Analysis of Categorical Data

Many experiments result in measurements that are **qualitative** or **categorical** rather than *quantitative*:

- Employees can be classified into one of five income bracket.
- Mice might react in one of three wyas when subjected to a stimulus.
- Motor vehicles might fall into one of four vehicle types.
- Paintings could be classified into one of k categories according to style and period.
- The quality of surgical incisions could be most meaningfully be identified as excellent, very good, good, fair, or poor.
- Manufactured items are acceptable, seconds, or rejects.

Example 14.1 A group of rats, one by one, proceed down a ramp to one of three doors. Let p_i denote the probability that a rat will choose the ith door, for $i = 1, 2, 3$. We wish to test the hypothesis that the rates have no preference concerning the choice of a door. Thus, the appropriate null hypothesis is

$$H_0 : p_1 = p_2 = p_3 = \frac{1}{3}.$$

The alternative hypothesis is

$$H_1 : \text{the opposite of } H_0.$$

To conduct such test, we suppose that the rates were sent down the ramp $n = 90$ times and that the three observed cell frequencies were $n_1 = 23$, $n_2 = 36$, and $n_3 = 31$. Consider the significance level $\alpha = 0.05$.

Example 14.2 Historically, the proportions of all Caucasians in the United States with blood phenotypes A, B, AB, and O are .41, .10, .04, and .45, respectively. To determine whether current population proportions still match these historical values, a random sample of 200 American Caucasians were were selected, and their blood phenotypes were recorded. The observed numbers with each phenotype are given by $n_1 = 90$ (for blood phenotypes A), $n_2 = 18$ (B), $n_3 = 12$ (AB), and $n_4 = 81$ (O). Is there sufficient evidence, at the 0.05 level of significance, to claim that current proportions differ from the historic values?

Example 14.3 A city expressway with four lanes in each direction was studied to see whether drivers preferred to drive on the inside lanes. A total of 1000 automobiles were observed during the heavy early-morning traffic, and their respective lanes were recorded. The results are: $n_1 = 294$ used lane 1; $n_2 = 276$ used lane 2; $n_3 = 238$ used lane 3; $n_4 = 192$ used lane 4. Do the data present sufficient evidence to indicate that some lanes are preferred over others?

A goodness-of-fit test:

$$H_0 : p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0},$$

where $p_i > 0$ and $\sum_{i=1}^{k} p_i = 1$. The alternative hypothesis is

$$H_1 : \text{the opposite of } H_0.$$

A special case:

Example 14.1 (Continued)

Example 14.2 (Continued)

Example 14.3 (Continued)

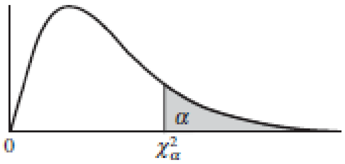
Example 14.4 The number of accidents Y per week at an intersection was checked for $n = 50$ weeks, with the results as shown in the Table. Test the hypothesis that the random variable Y has a Poisson distribution with $\lambda = 0.5$, assuming the observations to be independent. Use $\alpha = 0.05$.

y	Frequency
0	32
1	12
2	6
3 or more	0

Example 14.4 (continued) The number of accidents Y per week at an intersection was checked for $n = 50$ weeks, with the results as shown in the Table. Test the hypothesis that the random variable Y has a Poisson distribution, assuming the observations to be independent. Use $\alpha = 0.05$.

y	Frequency
0	32
1	12
2	6
3 or more	0

Table 6 Percentage Points of the χ^2 Distributions



df	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.950}$	$\chi^2_{0.900}$
1	0.0000393	0.0001571	0.0009821	0.0039321	0.0157908
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720
3	0.0717212	0.114832	0.215795	0.351846	0.584375
4	0.206990	0.297110	0.484419	0.710721	1.063623
5	0.411740	0.554300	0.831211	1.145476	1.61031
6	0.675727	0.872085	1.237347	1.63539	2.20413
7	0.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.16816
10	2.15585	2.55821	3.24697	3.94030	4.86518
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	3.07382	3.57056	4.40379	5.22603	6.30380
13	3.56503	4.10691	5.00874	5.89186	7.04150
14	4.07468	4.66043	5.62872	6.57063	7.78953
15	4.60094	5.22935	6.26214	7.26094	8.54675
16	5.14224	5.81221	6.90766	7.96164	9.31223
17	5.69724	6.40776	7.56418	8.67176	10.0852
18	6.26481	7.01491	8.23075	9.39046	10.8649
19	6.84398	7.63273	8.90655	10.1170	11.6509
20	7.43386	8.26040	9.59083	10.8508	12.4426
21	8.03366	8.89720	10.28293	11.5913	13.2396
22	8.64272	9.54249	10.9823	12.3380	14.0415
23	9.26042	10.19567	11.6885	13.0905	14.8479
24	9.88623	10.8564	12.4011	13.8484	15.6587
25	10.5197	11.5240	13.1197	14.6114	16.4734
26	11.1603	12.1981	13.8439	15.3791	17.2919
27	11.8076	12.8786	14.5733	16.1513	18.1138
28	12.4613	13.5648	15.3079	16.9279	18.9392
29	13.1211	14.2565	16.0471	17.7083	19.7677
30	13.7867	14.9535	16.7908	18.4926	20.5992
40	20.7065	22.1643	24.4331	26.5093	29.0505
50	27.9907	29.7067	32.3574	34.7642	37.6886
60	35.5346	37.4848	40.4817	43.1879	46.4589
70	43.2752	45.4418	48.7576	51.7393	55.3290
80	51.1720	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7541	65.6466	69.1260	73.2912
100	67.3276	70.0648	74.2219	77.9295	82.3581