

UCLA STAT 13
**Introduction to Statistical Methods for
the Life and Health Sciences**

● **Instructor:** Ivo Dinov,
Asst. Prof. In Statistics and Neurology

● **Teaching Assistants:**
Ming Zheng, Annie Che
UCLA Statistics

University of California, Los Angeles, Winter 2004
http://www.stat.ucla.edu/~dinov/courses_students.html

STAT 13, UCLA, Ivo Dinov Slide 1

Chapter 11: Tables of Counts

We discussed means and mean differences in Ch. 10 and developed a statistical toolbox for analyzing quantitative variables.

Now we want to develop a similar approach for analyzing qualitative variables.

Table-of-measurements → tables-of-counts;
Means → proportions
T/F-tests for inference on qualitative variables →
Chi-square (χ^2) tests for categorical data.

STAT 13, UCLA, Ivo Dinov Slide 2

Chapter 11: Tables of Counts

- One-dimensional tables and goodness of fit
- Two-way tables of counts
Chi-square test of homogeneity
Chi-square test of independence
2 by 2 tables
- The perils of collapsing tables

STAT 13, UCLA, Ivo Dinov Slide 3

**1-dimensional tables –
classify n -individuals in J -categories**

Qualitative (factors), class variables define class/group membership (marital-status, blood-type, etc.)

Frequency tables can be used to Summarize discrete/qualitative var's.

Category ...	P_1	P_2	...	P_j	...	P_J
Probability	O_1	O_2	...	O_j	...	O_J
Observed count	E_1	E_2	...	E_j	...	E_J
Expected count	$E_j = n p_j$					

STAT 13, UCLA, Ivo Dinov Slide 4

1-dimensional tables cont.

expected cell count = total × specified cell probability

The T and F statistics are used for inference about quantitative variables. χ^2 statistics is used for analysis of categorical data.

- When H_0 gives the probabilities of landing in each cell completely (no parameters to be estimated), $P(\text{cell}_1)=p_1, P(\text{cell}_2)=p_2, \dots, P(\text{cell}_j)=p_j$, and $\sum p_k=1$.
- Thus, having $J-1$ probabilities fixed determines the last probability.

$df = \text{number of categories} - 1$

STAT 13, UCLA, Ivo Dinov Slide 5

Chi-Square Test – goodness of fit test

- The Chi-square test statistic (χ^2) has observed value

$$\chi_0^2 = \sum_{\text{all cells in the table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

- The P -value for the test is
 $P\text{-value} = pr(X^2 \geq \chi_0^2)$ where $X^2 \sim \text{Chi-square}(df)$

To test a null-hypothesis, H_0 , we compare the observed counts in the table to the expected (theoretical) counts. For this reason this test is called a goodness-of-fit test – observed/expected count fit.

STAT 13, UCLA, Ivo Dinov Slide 6

Example of 1D table – Three blood types

TABLE 11.1.1 Proportions of Three Blood Types

	A	AB	B	Total
No. Observed	39	70	42	151
Proportion Observed	0.258	0.464	0.278	1.000

Slide 7 STAT 13, UCLA, Jon Dineen

Example of 1D table – rolling a single die

TABLE 11.1.2 210 Rolls of a Die

Outcome	1	2	3	4	5	6	Total
Count	26	40	37	26	43	38	210
Proportion	0.124	0.190	0.176	0.124	0.205	0.181	1.000

Why aren't these probabilities all equal?!?
 Are they supposed to?
 What are the expected probabilities (1/6)?
 χ^2 statistics is $\chi_0=7.54$, $df=5$, $P\text{-value}=0.18$

Slide 8 STAT 13, UCLA, Jon Dineen

Exit poll – sampling voters as they leave polling booths. Exit polls of 10,000 voters.

(a) Table of exit-poll sample and population Age distributions

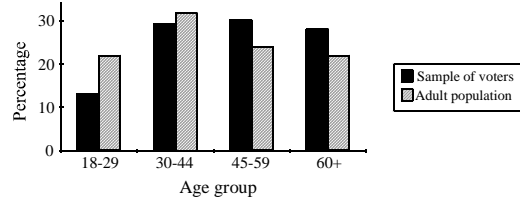
	Age group				Total
	18-29	30-44	45-59	60+	
Sample : (Percentages)	13	29	30	28	100
Population : (Percentages)	22	32	24	22	100

Are there differences between the population age groups and the exit-poll sample age groups?
 Younger voter underrepresented voters.
 Real differences or just due to sampling variation?

Slide 9 STAT 13, UCLA, Jon Dineen

Exit poll – Bar-plot of population/sample groups

(b) Plot of exit-poll sample and population Age distributions



H_0 : True proportions in the 4 age groups in the voter sample and the whole population are the same!

Slide 10 STAT 13, UCLA, Jon Dineen

Exit poll – Bar-plot of population/sample groups

(c) Table of observed and expected counts

	Age group				Total
	18-29	30-44	45-59	60+	
Observed count	1300	2900	3000	2800	10,000
Expected count	2200	3200	2400	2200	10,000

(Note: Counts approximate due to the rounding of percentages in the report.)

Figure 11.1.1 Comparing the age distributions for voters and the population.

H_0 : $p_{18-29} = 0.22$; $p_{30-34} = 0.32$; $p_{45-59} = 0.32$; $p_{60+} = 0.32$;

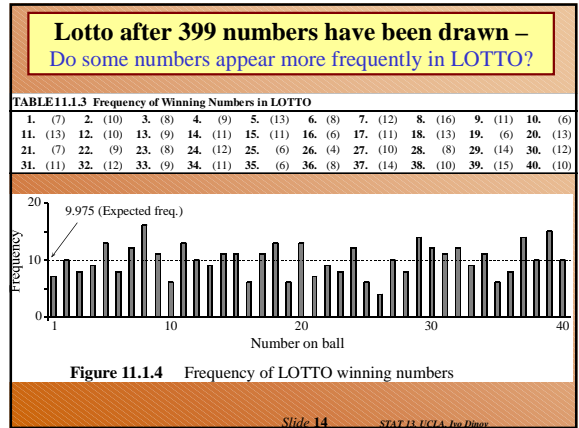
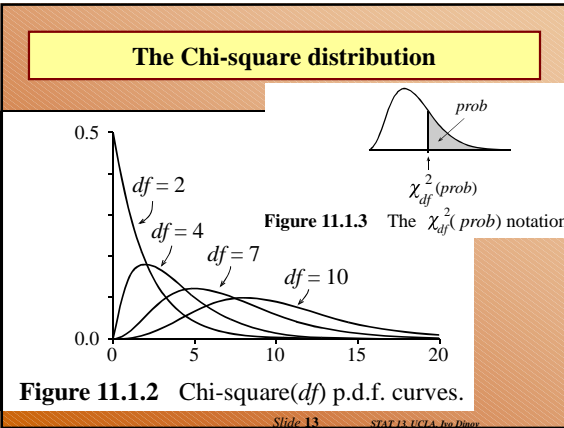
Slide 11 STAT 13, UCLA, Jon Dineen

Exit poll – Bar-plot of population/sample groups

$$\chi_0^2 = \sum_{\text{all cells in the table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = 709.94$$

$df = \text{number of groups} - 1 = 4 - 1 = 3$
 $P\text{-value} = 0.000$, very small, indicating extremely strong evidence against the null-hypothesis. The 95% CI for each age groups are:
 [12.3 : 13.7]; [28.1 : 30.0]; [29.1 : 30.9]; [27.1 : 28.9]

Slide 12 STAT 13, UCLA, Jon Dineen



Lotto after 399 numbers have been drawn – Do some numbers appear more frequently in LOTTO?

Number-range: [1:40]
 Number of balls selected at each draw: 7
 Number of samples: 57
 Total number of balls selected: $57 \times 7 = 399$,
 Expected value of each number: $399/40 = 9.975$
 Observed χ^2 statistics is $\chi_0^2 = 30.97$
 $df = 40 - 1 = 39$
 P-value = 0.817
 Conclusion: No evidence for departure from the null hypothesis.

Slide 15 STAT 13, UCLA, Joe Dimez

Review

- The test statistic for the Chi-square test compares observed and expected frequencies. In what sense are the *expected* frequencies expected? (Expected frequencies are the frequencies expected in H_0 were true.)
- What shape does the Chi-square distribution generally have? What happens to its shape as the degrees of freedom increase? (Skewed unimodal, becomes symmetric and Normal approximates it well for large df .)
- What values of the Chi-square test statistic (large or small) provide evidence against the null hypothesis? Why? (Large values, since P-value is small. See density curve.)

Slide 16 STAT 13, UCLA, Joe Dimez

Review

- For one-dimensional tables, how do you compute the degrees of freedom df ? ($df = \text{number of cells/groups} - 1$.)
- Do the expected counts have to be whole numbers? (No, expected counts = number of samples \times cell-probability.)

Slide 17 STAT 13, UCLA, Joe Dimez