M)
      z = -0.195
      Prob > |z| = 0.8453

```

Wilcoxon-Mann-Whitney test

- The results suggest that there is no statistically significant difference between the underlying distributions of the age of males and that of females ($z = -0.195$, $p = 0.8453$).

Summary of Parametric and Non-parametric test

Non-Parametric	Rationale	Parametric	Rationale
Wilcoxon Signed Ranks Tests	Tests hypothesis about the median of the population a sample was taken from	One Sample T-Test	To test a hypothesis about the mean of the population a sample was taken from
Wilcoxon Rank-Sum Test	Compare to samples to find out if they have identical population medians	Two Sample T-Tests	Tests if two samples have identical population means
Chi-Squared Test	Examine differences across cells of a table	Z-test, F-tests	
Kolmogorov-Smirnov Test	Explore the likelihood that the sample is emanating from a certain distribution	ANOVA	Find out if two or more sample means are significantly different.
Kruskal-Wallis Test	Find out if two or more sample medians are significantly different	Linear Correlation	Examine the direction of relationship for continuous data
Spearman Rank Correlation	Examine the direction of relationship for rank data	Regression	



Useful web addresses for Stata notes

- <http://www.ats.ucla.edu/stat/stata/whatstat/>
- http://www.ssc.wisc.edu/sscc/pubs/stata_students2.htm
- <http://data.princeton.edu/stata/default.html>
- <http://fmwww.bc.edu/GStat/docs/StataIntro.pdf>
- http://www.wiwi.uni-muenster.de/ioeb/Downloads/Forschen/Pfaff/Introduction_to_Stata_with_50+_Basic_Commands.pdf





Now read
Session 7.2
Notes!