Lecture 10

Categorical Data Contingency Tables, and Non-parametric Methods

Categorical data

- We will consider statistical problems based on data such that each observation can be classified as belonging to one of a finite number of possible categories or types. Observations of this type are called categorical data.

An example of blood type

- Counts of blood types in a number of people in an area:

Type	A	B	AB	O
Obs. Number	2162	738	228	2876

- How can we test the null hypothesis that the theoretical probabilities are the probabilities with which the observed data were sampled?

Type	A	B	AB	O
Theo. Probability	1/3	1/8	1/24	1/2

Test of Goodness of Fit

The χ^2 test

- Suppose that a large population consists of items of k different types, and let p_i denote the probability that an item selected at random will be of type i (i=1, …, k).

$$p_i \geq 0, i = 1, \cdots, k; \sum_{i=1}^{k} p_i = 1$$

- Let $p_1^0, \cdots, p_k^0$ be specific numbers such that

$$p_i^0 > 0, i = 1, \cdots, k; \sum_{i=1}^{k} p_i^0 = 1$$

The χ^2 test

- Suppose that the following hypotheses are to be tested:
- H_0: $p_i = p_i^0$, for i=1,…,k
- H_1: $p_i \neq p_i^0$, for at least one value of i
- We assume that a random sample of size n is to be taken from the given population. N_i denote the number of observations in the random sample that of type i.

$$\sum_{i=1}^{k} N_i = n$$

The χ^2 statistic

- The following statistic

$$Q = \sum_{i=1}^{k} \frac{\left(N_i - np_i^0\right)^2}{np_i^0}$$

has the property that if H_0 is true and the sample size n $\rightarrow \infty$, then Q converges in the distribution to the χ^2 distribution with *k*-1 degrees of freedom.

Goodness of fit

- The goodness of fit statistic, denoted by χ^2 as a sample statistic, is

$$Q = \sum_{i=1}^{k} \frac{\left(N_i - np_i^0\right)^2}{np_i^0}$$

- i.e. $\sum_{\text{All cells}} \frac{(\text{Observed cell count - Expected cell count})^2}{\text{Expected cell count}}$

- The χ^2 is the sum of the quantities for all *k* cells

The blood type example

- Blood types in a number of people in an area. H_0: $p_i = p_i^0, i = 1, \cdots, 4$

Type	A	B	AB	O	Total
p_i^0	1/3	1/8	1/24	1/2	1
N_i	2162	738	228	2876	6004

- The four expected counts under H_0 are

$$np_1^0 = 6004 \times \frac{1}{3} = 2001.3, np_2^0 = 6004 \times \frac{1}{8} = 750.5$$

$$np_3^0 = 6004 \times \frac{1}{24} = 250.2, np_4^0 = 6004 \times \frac{1}{2} = 3002.0$$

The blood type example

- The χ^2 test statistic is then

$$Q = \frac{(2162 - 2001.3)^2}{2001.3} + \frac{(738 - 750.5)^2}{750.5} + \frac{(228 - 250.2)^2}{250.2} + \frac{(2876 - 3002.0)^2}{3002.0} = 20.37$$

- In the case of χ^2 goodness of fit test, the p-value equals $1 - \chi^2_{k-1}(Q)$
- Here k=4, p-value is 1.42×10^{-4}. If the significance level α=0.05, we will reject H_0.

Likelihood Ratio Tests for Proportions

Likelihood ratio tests for proportions

- Although χ^2 tests are commonly used in such examples, we could actually use parametric tests in these examples.
- H_0: $p_i = p_i^0$, for i=1,…,k
- H_1: $p_i \neq p_i^0$, for at least one value of i
- So the likelihood function is

$$f(\mathbf{x}\,|\,\mathbf{p}) = \binom{n}{N_1,\cdots,N_k} p_1^{N_1} \cdots p_k^{N_k}$$

Likelihood ratio tests for proportions

- If H_0 is true, there is only one possible for the likelihood function, namely

$$\binom{n}{N_1,\cdots,N_k}\left(p_1^0\right)^{N_1}\cdots\left(p_k^0\right)^{N_k}$$

- It is not difficult to show that the MLE of p_i is $\hat{p}_i = N_i/n, i = 1,\cdots,k$
- The large-sample likelihood ratio test statistic is then

$$-2\log\Lambda(\mathbf{x}) = -2\sum_{i=1}^{k} N_i \log\left(\frac{np_i^0}{N_i}\right)$$

Likelihood ratio tests for proportions

- The large-sample test reject H_0 at level of significance of α_0 if this statistic is greater than the $1-\alpha_0$ quantile of the χ^2 distribution with k-1 degrees of freedom.
- In LRT, k-1 can be seen as the difference in the number of independent parameters to be estimated under the two hypotheses. 0 under H_0, k-1 under H_A, so $df=k$-1.

The example on blood type

- Blood types (continued). The values of np_i^0 has been calculated.
- Then the likelihood ratio test statistic is

$$-2\log\Lambda(\mathbf{x}) = -2\left[2162\log\left(\frac{2001.3}{2162}\right) + 738\log\left(\frac{750.5}{738}\right) + 228\log\left(\frac{250.2}{228}\right) + 2876\log\left(\frac{3002.0}{2876}\right)\right] = 20.16$$

- The p-value is the probability that a χ^2 random variable with three degrees of freedom is greater than 20.16, namely $1.57\text{x}10^{-4}$. This is nearly the same as the p-value from the χ^2 test.

An example in genetics: A resistance gene is linked with a molecular marker

F2 population	Resistant			Susceptible		
Marker type	A	H	B	A	H	B
Sample size	572	1161	14	3	22	569

Marker types A and B are parental types; H is the type of F1 hybrid

Resistant and susceptible can be fitted by the 3:1 ratio (one dominance gene locus): χ^2=0.17 (P=0.68). Marker types A, H, and B can be fitted by the 1:2:1 ratio (one co-dominance gene locus) : χ^2=0.32 (P=0.57)

But Resistance and Marker are not independent, i.e. can not be fitted by the 3:6:3:1:2:1 ratio.

Goodness-of-Fit for Composite Hypotheses

Goodness-of-fit for composite hypotheses

- We can extend the goodness-of-fit test to deal with the case in which the null hypothesis is that the distribution of our data belongs to a particular parametric family. The alternative hypothesis is that the data have a distribution that is not a member of that parametric family.
- There are two changes to the test procedure in going from the case of a simple null hypothesis to the case of a composite null hypothesis. First, in the test statistic Q, the probabilities p_i^0 are replaced by estimated probabilities based on the parametric family. Second, the degrees of freedom are reduced by the number of parameters.

The χ^2 test for composite null hypotheses

- H_0: there exists a value of $\theta \in \Omega$ such that $p_i = \pi_i(\theta)$, for i=1,…,k
- H_1: the hypothesis H_0 is not true
- Here $\theta=(\theta_1, \theta_2,\ldots, \theta_s)$ where $s<k$-1.
- The assumption that $s<k$-1 guarantees that the hypotheses H_0 actually restricts the values of p_1, …, p_k to a proper subset of the set of all possible values of these probabilities.

χ^2 Statistic for composite null hypotheses

- The following statistic

$$Q = \sum_{i=1}^{k} \frac{\left(N_i - n\pi_i(\hat{\theta})\right)^2}{n\pi_i(\hat{\theta})}$$

has the property that if H_0 is true and the sample size n $\rightarrow \infty$, then Q converges in the distribution to the χ^2 distribution with *k*-1-*s* degrees of freedom.

Example

- Genetics. Consider a gene that has two different alleles. Each individual in a given population must have one of three possible genotypes. If the alleles arrive independently from the two parents, and if every parent has the same probability θ of passing the first allele to each offspring, then the probabilities p_1, p_2 and p_3 of the three different genotypes are

$$p_1 = \theta^2, p_2 = 2\theta(1-\theta), p_3 = (1-\theta)^2$$

Example

- It is desired to test the null hypothesis H_0 that the probabilities p_1, p_2 and p_3 can be represented in this form.
- For this data, k=3 and s=1.
- Therefore, when H_0 is true, the distribution of the statistic Q defined by

$$Q = \sum_{i=1}^{k} \frac{\left(N_i - n\pi_i(\hat{\theta})\right)^2}{n\pi_i(\hat{\theta})}$$

which will be approximately the χ^2 distribution with 1 degrees of freedom.

Determining the maximum likelihood estimates

- When H_0 is true, the likelihood function L(θ) for the observed numbers $N_1, \ldots, N_k$ will be

$$L(\theta) = \binom{n}{N_1, \cdots, N_k} [\pi_1(\theta)]^{N_1} \cdots [\pi_k(\theta)]^{N_k}$$

- Thus,

$$\log L(\theta) = \log \binom{n}{N_1, \cdots, N_k} + \sum_{i=1}^{k} N_i \log \pi_i(\theta)$$

- The MLE $\hat{\theta}$ will be the value of θ for which log L(θ) is a maximum. The multinomial coefficient does not affect the maximization and we shall ignore it.

Example

- Genetics (continued). *k*=3.

$$\log L(\theta) = \sum_{i=1}^{3} N_i \log \pi_i(\theta)$$

$$= N_1 \log\left(\theta^2\right) + N_2 \log\left(\theta(1-\theta)\right) + N_3 \log\left((1-\theta)^2\right)$$

$$= \left(2N_1 + N_2\right)\log\theta + \left(2N_3 + N_2\right)\log(1-\theta) + N_2 \log 2$$

- It can be founded by differentiation that the value of θ for which log L(θ) is a maximum is

$$\hat{\theta} = \frac{2N_1 + N_2}{2\left(N_1 + N_2 + N_3\right)}$$

- The value of the statistic Q can now be calculated from the observed numbers N_1, N_2 and N_3.

An example in MN blood type

We want to test if the population in HW equilibrium: If (M, N)=(p, q), then (MM, MN, NN)=(p^2, $2pq$, q^2)

Genotype	MM	MN	NN	Total
Sample size	N_1=233	N_2=385	N_3=129	N=747
Predicted by Hardy-Weinberg Law	242.70	366.18	138.38	747

- (p, q) = (0.57, 0.43)
- (p^2, $2pq$, q^2) = (0.3249, 0.4902, 0.1849)
- χ^2=1.96 (df=1, P=0.16), i.e. the population is in HW equilibrium

Contingency Tables

Contingency tables

- When each observation in our sample is a bivariate discrete random vector (a pair of discrete random variables), then there is a simple way to test the hypothesis that the two random variables are independent. The test is another form of χ^2 test.

Definition

- A table in which each observation is classified in two or more ways is called a contingency table.
- For example, a two-way contingency table.

	Candidate preferred			
Curriculum	**A**	**B**	**Undecided**	Totals
Engineering and science	24	23	12	59
Humanities and social sciences	24	14	10	48
Fine arts	17	8	13	38
Industrial and public administration	27	19	9	55
Totals	92	64	44	200

Two-way contingency table

- Consider a two-way contingency table containing R rows and C columns. For $i=1,\ldots,R$ and $j=1,\ldots,C$, let p_{ij} denote the probability that an individual selected at random from a given population will be classified in the i^{th} row and the j^{th} column of the table.

$$p_{i+} = \sum_{j=1}^{C} p_{ij}, \; p_{+j} = \sum_{i=1}^{R} p_{ij}$$

$$\sum_{i=1}^{R}\sum_{j=1}^{C} p_{ij} = \sum_{i=1}^{R} p_{i+} = \sum_{j=1}^{C} p_{+j} = 1$$

Two-way contingency table

- Let N_{ij} denote the number of individuals who are classified in the i^{th} row and the j^{th} column of the table.

$$N_{i+} = \sum_{j=1}^{C} N_{ij}, N_{+j} = \sum_{i=1}^{R} N_{ij}$$

$$\sum_{i=1}^{R}\sum_{j=1}^{C} N_{ij} = \sum_{i=1}^{R} N_{i+} = \sum_{j=1}^{C} N_{+j} = n$$

- Hypotheses to be tested:
- H_0: $p_{ij}=p_{i+}p_{+j}$, for $i=1,\ldots,R$, $j=1,\ldots,C$
- H_1: the hypothesis H_0 is not true

The χ^2 test of independence

- Under H_0, the unknown parameters p_{ij} of these RC cells have been expressed as functions of the unknown parameters p_{i+} and p_{+j} .
- Since $\sum_{i=1}^{R} p_{i+} = \sum_{j=1}^{C} p_{+j} = 1$

the actual number of unknown parameters to be estimated when H_0 is true is $s=(R-1)+(C-1)=R+C-2$

The χ^2 test of independence

- Let $\hat{E}_{ij}$ denote the MLE of the expected number of observations that will be classified in the *i*th row and the *j*th column of the table when H_0 is true.

$$Q = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{\left(N_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}}$$

- Q has the property that if H_0 is true and the sample size n $\rightarrow \infty$, then Q converges in the distribution to the χ^2 distribution with *RC*-1-*s*=(*R*-1)(*C*-1) degrees of freedom.

Estimator $\hat{E}_{ij}$

- When H_0 is true, $p_{ij}=p_{i+}p_{+j}$
- Let $\hat{p}_{i+}$ and $\hat{p}_{+j}$ denote the MLE of p_{i+} and p_{+j}. Then

$$\hat{E}_{ij} = n\hat{p}_{i+}\hat{p}_{+j} = n\left(\frac{N_{i+}}{n}\right)\left(\frac{N_{+j}}{n}\right) = \frac{N_{i+}N_{+j}}{n}$$

- Substitute this value into the equation of Q. The null hypothesis H_0 should be rejected if Q>=*d*, where *d* is an appropriately chosen constant.

Example

- College survey (continued). From the data, we know that $N_{1+}=59$, $N_{2+}=48$, $N_{3+}=38$ and $N_{4+}=55$. Also $N_{+1}=92$, $N_{+2}=64$ and $N_{+3}=44$. n=200, $R=4$ and $C=3$. Expected cell counts are

	Candidate preferred			
Curriculum	A	B	undecided	Totals
Engineering and science	27.14	18.88	12.98	59
Humanities and social sciences	22.08	15.36	10.56	48
Fine arts	17.48	12.16	8.36	38
Industrial and public administration	25.30	17.60	12.10	55
Totals	92	64	44	200

Example

- Compare the expected values with the true values N_{ij}.

$$Q = \frac{(24-27.14)^2}{27.14} + \frac{(23-18.88)^2}{18.88} + \frac{(12-12.98)^2}{12.98}$$
$$+ \frac{(24-22.08)^2}{22.08} + \frac{(14-15.36)^2}{15.36} + \frac{(10-10.56)^2}{10.56}$$
$$+ \frac{(17-17.48)^2}{17.48} + \frac{(8-12.16)^2}{12.16} + \frac{(13-8.36)^2}{8.36}$$
$$+ \frac{(27-25.30)^2}{25.30} + \frac{(19-17.60)^2}{17.60} + \frac{(9-12.10)^2}{12.10} = 6.68$$

Example

- Since R=4 and C=3, the corresponding tail area is to be found from a table of χ^2 distribution with (R-1)(C-1)=6 degrees of freedom.
- P-value is 0.35. Therefore, we would only reject H_0 at level α_0 if α_0>=0.3. That is, we cannot say the candidate preference depends on curriculum.

Simpson’s Paradox

Simpson's Paradox

- When tabulating discrete data, we need to be careful about aggregating groups.
- Suppose that a survey has two questions. If we construct a single table of responses to the two questions that includes both men and women, we might get a very different picture than if we construct separate tables for the responses of men and women.

An example of the paradox

- Results of experiment comparing two treatments

All patients	Improved	Not improved	Percent improved
New treatment	20	20	50
Standard treatment	24	16	60

- Disaggregated by sex

Men only	Improved	Not improved	Percent improved
New treatment	12	18	40
Standard treatment	3	7	30

Women only	Improved	Not improved	Percent improved
New treatment	8	2	80
Standard treatment	21	9	70

An example of the paradox

- According to the first table, the new treatment is superior to the standard treatment both for men and for women,
- According to the second and third tables, the new treatment is inferior to the standard treatment when all the subjects are aggregated.
- This type of result is known as Simpson’s paradox.

The paradox explained

- In the example, women have a higher rate of improvement from the disease than men have, regardless of which treatment they receive.
- Furthermore, most of the women in the sample receive the standard treatment while most of the men received the new treatment.

The paradox explained

- The new treatment looks bad in the aggregated table because most of the people who weren't going to respond well to either treatment got the new treatment while most of the people who were going to respond well to either treatment got the standard treatment.
- Even though the numbers of men and women in the experiment were equal, a high proportion of the women and a low proportion of men received the standard treatment.
- Since women have a much higher rate of improvement than men, it is found in the aggregated table than the standard treatment manifests a higher overall rate of improvement than does the new treatment.

The paradox explained

- Simpson's paradox demonstrates dramatically the dangers in making inferences from an aggregated table.
- **To avoid the paradox, the proportions of men and women among the subjects who receive the new treatment must be the same, or approximately the same, as the proportions of men and women among the subjects who receive the standard treatment.**
- It is not necessary that there be equal numbers of men and women in the sample.

Express Simpson's Paradox in probability terms

- A: the event that a subject chosen for the experiment will be a man.
- A^C: the event that the subject will be a woman.
- B: the event that a subject will receive the new treatment.
- B^C: the event that the subject will receive the standard treatment.
- I: the event that a subject will improve.

Express Simpson's paradox in probability terms

- Simpson's paradox reflects the fact that it is possible for all three of the following inequalities to hold simultaneously:

$$\Pr\left(I \middle| A \bigcap B\right) > \Pr\left(I \middle| A \bigcap B^C\right)$$

$$\Pr\left(I \middle| A^C \bigcap B\right) > \Pr\left(I \middle| A^C \bigcap B^C\right)$$

$$\Pr\left(I \middle| B\right) < \Pr\left(I \middle| B^C\right)$$

Express Simpson's paradox in probability terms

- The discussion that we have just given in regard to the prevention of Simpson's paradox can be expressed as follows:
- If $\Pr(A|B) = \Pr(A|B^C)$, then it is not possible for all three inequalities to hold.
- Similarly, if $\Pr(B|A) = \Pr(B|A^C)$, then it is not possible for all three inequalities to hold.

Summary

- Simpson’s paradox occurs when the relationship between the two categorical variables in every part of a disaggregated table is the opposite of the relationship between those same two variables in the aggregated table.
- Be careful with “structure” for pooling or combined analysis!

Nonparametric Statistics

Nonparametric statistics

- In some problems, we have one specific distribution in mind of the data we will observe.
- If that one distribution is not appropriate, we do not necessarily have a parametric family of alternative distributions in mind.
- In these cases, and others, we can still test the null hypothesis that the data come from the one specific distribution against the alternative hypothesis that the data do not come from that distribution.

Nonparametric statistics

- We shall not assume that the available observations come from a particular parametric family of distributions.
- Rather, we shall study inferences that can be made about the distribution from which the observations come, without making special assumptions about the form of that distribution.

Example

- As one example, we might simply assume that the observations form a random sample from a continuous distribution, without specifying the form of this distribution any further, and we might then investigate the possibility that this distribution is a normal distribution.

Definition

- Problems in which the possible distributions of the observations are not restricted to a specific parametric family are called nonparametric problems.
- The statistical methods that are applicable in such problems are called nonparametric methods.

Order Statistics

- Sample: $x_1, x_2, \ldots, x_n$
- Ordered sample: $x_1^* < x_2^* < \ldots < x_n^*$
- Quantile statistics
  - 25% quantile
  - 25% quantile, also called Median, could be a better estimate of population mean in the case of non-normal distributions
  - 75% quantile
- Rank: position of x_k in order statistics, represented by r_k.

Nonparametric methods

- Sign test
- Rank test
- Permutation test
- Rank Correlation Coefficient, also called Spearman's Rank Correlation

Let’s work on previous examples together!

An example in ABO blood type

- Blood types in a number of people in an area. H_0: $p_i = p_i^0, i = 1, \cdots, 4$

Type	A	B	AB	O	Total
p_i^0	1/3	1/8	1/24	1/2	1
N_i	2162	738	228	2876	6004

- The four expected counts under H_0 are

$$np_1^0 = 6004 \times \frac{1}{3} = 2001.3, np_2^0 = 6004 \times \frac{1}{8} = 750.5$$

$$np_3^0 = 6004 \times \frac{1}{24} = 250.2, np_4^0 = 6004 \times \frac{1}{2} = 3002.0$$

An example in genetics: A resistance gene is linked with a molecular marker

F2 population	Resistant			Susceptible		
Marker type	A	H	B	A	H	B
Sample size	572	1161	14	3	22	569

Marker types A and B are parental types; H is the type of F1 hybrid

Resistant and susceptible can be fitted by the 3:1 ratio (one dominance gene locus): χ^2=0.17 (P=0.68). Marker types A, H, and B can be fitted by the 1:2:1 ratio (one co-dominance gene locus) : χ^2=0.32 (P=0.57)

But Resistance and Marker are not independent, i.e. can not be fitted by the 3:6:3:1:2:1 ratio:

An example in MN blood type

We want to test if the population in HW equilibrium: If (M, N)=(p, q), then (MM, MN, NN)=(p^2, $2pq$, q^2)

Genotype	MM	MN	NN	Total
Sample size	N_1=233	N_2=385	N_3=129	**N=747**
Predicted by Hardy-Weinberg Law	**242.70**	**366.18**	**138.38**	**747**

- (p, q) = (0.57, 0.43)
- (p^2, $2pq$, q^2) = (0.3249, 0.4902, 0.1849)
- χ^2=1.96 (df=1, P=0.16), i.e. the population is in HW equilibrium

An example in ABO blood type

when we don't have the expected frequencies

Type	A	B	AB	O	Total
N_i	2162	738	228	2876	6004

- H_0: this is a randomly mated population, i.e. the population is in HW equilibrium.
- Need to estimate p(A), p(B), p(O) first.
- By HWE, we mean:
  - $p(AA)=p(A)^2$, $p(AO)=2*p(A)*P(O)$
  - $p(BB)=p(A)^2$, $p(BO)=2*p(B)*P(O)$
  - $p(AB)=2*p(A)*P(B)$
  - $p(OO)=p(O)^2$

Independence test of contingency tables

Curriculum	Candidate preferred		
	A	B	Undecided
Engineering and science	24	23	12
Humanities and social sciences	24	14	10
Fine arts	17	8	13
Industrial and public administration	27	19	9