

Final Review problems (solutions available): 6.9, 6.12, 6.20, 6.27, 6.29, 6.33, 6.73

Final Review problems (solutions not available yet): 10.18, 10.34, 10.35, 10.46

Chapter 10 Categorical Data

- In this chapter, we consider variables that are discrete or categorical (e.g., number of accidents per month at an intersection, or blood type (A+, B, O, etc.)
- We have already worked with the binomial distribution, one of the key distributions for the analysis of these data. Sections 10.2 and 10.3 discuss the two most important analyses

10.2 Inferences About a Population Proportion π

Case Study (from Ramsey, F.L. and Schafer, D.W., 1997. *The Statistical Sleuth, 2nd Ed.* p. 549). The Salk polio vaccine trials of 1954 included a double-blind experiment in which elementary school children were assigned at random to one of two groups: injection with the Salk polio vaccine, or injection with a placebo. (Parents consented to including their children in the study).

- The objective was to assess the vaccine as a cause of infantile paralysis

	Infantile Paralysis		
	Yes	No	Total
Placebo	142	199858	200000
Vaccine	56	199944	200000
	198	199802	400000

- Is this a randomized experiment or an observational study?
- What is the population to which inferences can be made?
- Can causation be assessed with these data?

A test is set up as follows

- Let π_1 denote the probability of infantile paralysis among children that are administered a placebo, and π_2 denote the probability of infantile paralysis among children that are administered the Salk vaccine

1) $H_0 : \pi_1 - \pi_2 = 0$ versus $H_a : \pi_1 - \pi_2 < 0$ is tested using

Group	$\hat{\pi}$	$\hat{\sigma}_{\hat{\pi}}$
Placebo	0.00071	0.000060
Vaccine	0.00028	0.000037

- Without carrying out a formal test, it is clear that there is no evidence against H_0 and in favor of H_a (because $\hat{\pi}_1 - \hat{\pi}_2 = 0.00043 > 0$)
- **Case Study** (from Ramsey, F.L. and Schafer, D.W., 1997. *The Statistical Sleuth*, p. 517). D.E. Crews. 1988. *Cardiovascular mortality in American Samoa*, *Human Biology*, **60**, 417-433.
- Obesity is known to be associated with increased risk of cardiovascular disease in western societies. Is this because of the strain of excessive weight or social stigma?
- Crews (1998) addressed this question by comparing the proportions of cardiovascular death among two samples of obese and non-obese American Samoan women. These data are summarized in tabular form in a contingency table:

	CVD Death		Total
	Yes	No	
Obese	16	2045	2061
Not Obese	7	1044	1051
	23	3089	3112

- The proportion of deaths attributable to CV (between 1976 and 1981) were

$$\hat{\pi}_o = \frac{16}{2061} = 0.00776$$

and

$$\hat{\pi}_n = \frac{7}{1051} = 0.00666.$$

- These estimates are based on relatively small numbers: 16 and 7. Can we trust them?
- Our level of trust corresponds to the precision of these values 0.00776 and 0.00666 as estimates of the population death rates. That is, Crews is really after the rate of CV death rates in the population of *all* obese and non-obese American Samoan women, not just the 3112 in this study

10.2 Inferences About a Population Proportion π

- Recall that a binomial experiment consists of n independent trials, each producing a dichotomous response (S or F), and with the same probability of success (denoted by π) on each trial
- We define (a binomial random variable) as

$$Y = \# \text{ of successes}$$

Recall,

$$P(y) = P(Y = y) = \begin{cases} \frac{n!}{y!(n-y)!} \pi^y (1-\pi)^{n-y}, & y = 0, 1, \dots, n, \\ 0, & \text{otherwise.} \end{cases}$$

- Example: roll a die 10 times, and let $Y = \#$ of 1's. Then, $\pi = 1/6$, and

$$P(0) = \frac{10!}{0!10!} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^{10} = \frac{10!}{1 \times 10!} \left(\frac{5}{6}\right)^{10} = 0.161.$$

- **Estimation of π .** Suppose that π is unknown, and the objective is to estimate π
- Observe n independent trials and Y successes. Then, the estimate of π is the *sample proportion of successes*

$$\hat{\pi} = \frac{Y}{n},$$

- $\hat{\pi}$ is a statistic, and hence it has a *sampling distribution*. The exact sampling distribution is not very useful, however, the Central Limit Theorem provides a useful approximate distribution:

$$\hat{\pi} \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right),$$

(because $\hat{\pi}$ is really a mean of n $\text{Bin}(1, \pi)$ observations)

- The approximation works only if both $n\pi > 5$ and $n(1-\pi) > 5$

- The *standard error* of $\hat{\pi}$ is

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

- An estimate of the standard error of $\hat{\pi}$ is

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

- **A Confidence Interval for π .** Recall that an approximate $100(1 - \alpha)$ confidence interval for μ when sampling from $N(\mu, \sigma)$ is

$$\bar{Y} \pm z_{\alpha/2} \hat{\sigma}_{\bar{y}}.$$

- Similarly, an approximate $100(1 - \alpha)$ confidence interval for π when sampling from $\text{Bin}(n, \pi)$ is

$$\hat{\pi} \pm z_{\alpha/2} \hat{\sigma}_{\hat{\pi}},$$

provided that both $n\hat{\pi} > 5$ and $n(1 - \hat{\pi}) > 5$.

- When $\hat{\pi} = 0$, the CI is $(0, 1 - [\alpha/2]^{1/n})$
- When $\hat{\pi} = 1$, the CI is $([\alpha/2]^{1/n}, 1)$
- **Sample size calculations.** For a $100(1 - \alpha)$ confidence interval, the minimum number of observations necessary to insure a half-width of E is

$$n = z_{\alpha/2}^2 \frac{\pi(1 - \pi)}{E^2}$$

(See Ott and Longnecker, p. 474)

- If π is unknown, then we will assume that $\pi = 0.5$. This is because 0.5 insures the largest possible sample size. E.g., for a 95% CI with half-width $E = 0.1$, we need

$$n = 1.96^2 \frac{0.5 \times 0.5}{0.1^2} \doteq 96$$

observations

Large Sample Hypothesis Testing

- Often, it is of interest to test whether π is equal to some value against the alternative that it is different. Specifically, we wish to test whether π is equal to a particular value, say π_0 .

- We consider three alternate hypotheses versus the null hypothesis

$$H_0 : \pi = \pi_0.$$

- The three versions of H_a are

1) $H_a : \pi \neq \pi_0$, or

2) $H_a : \pi > \pi_0$, or

3) $H_a : \pi < \pi_0$.

- The test is a large-sample test, that is, it should only be used when the sample size is large. Specifically, we require that $n\pi > 5$ and $n(1 - \pi) > 5$

- If those conditions are met, we use the test statistic

$$z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}}$$

where $\sigma_{\hat{\pi}} = \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}$.

- For a α -level test, the rejection regions are:

1. $R = \{z \mid |z| > z_{\alpha/2}\}$

2. $R = \{z \mid z > z_{\alpha}\}$

3. $R = \{z \mid z < -z_{\alpha}\}$,

where z_{α} is the upper α critical value from the $N(0,1)$ distribution. Hence, $\alpha = P(Z > z_{\alpha})$

Example Suppose that probability of death because of CVD in obese American women during the same period was known to be 0.01. Is there sufficient evidence to conclude that the CVD death rate is lower among American Samoan women?

- Set $\pi_0 = 0.01$, and define $\pi =$ probability that a randomly selected obese American Samoan dies of CVD during a five year interval.

- The test is set up by stating the hypotheses

$$H_0 : \pi = \pi_0 \text{ versus } H_a : \pi < \pi_0$$

- Next, we need to check that the large sample conditions are met:

$$n\pi = 2061 \times 0.01 = 20.6 > 5$$

and

$$n(1 - \pi) \gg 5$$

- The conditions are met, so the test statistic is

$$z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}} = \frac{0.00776 - 0.01}{0.00193} = -\frac{0.00224}{0.00193} = -1.15.$$

- Instead of a formal decision, I'll compute the p -value:

$$p\text{-value} = P(Z < -1.15) = 0.12$$

- These data provide marginal, or weak, statistical evidence that $\pi < 0.01$ during the period of interest

10.5 Inferences about the Difference Between Two Population Proportions $\pi_1 - \pi_2$

- Setting: Two independent samples have been randomly sampled from two possibly different populations and some binary attribute has been recorded (e.g., incidence of CVD death) on each sample unit (or individual)
- Generically, we measure a binary attribute on each sample unit. Generically, the attribute has two states, S (success) and F (failure)
- From population 1, we obtain n_1 observations. Of these, Y_1 are S 's. The estimate of the population proportion of S 's is

$$\hat{\pi}_1 = \frac{Y_1}{n_1}.$$

- From population 2, we obtain n_2 observations. Of these, Y_2 are S 's. The estimate of the population proportion of S 's is

$$\hat{\pi}_2 = \frac{Y_2}{n_2}.$$

- To compare π_1 and π_2 using the sample data, we use procedures that parallel the situation when analyzing two population means.

- Recall that for large samples, $\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$.

- For large samples,

$$\hat{\pi}_1 - \hat{\pi}_2 \sim N\left(\pi_1 - \pi_2, \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}\right).$$

The standard error of the difference $\hat{\pi}_1 - \hat{\pi}_2$ is estimated by the "plug-in" estimator

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}.$$

- This normal approximation can be used provided that $n_1\hat{\pi}_1 > 5$, $n_1(1 - \hat{\pi}_1) > 5$, $n_2\hat{\pi}_2 > 5$ and $n_2(1 - \hat{\pi}_2) > 5$.

- If so, then an approximate $100(1 - \alpha)$ confidence interval for $\pi_1 - \pi_2$ is

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2} \hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2}.$$

- For example, in the CVD study,

$$\hat{\pi}_o = \frac{16}{2061} = 0.00776, \hat{\sigma}_{\hat{\pi}_o} = 0.00193,$$

and

$$\hat{\pi}_n = \frac{7}{1051} = 0.00666, \hat{\sigma}_{\hat{\pi}_n} = 0.00251$$

yield

$$\begin{aligned}\hat{\sigma}_{\hat{\pi}_o - \hat{\pi}_n} &= \sqrt{\frac{\hat{\pi}_o(1 - \hat{\pi}_o)}{n_o} + \frac{\hat{\pi}_n(1 - \hat{\pi}_n)}{n_n}} \\ &= \sqrt{\frac{0.00776(1 - 0.00776)}{2061} + \frac{0.00666(1 - 0.00666)}{1051}} \\ &= 0.00317\end{aligned}$$

- An approximate 95% CI for $\pi_o - \pi_n$ is

$$0.00776 - 0.00666 \pm 1.96 \times 0.00317 = -0.0051 \text{ to } 0.0073.$$

- Because the interval contains 0, I conclude that there is little statistical evidence of a difference between π_1 and π_2 .

Large Sample Hypothesis Testing

- Adopting the usual setup, the hypotheses of interest are

$H_0 : \pi_1 - \pi_2 = 0$ versus

1) $H_a : \pi_1 - \pi_2 \neq 0$, or

2) $H_a : \pi_1 - \pi_2 > 0$, or

3) $H_a : \pi_1 - \pi_2 < 0$.

- Provided that $n_1\hat{\pi}_1 > 5$, $n_1(1 - \hat{\pi}_1) > 5$, $n_2\hat{\pi}_2 > 5$ and $n_2(1 - \hat{\pi}_2) > 5$, the test statistic is

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\hat{\sigma}_{\hat{\pi}}} \sim N(0,1)$$

- How should $\hat{\sigma}_{\hat{\pi}}$ be computed? In a test situation, we always assume H_0 to be true, and compute the necessary test statistics from that standpoint

- Here, H_0 implies that $\pi_1 - \pi_2 = 0 \Rightarrow \pi_1 = \pi_2 = \pi$. So, for efficiency, we combine samples. The number of S 's is $Y_1 + Y_2$ out of $n_1 + n_2$ observations.

- Using these sample statistics, $\hat{\pi}$ is estimated by

$$\hat{\pi} = \frac{Y_1 + Y_2}{n_1 + n_2},$$

- Then, the standard error (or deviation) of $\hat{\pi}$ is

$$\begin{aligned}\sigma_{\hat{\pi}} &= \sqrt{\frac{\pi(1-\pi)}{n_1} + \frac{\pi(1-\pi)}{n_2}} \\ &= \sqrt{\pi(1-\pi) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)},\end{aligned}$$

and the estimate of $\sigma_{\hat{\pi}}$ is

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\hat{\pi}(1-\hat{\pi}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

- Finally, the test statistic is

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\hat{\sigma}_{\hat{\pi}}} = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}},$$

where $\hat{\pi}_1 = \frac{Y_1}{n_1}$, $\hat{\pi}_2 = \frac{Y_2}{n_2}$, and

$$\hat{\pi} = \frac{Y_1 + Y_2}{n_1 + n_2}.$$

- For a α -level test, the rejection regions are:

1) $R = \{z \mid |z| > z_{\alpha/2}\}$

2) $R = \{z \mid z > z_{\alpha}\}$

3) $R = \{z \mid z < -z_{\alpha}\}.$

- Example: for the CVD study, $\hat{\pi}_1 = 0.00776$, $\hat{\pi}_2 = 0.00666$, and

$$\hat{\pi} = \frac{16 + 7}{2061 + 1051} = \frac{23}{3112} = 0.00739.$$

Also,

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\hat{\pi}(1 - \hat{\pi}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{0.00739(1 - 0.00739) \left(\frac{1}{2061} + \frac{1}{1051} \right)}$$

$$= 0.00325.$$

- The test of $H_0 : \pi_o - \pi_n = 0$ versus $H_0 : \pi_o - \pi_n > 0$ is given by

$$z = \frac{\hat{\pi}_o - \hat{\pi}_n}{\hat{\sigma}_{\hat{\pi}}} = \frac{0.00776 - 0.00666}{0.00325} = 0.34.$$

- The p -value is $P(\hat{\pi}_o - \hat{\pi}_n \geq 0.0011 \mid \pi_o = \pi_n)$, or

$$p\text{-value} = P(Z > 0.34) = 0.37$$

- My conclusion is that there is no evidence of differences in CVD rates between the two populations, obese and non-obese American Samoan women.
- Note that the relationship between obesity and CVD in American women is not directly addressed in this study.
- Warning: sometimes the importance of differences in proportions depends on the size of the proportions
 1. Suppose that the probability of contracting malaria are 0.5 and 0.48 when drugs A and B are taken, respectively. This difference is not much in a practical sense.
 2. Suppose instead that the probabilities of contracting malaria are 0.04 and 0.02 when drug A and B are taken, respectively. It may be argued that this difference is substantial in a practical sense, as the odds of contracting malaria are twice as great when taking A compared to B. (The odds are $0.04/0.02 = 2$).
- The analysis of odds, and odds ratios, is discussed at length in books on categorical data analysis. E.g.,
 1. Fienberg, S. 1987. The Analysis of Cross-Classified Data, 5th Ed. MIT Press.
 2. Agresti, A. 1990. Categorical Data Analysis. Wiley.

Case Study (from Ramsey, F.L. and Schafer, D.W., 1997. *The Statistical Sleuth, 2nd Ed.* p. 549). The Salk polio vaccine trials of 1954 included a double-blind experiment in which elementary school children were assigned at random to one of two groups: injection with the Salk polio vaccine, or injection with a placebo. (Parents consented to including their children in the study).

- The objective was to assess the vaccine as a cause of infantile paralysis

	Infantile Paralysis		Total
	Yes	No	
Placebo	142	199858	200000
Vaccine	56	199944	200000
	198	199802	400000

- This an experiment
- The population to which inferences can be made are elementary school children (with appropriate regional, age, etc limitations)
- Causation can be assessed with these data, because with randomized assignment of individuals to group, there can no reasonable alternate explanation of differences between groups besides the treatment

A test is set up as follows

- Let π_1 denote the probability of infantile paralysis among children that are administered a placebo, and π_2 denote the probability of infantile paralysis among children that are administered the Salk vaccine

1) $H_0 : \pi_1 - \pi_2 = 0$ versus $H_a : \pi_1 - \pi_2 < 0$ is tested using

Group	$\hat{\pi}$	$\hat{\sigma}_{\hat{\pi}}$
Placebo	0.00071	0.000060
Vaccine	0.00028	0.000037

- Because $\hat{\pi}_1 - \hat{\pi}_2 = 0.00043$, there is no evidence against H_0 and in favor of H_a .

- For completeness, though,
- 2) Using these sample statistics, $\hat{\pi}$ is estimated by

$$\hat{\pi} = \frac{142 + 56}{400000} = 0.000495,$$

- Then, the standard deviation of $\hat{\pi}$ is

$$\begin{aligned} \sigma_{\hat{\pi}} &= \sqrt{0.000495 \times 0.999505 \times \frac{2}{200000}} \\ &= 0.000141, \end{aligned}$$

- The test statistic is

$$z = \frac{0.00071 - 0.00028}{0.000141} = 3.57$$

and $p\text{-value} = P(Z < 0.3.57) \approx 0.99$.

3) What is the conclusion? How do we interpret such a large p -value?

10.3 Inferences About Several Proportions: Chi-Square Goodness-of-Fit Test

- Recall, the binomial random variable counts the number of outcomes k out of n that fall into one (S) of two classes. A key component is π , the probability that any single trial will produce an outcome in class S
- Sometimes, we have more than two classes and we are interested in the probabilities of one outcome falling in each class
- For example, if I randomly sample one individual among the general U.S. population, what is the probability that his or her blood type is A (or O)?
- There are 4 blood types, A, B, AB, O
- The probabilities of randomly selecting these types from the general U.S. population are $P(O) = 0.45$, $P(A) = 0.40$, $P(B) = 0.11$, $P(AB) = 0.04$
- Is this distribution the same among Blacks, Hispanics and Whites?
- To address questions of this type, we consider a generalization of the binomial distribution, the multinomial distribution

- A **multinomial experiment** consists of
 1. n independent trials
 2. each resulting in one of k possible outcomes (e.g., A, B, AB, O $\Rightarrow k = 4$)
 3. probabilities of each outcome are the same, $\pi_1, \pi_2, \dots, \pi_k$, on each trial, where $\pi_1 + \pi_2 + \dots + \pi_k = 1$.
- The multinomial random variable is a multivariate random variable that counts the number of outcomes in each of the k classes. The random variable is an ordered n -tuple denoted by

$$(n_1, n_2, \dots, n_k),$$

where $n = n_1 + n_2 + \dots + n_k$.

- Then, the probability of obtaining n_1 individuals belonging to class 1, and n_2 individuals belonging to class 2, \dots , n_k individuals belonging to class k , is

$$P(n_1, n_2, \dots, n_k) = \frac{n!}{n_1! n_2! \dots n_k!} \pi_1^{n_1} \pi_2^{n_2} \dots \pi_k^{n_k}$$

- If there are two classes ($k = 2$), then the multinomial random variable corresponds to a binomial random variable $X \sim \text{Bin}(n, \pi)$.

2. Outcomes are S and F

3. $\pi_1 = \pi$, the probability of a success, and $\pi_2 = 1 - \pi_1$,

- Then,

$$P(n_1, n_2,) = \frac{n!}{n_1! n_2!} \pi_1^{n_1} \pi_2^{n_2} = \frac{n!}{n_1! (n - n_1)!} \pi_1^{n_1} (1 - \pi_1)^{n - n_2}.$$

- Example: Blood type. Consider 3 donors. What is the probability that the 3 most common types are represented in a random sample of size 3 drawn from the general population? The outcome of interest is

$$(n_1, n_2, n_3, n_4) = (1, 1, 1, 0).$$

- The probability of this outcome is

$$\begin{aligned} P(1, 1, 1, 0) &= \frac{3!}{1!1!1!0!} 0.45^1 0.40^1 0.11^1 0.04^0 . \\ &= 6 \times 0.45 \times 0.40 \times 0.11 \times 1 \\ &= 0.116 \end{aligned}$$

- Usually, we are interested in testing whether a particular model, i.e., set of probabilities, is *incorrect* for some process or population
- We use the chi-square statistic to test the model fit
- Method:
 - 1) Assume the model is correct, and determine the expected number of observations in each cell, category, or class, given the total number of observations n .
 - 2) Compare the expected and observed numbers using the chi-square statistic
 - 3) If the expected and observed numbers are different, then there is evidence that the model is not correct
- Let $E_i = n\pi_i$ denote the expected cell count for cell i
- Let n_i denote the observed cell count for cell i
- Previous example: $E_1 = 3 \times 0.45 = 1.35$
- The chi-square (goodness-of-fit) statistic is

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - E_i)^2}{E_i}.$$

We use this statistic to test whether the data fit the model. If the fit is poor, then we have evidence against the model

- Example. Suppose from a sample of 100 Hispanics, we observe the following counts.

Type	n_i
A	43
B	30
O	22
AB	5

- Is there evidence that the distribution of blood types among Hispanics is different from the general population? The expected counts (E) are shown below

Type	n_i	E_i
O	43	45
A	30	40
B	22	11
AB	5	4

- $$\chi^2 = \frac{(43 - 45)^2}{45} + \frac{(30 - 40)^2}{40} + \frac{(22 - 11)^2}{11} + \frac{(5 - 4)^2}{4}$$

$$= 13.83$$

- The probability of obtaining such a large such a large χ^2 value is $P(\chi_3^2 \geq 13.83) = 0.0031$. Hence, there is strong evidence that the general population distribution is different that the Hispanic distribution
- Essentially all of the lack-of-fit is coming from two cells: B and O. Note that

$$\frac{(30 - 40)^2}{40} = 2.5 \text{ and } \frac{(22 - 11)^2}{11} = 11;$$

so 13.5 of the total χ^2 (13.83) is attributable to these two cells

Formal Test

- $H_0 : \pi_i = \pi_{i0}, i = 1, \dots, k$, versus $H_a : \text{at least one } \pi_i \text{ is not equal to } \pi_{i0}$.
- The H_0 model specifies the value of each π_{i0} .
- Under H_0 , the statistic

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - E_i)^2}{E_i}$$

is approximately χ_{k-1}^2 in distribution. We say that the degrees of freedom are $\text{df} = k - 1$.

- If the E_i 's are sufficiently large (all E_i 's > 1 , and no more than 20% are less than 5), then we can approximate the distribution of χ^2 by the χ_{k-1}^2 distribution. Otherwise, we must use exact methods
- Rejection region for a α -level test. Set $R = \left\{ \chi^2 \mid \chi^2 > \chi_{k-1, \alpha}^2 \right\}$ where $\chi_{k-1, \alpha}^2$ satisfies

$$P(\chi_{k-1}^2 > \chi_{k-1, \alpha}^2) = \alpha.$$

Example. Genetic theory calls for a 9 : 3 : 3 : 1 distribution of traits in classes A, B, C, and D, respectively. Observed counts from a sample of $n = 160$ individuals are

$$n_1 = 99, n_2 = 33, n_3 = 24, \text{ and } n_4 = 4,$$

respectively.

- Thus, $n = 99 + 33 + 24 + 4 = 160$.
- To determine statistical evidence against the theorized distribution, we compute

$$\pi_1 = \frac{9}{16}, \pi_2 = \frac{3}{16} = \pi_3, \pi_4 = \frac{1}{16}.$$

- The large sample conditions are easily satisfied because $n\pi_4 = 160 \times \frac{1}{16} = 10$. Then,

$$E_1 = n\pi_1 = 160 \times \frac{9}{16} = 90,$$

$$E_2 = n\pi_2 = 160 \times \frac{3}{16} = 30 = E_3,$$

and

$$E_4 = 10.$$

Then,

$$\chi^2 = \frac{(99 - 90)^2}{90} + \frac{(33 - 30)^2}{30} + \frac{(24 - 30)^2}{30} + \frac{(4 - 10)^2}{10} = 6.0.$$

• Under $H_0 : \pi_1 = \frac{9}{16}, \pi_2 = \frac{3}{16} = \pi_3, \pi_4 = \frac{1}{16}$, $\chi^2 \sim \chi_3^2$, and $P(\chi_3^2 > 6.0) \doteq 0.11$

• I conclude that there is some statistical evidence against the theorized distribution.

Remarks

- Care must be taken in the construction of the hypotheses. I tested evidence *against* the model stating that there is a 9 : 3 : 3 : 1 distribution of classes among the population.
- If I believe this model to be correct, and I am trying to establish evidence supporting this model, then I have made a logical error. Two arguments:
 1. The hypothesis testing framework sets H_a as the research hypothesis (the hypothesis that we believe to be true) and H_0 as a counter hypothesis. This is backwards if we want to confirm the model because H_0 states that the model is correct.
 2. Suppose that I want to show that H_0 is correct. Then, because I believe the model is correct, I do not want to reject H_0 . The best way to insure that outcome is to take a very small sample. That approach is inconsistent with the principles of scientific inquiry

Homework: For Feb. 12 (Wens.)

p. 501: 10.57 (Assigned for today)

p. 524, 10.78

(From Agresti, A. 1990. *Categorical Data Analysis*. Wiley, and a report from the Physician's Health Study Research Group at Harvard Medical School¹. A randomized clinical trial testing whether aspirin taken regularly reduces mortality from cardiovascular disease produced the data in the following table. Every day, physician's participating in the study took either one aspirin tablet or a placebo. The physicians did not know whether the table was aspirin or placebo.

1) Determine if myocardial infarction and aspirin use is independent.

- 2) Determine if there is evidence of an relationship between aspirin use and attack after combining the fatal and non-fatal classes.
- 3) Explain your conclusions in non-technical terms.

	Myocardial Infarction		
	Fatal Attack	Non-Fatal Attack	No Attack
Placebo	18	171	10,845
Aspirin	5	99	10,933

¹Preliminary Report: Findings from the Aspirin Component of the Ongoing Physician's Health Study. *N. England J. Med.* **318**:262-264, 1988.

10.5 The Poisson Distribution

- The Poisson distribution describes counts obtained from random processes in space and time.
- Ott and Longnecker (p. 497-494) provide other motivations and descriptions besides the following application
- The usual application of the Poisson distribution is as a model of the number of events occurring in a fixed length of time, or in a fixed area
- **Examples**
 - 1) the number of γ -particles that are detected, per second, by a Geiger counter held in a fixed position
 - 2) the number of microbes in 1 cc of fluid randomly drawn from a well-mixed sample of effluent
 - 3) the number of knapweed plants per m² at a random location on the West face of Mt. Sentinel
- Often, we are interested in determining if there are deviation from a randomness, spatially or temporally with respect to the number of events, or organisms
- We test whether a set of observed counts is consistent with a random distribution using the Poisson distribution
- In this case, the Poisson distribution is consistent with a random (spatial or temporal) distribution. That is, if the objects or events occur randomly in space (or time), and we count the number of objects or events in a set of units (or time intervals), the the distribution of counts will be Poisson

- Let Y denote a Poisson random variable (i.e., $Y \sim P(\mu)$).
- An very important property of Y is

$$E(Y) = \mu = \sigma^2 = \text{Var}(Y)$$

- The probability density function of Y is

$$P(Y = y) = \begin{cases} \frac{\mu^y e^{-\mu}}{y!}, & y \in \{0, 1, 2, \dots\} \\ 0, & \text{otherwise.} \end{cases}$$

- We use a chi-square test to determine the appropriateness of the Poisson model for a set of data. Specifically, we test

H_0 : the data are observations from a Poisson process, and

H_a : the data are not Poisson

- We use the chi-square goodness-of-fit test by setting up a set of categories corresponding to the count values $0, 1, 2, \dots, k$
- Category k is usually an open-ended; e.g., ≥ 4
- For example, a table of expected and observed counts is

	Count				
	0	1	2	3	≥ 4
n_i	3	4	5	1	3
E_i	2.17	4.20	4.06	2.61	1.97

- **Example** Are the number of accidents occurring at an intersection per month, random, or is there a pattern, such as more accidents in some months than others? Said another way, is it reasonable to believe that this count is Poisson; that is, do accidents occur randomly, without a pattern?

- The data are monthly counts for a year:
(0,3,2,2,0,1,4,1,1,2,0,7,2,1,2).

Step 1 - summarize the data

- Estimate μ by the average count, i.e., given counts y_1, \dots, y_n ,

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

- For these data, $n = 15$ and

$$\hat{\mu} = \frac{28}{15} = 1.93$$

- Compute the frequency of each observed count.

y_i	0	1	2	3	≥ 4
n_i	3	4	5	1	2

Step 2 - compute the *estimated expected* counts according to

$$\hat{E}_i = n\hat{P}(Y = i) = n \times \frac{\hat{\mu}^i e^{-\hat{\mu}}}{i!}$$

- For example,

$$\hat{E}_0 = nP(Y = 0) = 15 \times \frac{1.93^0 e^{-1.93}}{0!} = 15 \times 0.145 = 2.17.$$

- The table of estimated expected and observed counts is

	Count				
	0	1	2	3	≥ 4
n_i	3	4	5	1	3
\hat{E}_i	2.17	4.20	4.06	2.61	1.97

Step 3 - compute the approximate chi-square (goodness-of-fit) test statistic

$$\begin{aligned}\chi^2 &= \sum_{i=1}^k \frac{(n_i - \hat{E}_i)^2}{\hat{E}_i} \\ &= \frac{(3 - 2.17)^2}{2.17} + \frac{(4 - 4.2)^2}{4.2} + \dots + \frac{(2 - 1.97)^2}{1.97} \\ &= 1.54\end{aligned}$$

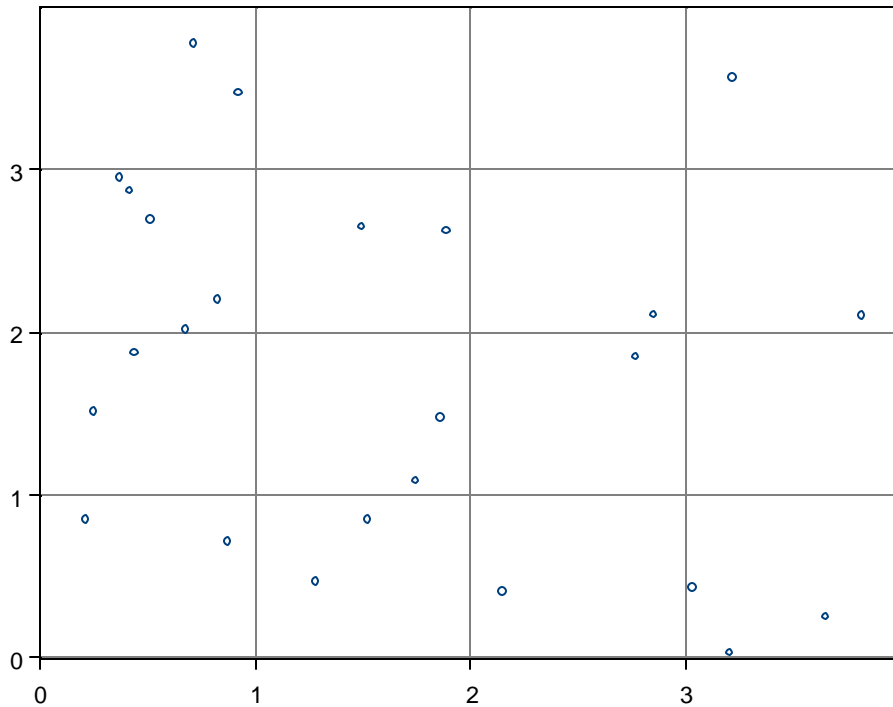
- Under H_0 , $\chi^2 \sim \chi_{df}^2$, where $df = \text{number of cells} - 2$. In this case, $df = 3$, and $P(\chi_3^2 > 1.54) \doteq 0.67$
- The degrees of freedom are the number of cells $- 2$, because there is a linear constraint in the table (the sum of the cell values is n), and we must estimate the Poisson mean μ to compute the \hat{E}_i 's. Consequently, the expected values are slightly more like the observed values than if we used a set of hypothesized values $\pi_{10}, \dots, \pi_{k0}$
- Warning: the approximation $\chi^2 \sim \chi_{df}^2$ is only good if all the expected cells counts are at least 1 and at least 80% of the 5 or larger
- In the example above, is the approximation $\chi^2 \sim \chi_{df}^2$ reliable? No.

Case Study The Spatial Distribution of Fire Ignitions on National Forest Lands in the Northern Region (J. Jones, USDA Forest Service; R. Redmond and J. Shumacher, MT Coop. Wildl. Res. St'n)

- An important first step was to determine if the spatial distribution of lightning induced fire ignitions is nonrandom. If they are nonrandom, then there interest in attempting to analyze factors affecting the spatial distribution
- The analysis was carried out for each decade: 1960's, 1970's, 1980's, 1990's, and for each type of fire ignition, those caused by lightning versus those cause by people.
- They counted the number of fires that were observed in each cell of a 5 km lattice, or grid, superimposed over National Forest lands in the Northern Region
- If the spatial distribution of fire ignitions is random, then statistical theory says that this distribution is Poisson. Conversely, if the distribution is not Poisson, then the distribution is not random

- These facts give the standard test for random dispersal. Specifically, they tested whether the number of fires are Poisson in distribution. If this hypothesis were rejected, then they would conclude that the spatial distribution of fires is not random

Figure 1. 25 points. Random or not?



- They used only those count values for which the model of randomness predicts an expected cell count of least 5. All other cells were ignored. This procedure is conservative and simpler than the other usual approach of combining, as one cell, cells with cells counts of less than 5
- The table below list number of fires (Count), the observed number of cells with that number of fires (Obs), the expected number of cells with that number of fires (Expected), and the contribution of the cell to the chi-square statistic (χ^2 contribution)

Table 1. Numbers of lightning-caused fires between 1960 and 1970. $\bar{Y} = 1.79$; $\hat{\sigma}^2 = 5.88$.

Count	Obs	Expected	χ^2 contribution
0	1309	570.7	955.3
1	825	1024.2	38.7
2	422	919.0	268.8
3	285	549.8	127.5
4	175	246.7	20.8
5	138	88.5	27.6
6	97	26.5	187.7
7	62	6.8	448.9

- The table shows that there is a high degree of organization in the spatial distribution of ignitions. There are far more cells with zeros than expected, and far more cells with many ignitions than expected
- It is not really necessary, but for completeness, a goodness-of-fit test assessing evidence against the null hypothesis of randomness is $\chi^2 = 2075.5$ (df = 6), and the p -value is less than 0.0001

10.6 Contingency Tables: Test of Independence

Setting:

- The data consist of n measurements on *two* categorical variables; e.g., HIV status (positive,negative) and behavioral risk category (high,low)
- Note that both variables are categorical. There is no obvious ordering of the levels of either variable
- Our interest lies in analyzing the association between the two variables: for example, is HIV status and behavior related? Said another way, is it possible to effectively predict one variable (e.g., HIV status) from the other?

Contingency Tables

- A device for summarizing the pattern of association between the two variables
- Let r denote the number of categories for the first variable, and c denote the number of categories for the second variable.
- The contingency table consists of r rows and c columns. In row i and column j , we place the number of observations recorded with level i of the first (row) variable and with level j of the second (column) variable

Example: Daly, M. and Wilson, M. 1988. "Evolutionary Social Psychology and Family Homicide", *Science*, **242**, 519-24.

- The Table below lists the number of homicides of children committed by their parents. These data can be used to investigate whether there is a association between the parent and child genders, and the age classification of the child.

Gender of parent/child	Age Classification of the Child				
	Infantile (0-1)	Oedipal (2-5)	Latency (6-10)	Circumpubertal (11-16)	Adult (≥ 17)
Male/Male	24	21	21	29	104
Male/Female	17	27	10	14	47
Female/Male	53	21	19	9	8
Female/Female	50	27	5	4	15

- There are a variety of models that may be considered for *cross-classified* data. We will consider the simplest and most-widely used: independence.
- The alternative to independence is dependence. If the variables are dependent, then there is an association between the two, and some classes of the row variable are more likely to occur with particular classes of the column variable
- For example, a cursory look at the table suggests that when male parents kill their children, it is more likely that the victims are older than the victims of female murderers. In other words, there appears to be an association between gender and age class

- We identify the counts, and probabilities using a row subscript (i), and a column subscript (j):

$$\begin{array}{cccc}
 n_{11} & n_{12} & \cdots & n_{1c} \\
 n_{21} & n_{22} & \cdots & n_{2c} \\
 \vdots & \vdots & & \vdots \\
 n_{r1} & n_{r2} & \cdots & n_{rc}
 \end{array}$$

- A generic entry is $n_{\text{row,column}} = n_{ij}$.
- A specific entry is $n_{12} = 21$, the number of murders of Oedipal male children by male parents

- The total number of observations is $n = \sum_{i=1}^r \sum_{j=1}^c n_{ij}$

- A test of

$$H_0 : \text{row and column variables are independent,}$$

versus

$$H_a : \text{row and column variables are not independent}$$

is obtained from the χ^2 goodness-of-fit test

- Specifically, independence implies a model. The underlying model is the multinomial distribution for the number of observations in each cell of the table. Recall, the multinomial random variable was denoted by

$$(n_1, n_2, \dots, n_k),$$

where $n = n_1 + n_2 + \dots + n_k$. Now, we can think of them as

$$(n_{11}, n_{12}, \dots, n_{rc})$$

or in table form as

$$\begin{array}{cccc} n_{11} & n_{12} & \cdots & n_{1c} \\ n_{21} & n_{22} & \cdots & n_{2c} \\ \vdots & \vdots & & \vdots \\ n_{r1} & n_{r2} & \cdots & n_{rc} \end{array}$$

- Similarly, the multinomial probabilities are denoted by

$$(\pi_{11}, \pi_{12}, \dots, \pi_{rc}),$$

though it may more useful to think of them in table form

$$\begin{array}{cccc} \pi_{11} & \pi_{12} & \cdots & \pi_{1c} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2c} \\ \vdots & \vdots & & \vdots \\ \pi_{r1} & \pi_{r2} & \cdots & \pi_{rc} \end{array}$$

- To develop the test, recall that events A and B are independent if and only if

$$P(A \cap B) = P(A)P(B).$$

- If the row and column variables are independent, then the probability that a single murder will be classified as category i of the row variable (event A) and category j of the column variable (event B) is

$$\pi_{ij} = \pi_i \pi_j,$$

where

$\pi_i = P(\text{an outcome belongs to class } i \text{ of the row variable}) = P(A)$, and

$\pi_j = P(\text{an outcome belongs to class } j \text{ of the column variable}) = P(B)$

$\pi_{ij} = P(\text{an outcome belongs to class } j \text{ of the column variable and class } i \text{ of the row variable}) = P(A \cap B)$

- If the row and column variables are *not* independent, then at least one cell product is incorrect, i.e., there is some i and j such that

$$\pi_{ij} \neq \pi_i \pi_j$$

- Therefore, the test of association is actually a test of independence given by

- $H_0 : \pi_{ij} = \pi_i \pi_j$, for all $i = 1, \dots, r$, and $j = 1, \dots, c$,

versus

$H_a : \pi_{ij} \neq \pi_i \pi_j$ for at least one π_{ij} .

- Under H_0 , the statistic

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - E_{ij})^2}{E_{ij}},$$

where E_{ij} 's are the expected cell counts under the independence model

- If H_0 is true, then χ^2 is approximately χ_{df}^2 in distribution, where

$$df = (r - 1)(c - 1),$$

provided that all E_{ij} 's are at least 1, and no more than 20% are less than 5

- The rejection region for a α -level test are all values of χ^2 which are larger than the upper-tail $1 - \alpha$ percentile of the χ_{df}^2 distribution.
- Let $\chi_{df,\alpha}^2$ denote the upper-tail $1 - \alpha$ percentile.
- Then, the rule is: reject H_0 if $\chi^2 > \chi_{df,\alpha}^2$. A p -value for the test is the probability of obtaining a value of χ^2 at least as large as was observed

Calculation of E_{ij} 's

- Denote the row and column totals by n_{i+} and n_{+j}
- For example,

$$n_{1+} = 24 + 21 + 21 + 29 + 104 = 199$$

and

$$n_{+1} = 24 + 17 + 53 + 50 = 144$$

- Also, $n = 525$
- Estimate π_i by

$$\hat{\pi}_i = n_{i+}/n, \text{ and } \hat{\pi}_j = n_{+j}/n;$$

then, under H_0 (independence)

$$\hat{\pi}_{ij} = \frac{n_{i+}}{n} \frac{n_{+j}}{n}.$$

- Then, $\hat{E}_{ij} = n\hat{\pi}_{ij}$ can be slightly simplified:

$$\hat{E}_{ij} = n \times \frac{n_{i+}}{n} \frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}.$$

- For example,

$$\hat{\pi}_{11} = \frac{n_{1+}n_{+1}}{n} = 199 \times \frac{144}{525} = 54.6$$

- Using this formula gives the table of expected counts:

Gender of parent/child	Age Classification of the Child				
	Infantile (0-1)	Oedipal (2-5)	Latency (6-10)	Circumpubertal (11-16)	Adult (≥ 17)
Male/Male	54.6	36.4	20.8	21.2	66.0
Male/Female	31.5	21.0	12.0	12.3	38.1
Female/Male	30.2	20.1	11.5	11.7	36.5
Female/Female	27.7	18.5	10.6	10.8	33.5

- The expected counts are sufficiently large, and the χ^2 statistic is $\chi^2 = 143.8$, based on $df = (4 - 1) \times (5 - 1) = 12$. Because $P(\chi^2_{12} > 143.8) < 0.0001$, there is very strong evidence that the independence model does not hold, and that there is an association between gender and age class

Overview of Categorical Data Methods

- Given a multinomial experiment, the multinomial random variable counts the number of outcomes in each of $k \geq 2$ categories.
- A binomial experiment is a multinomial experiment in which there are only two categories (S and F).
- The probability of observing one count in category j is π_j .
- Our primary interest is in testing

$$H_0 : \pi_j = \pi_{j0}, \text{ for each } j = 1, \dots, k$$

versus

$$H_a : \pi_j \neq \pi_{j0}, \text{ for at least one } j.$$

- π_{j0} is specified for each $j = 1, \dots, k$
- The test statistic is the goodness-of-fit chi-square (χ^2_{df}) statistic. The df depends on the application

- **Applications**

- 1) Multinomial data (e.g., genetic traits) - We attempt to disprove a model for the π_i 's. H_0 specifies that the data obey the model
- 2) Contingency tables, with rc categories. We test for association between two variables. H_0 specifies independence, that is, $\pi_{ij} = \pi_i\pi_j$ for all i and j
- 3) Poisson models. We test whether the data follow a Poisson distribution by forming k categories. From the Poisson distribution, and the data, we compute estimates $\hat{\pi}_i$