1 An Introduction to the Analysis of Categorical Data

1.1 What is meant by "categorical data"?

A *categorical* variable has a measurement scale that consists of a set of categories.

Examples:

- Attitude toward gun control: favor or oppose
- Gender: male or female
- Education: Did not graduate from high school, high school, some college, bachelor's degree, master's degree, Ph. D., professional degree
- Cholesterol level (high, moderate, low)
- Heart disease (yes, no)

1.2 Discrete (Count) Data

Often data consist of the counts of the number of events of certain types. We will study methods of analysis specifically designed for count data.

Examples:

- Number of deaths from tornadoes in a state during a given year: 0, 1, 2, 3, ...
- The number of automobile accidents on Texas Avenue during a week

1.3 Response and Explanatory Variables

The *response* variable is often called the *dependent* variable.

The *explanatory* variables are often called *predictors* or *independent* variables.

Examples: Distinguish between response and explanatory variables.

- Attitude toward gun control, Gender, Education
- Cholesterol level (high, moderate, low), Heart disease (yes, no)
- Number of deaths from tornadoes, number of killer tornadoes

1.4 Nominal and Ordinal Data

Categorical variables with ordered scales are *ordinal* variables. Otherwise they are *nominal*.

Examples: Identify whether each variable is nominal or ordinal.

- Education
- Political party affiliation (Democrat, Republican, Other)
- Cholesterol level (high, moderate, low)
- Hospital location (College Station, Bryan, Temple, Houston)

Methods designed for nominal variables:

- can be used for both nominal and ordinal variables.
- do not use the ordering for ordinal variables. This can cause a loss of power.
- give the same results no matter in which order the categories are listed.

Methods designed for ordinal variables:

- cannot be used with nominal variables.
- make use of the category ordering.
- would give different results if the categories were differently ordered.

1.5 Examples of Categorical Data Analyses

- In a sample of 50 adult Americans, only 14 correctly described the Bill of Rights as the first ten amendments to the U. S. Constitution. Estimate the proportion of Americans that can give a correct description of the Bill of Rights.
- In 1954, 401,974 children were participants in a large clinical trial for polio vaccine. Of these 201,229 were given the vaccine, and 200,745 were given a placebo. 110 of the children who received a placebo got polio and 33 of those given the vaccine got polio. Was the vaccine effective?
- The Storm Prediction Center (an agency of NOAA) tracks the number and characteristics of tornadoes. Construct a model relating the number of deaths to the number of tornadoes or to the number of killer tornadoes.
- A study was carried out to compare two treatments for a respiratory disorder. The goal was to compare the proportions of patients responding favorably to test and placebo. A confounding factor is that the study was carried out at two centers which had different patient populations. We wish to examine the association between treatment and response while adjusting for the effects of the centers.

- Stillbirth is the death of a fetus at any time after the twentieth week of pregnancy. A premature birth is the live birth of a child from the twentieth until the thirty-seventh week of pregnancy. Fit loglinear models to investigate the association among the following variables that were recorded in a study of stillbirth in the Australian state of Queensland:
  - Birth status(B) – stillbirth or live birth
  - Gender(G) – male or female
  - Gestational age(A) – $\leq 24, 25-28, 29-32, 33-36, 37-41$ weeks
  - Race(R) – Aborigine or white
- A female horseshoe crab with a male crab attached to her nest sometimes has other male crabs, called satellites, near her. Fit a logistic regression model to estimate the probability that a female crab has at least one satellite as a function of color, weight, and carapace width.
- 59 alligators were sampled in Florida. The response is primary food type: Fish, Invertebrate, and Other. The explanatory variable is length of the alligator in meters.

1.6 Some References

- Alan Agresti, *Analysis of Ordinal Categorical Data*
- Alan Agresti, *Categorical Data Analysis, Third Edition*
- Paul D. Allison, *Logistic Regression Using the SAS System*
- Christopher Bilder and Thomas Loughin, *Analysis of Categorical Data with R*
- Ronald Christensen, *Log-Linear Models and Logistic Regression, Second Edition*
- D. Collett, *Modelling Binary Data, Second Edition*
- P. Congdon, *Bayesian Models for Categorical Data*
- D. R. Cox and E. J. Snell, *Analysis of Binary Data, Second Edition*
- Julian Faraway, *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*
- Frank E. Harrell, Jr., *Regression Modelling Strategies*
- D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression, Second Edition*
- Bayo Lawal, *Categorical Data Analysis with SAS and SPSS Applications*
- P. McCullagh and J. A. Nelder, *Generalized Linear Models, Second Edition*
- Jeffrey S. Simonoff, *Analyzing Categorical Data*
- Maura E. Stokes, Charles S. Davis, and Gary G. Koch, *Categorical Data Analysis Using the SAS System, Third Edition*
- Gerhard Tutz, *Regression for Categorical Data*
- Daniel Zelterman, *Advanced Log-Linear Models Using SAS*

2 Models for Categorical Data

In this section we will study the basic distributions used in the analysis of categorical data:

- Binomial distribution
- Multinomial distribution
- Poisson distribution
- Models for overdispersed data

2.1 Binomial Probability Distribution

A binomial experiment satisfies the conditions:

1. Experiment consists of n trials.
2. Each trial can result in one of two outcomes (success, S, or failure, F).
3. Trials are independent.
4. The probability of S is the same for each trial: $P(S) = \pi$.

Define the random variable Y to be the number of successes in the binomial experiment. We call Y a **binomial random variable**. We write:

$$Y \sim \text{Binomial}(n, \pi)$$

n = fixed number of trials

π = probability of success

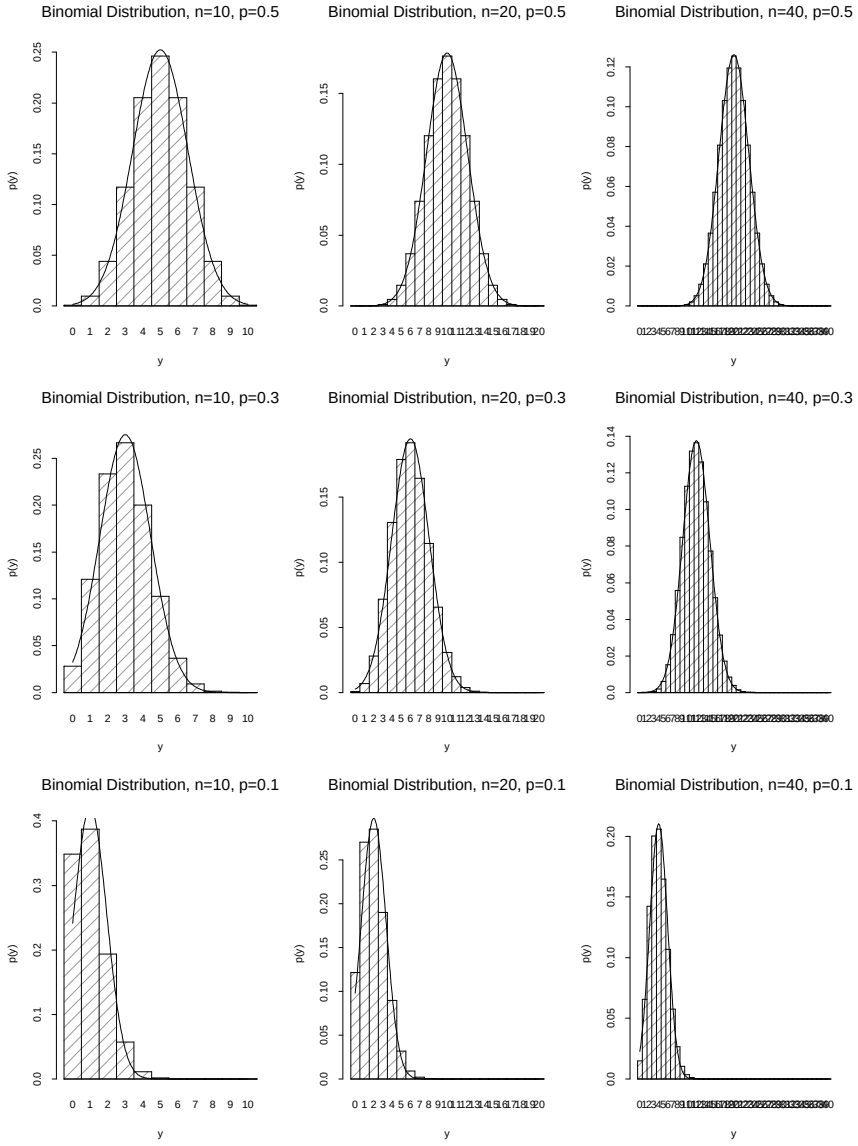
$$P(y) = Pr(Y = y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n$$

where

$$\binom{n}{y} = \frac{n!}{y!\,(n-y)!}$$

$$\begin{aligned} \mu &= E(Y) &&= n\pi \\ \sigma^2 &= \text{Var}(Y) &&= n\pi(1-\pi) \end{aligned}$$

Some Binomial Distributions



2.2 Multinomial Probability Distribution

A **multinomial experiment** satisfies the following conditions:

1. Experiment consists of n trials.
2. Each trial can result in one of c mutually exclusive categories.
3. Trials are independent.
4. The probability of the i^{th} category is the same for each trial: π_i, where $\pi_1 + \cdots + \pi_c = 1$.

The observable random variables are $(N_1, N_2, \ldots, N_c)$ where N_i = number of trials resulting in the i^{th} category, where $N_1 + \cdots + N_c = n$. The random variables $\boldsymbol{N} = (N_1, \ldots, N_c)$ are said to have a multinomial distribution, Mult$(n, \boldsymbol{\pi})$ with probability function

$$P(\boldsymbol{n}) = P[N_1 = n_1, \ldots, N_c = n_c] = \frac{n!}{n_1! n_2! \cdots n_c!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}.$$

Expected Value: $E(N_i) = n\pi_i = E_i$

Variance: $\mathrm{Var}(N_i) = n\pi_i(1 - \pi_i)$

Correlation: $\mathrm{Corr}(N_i, N_j) = -\sqrt{\frac{\pi_i \pi_j}{(1-\pi_i)(1-\pi_j)}}$

2.3 Poisson Distribution

Consider these random variables:

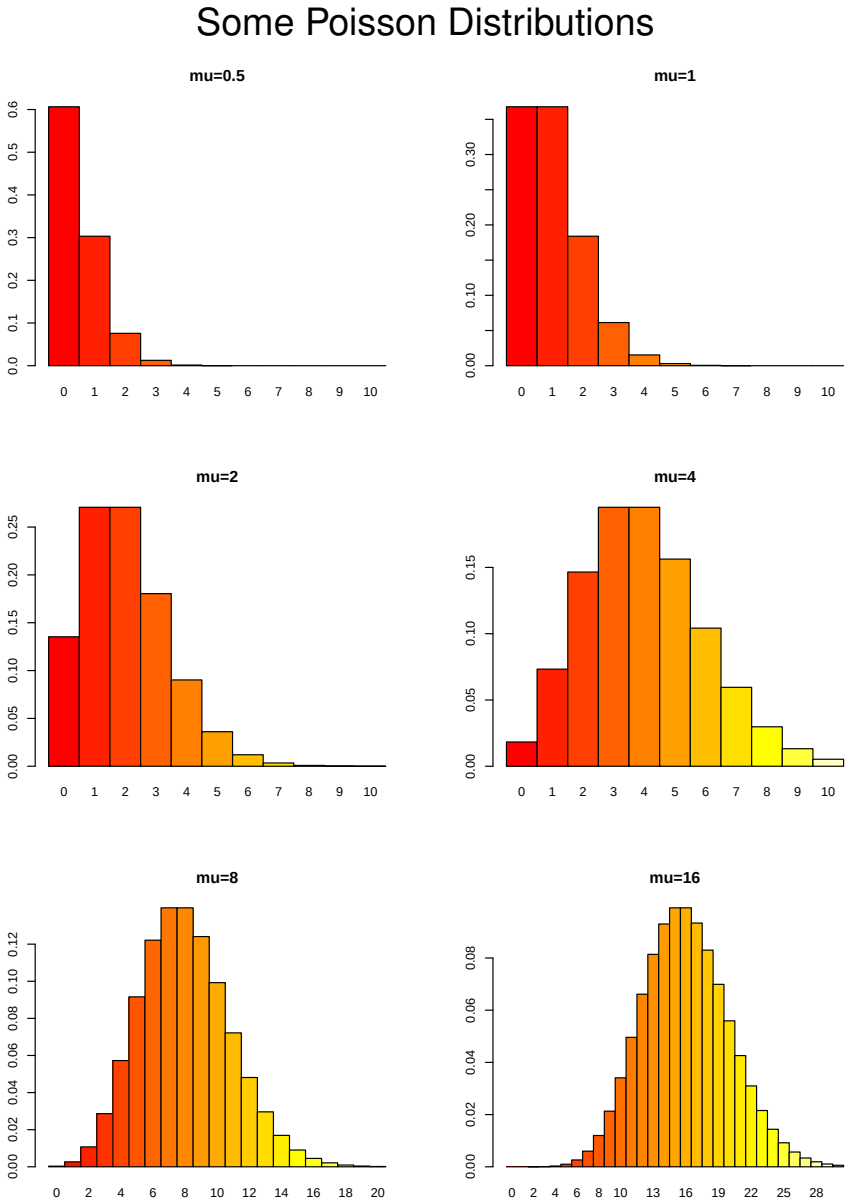
- Number of telephone calls received per hour.
- Number of days school is closed due to snow.
- Number of baseball games postponed due to rain during a season.
- Number of trees in an area of forest.
- Number of bacteria in a culture.

A random variable Y, the number of successes occurring during a given time interval or in a specified region (of length or size T), is called a *Poisson* random variable. The corresponding distribution:

$$Y \sim \text{Poisson}(\mu)$$

where $\mu = \lambda T$ and λ is the rate per unit time or rate per unit area.

$$\begin{aligned} P(y) &= Pr(Y = y) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \ldots, \quad \mu > 0 \\ E[Y] &= \mu \\ \text{Var}[Y] &= \mu \end{aligned}$$



2.4 Models for Overdispersed Data

Both the binomial distribution and the Poisson distribution have one parameter. This implies that the variance must have a particular mathematical relationship to the mean. When the data exhibit a sample mean and a sample variance that do not reflect this mathematical relationship, we should consider alternative models for the data.

For Poisson data, the population mean and population variance are equal. If the data exhibit a larger variance than the mean, the data are said to be *overdispersed*. This occurs when there are different Poisson models in the population resulting in a larger observed variance than expected. We will consider a couple of models that result in overdispersion.

2.4.1 The Zero-Inflated Poisson Model

Some data exhibit a greater than expected variance because there are too many observations taking the value of zero. We can model this as a mixture of two populations: a Poisson population that is selected with probability π and a population of zeros that is selected with probability $1 - \pi$.

The probability function for the zero-inflated Poisson r.v. is

$$P(x;\mu,\pi) = \begin{cases} 1-\pi+\pi e^{-\mu} & \text{if } x = 0 \\ \pi e^{-\mu}\mu^x/x! & \text{if } x > 0 \end{cases}$$

An easy calculation provides the mean and variance:

$$\begin{aligned} E(X) &= \pi\mu, \\ \mathrm{Var}(X) &= \pi\mu[1+\mu(1-\pi)]. \end{aligned}$$

We can compare the mean to the variance using an *index of dispersion*:

$$\frac{\mathrm{Var}(X)}{E(X)} = 1+\mu(1-\pi) > 1 \quad \text{if} \quad \pi \neq 0.$$

2.4.2 The Negative Binomial Distribution

Overdispersion can result from a different type of heterogeneity. Suppose that each observation X has a Poisson distribution with the mean of each observation being a random variable with a gamma distribution. The resulting r.v. has a *negative binomial distribution* with probability function

$$P(x;\mu,\nu) = \frac{\Gamma(x+\nu)}{x!\,\Gamma(\nu)}\left(\frac{\nu}{\nu+\mu}\right)^{\nu}\left(\frac{\mu}{\nu+\mu}\right)^{x},\ x=0,1,2,\ldots$$

where the gamma function is defined by

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}e^{-t}dt.$$

The mean and variance are:

$$\begin{aligned} E(X) &= \mu \\ \mathrm{Var}(X) &= \mu[1+D\mu] \quad \text{where} \quad D = 1/\nu \end{aligned}$$

We can consider D to be an index (or measure) of overdispersion.

3 Statistical Inference for Categorical Data

We will now suppose that we have observed data from one of the previous distributions where the value of the parameter is unknown. We will develop methods for statistical inference concerning the unknown parameters.

- We use the technique of maximum likelihood for estimation of parameters.
- We use likelihood-based tests for hypothesis testing. These tests include likelihood-ratio tests, Wald tests, and score tests.
- We use the corresponding confidence intervals for interval estimation. We will construction likelihood-ratio intervals, Wald intervals, and score intervals.

We now apply some of these techiques to the binomial, Poisson, and multinomial distributions. In later chapters, we will consider response data with one on the above distributions in the situation where we have one or more explanatory variables.

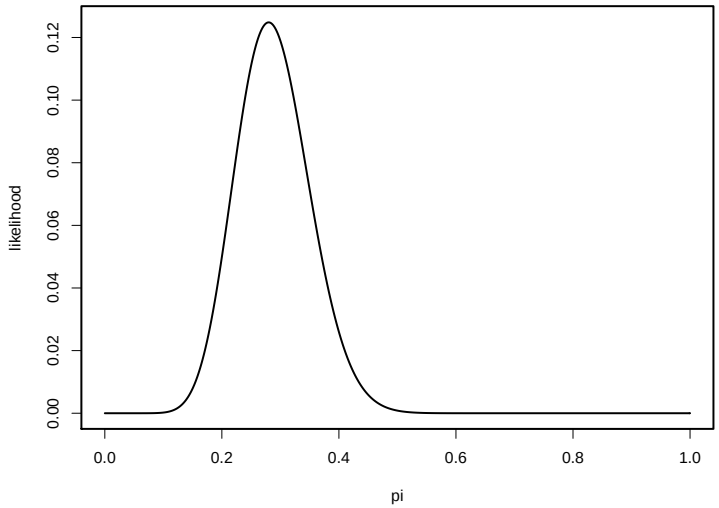
3.1 Inference for a Binomial Proportion

Suppose we observe $Y \sim$ Binomial(n, π). Our goal is to estimate π.

Example: In a sample of 50 adult Americans, only 14 correctly described the Bill of Rights as the first ten amendments to the U. S. Constitution. Estimate the proportion of Americans that can give a correct description of the Bill of Rights.

A common method of estimation is known as *maximum likelihood estimation*. We use the formula for the binomial distribution to write out the probability of 14 successes out of 50 trials.

$$P(14) = Pr(Y = 14) = \binom{50}{14} \pi^{14} (1 - \pi)^{50-14}.$$



The probability function $P(y)$ viewed as a function of the parameter π is called the likelihood function, $\ell(\pi)$.

The value of π that maximizes $\ell(\pi)$ is the maximum likelihood estimate (m.l.e.), $\hat{\pi} = \frac{14}{50} = 0.28$. We can verify that this value maximizes the likelihood using calculus.

We take the derivative of the natural logarithm of the likelihood, $L(\pi) = \log(\ell(\pi))$:

$$\frac{\partial L(\pi)}{\partial \pi} = \frac{\partial}{\partial \pi}\left[\log\binom{50}{14} + 14\log(\pi) + 36\log(1-\pi)\right] = \frac{14}{\pi} - \frac{36}{1-\pi}$$

We set this equal to zero and solve for π:

$$\frac{14}{\pi} - \frac{36}{1-\pi} = 0 \quad \Longrightarrow \quad \frac{14}{\pi} = \frac{36}{1-\pi} \quad \Longrightarrow \quad 14(1-\pi) = 36\pi \quad \Longrightarrow \quad \hat{\pi} = p = \frac{14}{50}$$

M.L.E. for the Population Proportion

A similar argument for an $Y \sim\sim$ Binomial(n, π) implies that the mle of π is the sample proportion, p:

$$\hat{\pi} = p = \frac{Y}{n}.$$

Properties of the M.L.E.

1. $E(\hat{\pi}) = \pi$
2. $\text{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$
3. The distribution of $\hat{\pi}$ is approximately normal with the above mean and variance.

Fact: If $Y \sim$ Binomial(n, π) and the probability histogram is not too skewed ($n\pi \geq 5$, $n(1-\pi) \geq 5$), then Y is approximately normally distributed. This implies that the m.l.e. $\hat{\pi} = \frac{Y}{n}$ is approximately normally distributed.

3.1.1 Tests Concerning a Population Proportion

We use large sample methodology to construct tests for a population proportion. Therefore, when $n\pi \geq 5$ and $n(1-\pi) \geq 5$, we can standardize the m.l.e. $\hat{\pi}$ to obtain an approximately standard normal rv:

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

Null hypothesis: $H_0 : \pi = \pi_0$

The score statistic uses the distribution under H_0 to obtain the standard error in the test statistic:

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

The Wald statistic uses the estimated standard deviation of $\hat{\pi}$ (also known as the standard error) to standardize $\hat{\pi}$:

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}$$

Alternative hypothesis	Rejection region for level α test	$P-$value
$H_A : \pi > \pi_0$	$Z \geq Z_\alpha$	$P[Z \geq Z_{obs}]$
$H_A : \pi < \pi_0$	$Z \leq -Z_\alpha$	$P[Z \leq Z_{obs}]$
$H_A : \pi \neq \pi_0$	$Z \geq Z_{\alpha/2}$ or $Z \leq -Z_{\alpha/2}$	$2P[Z \geq \lvert Z_{obs} \rvert]$

3.1.2 A Large-Sample Interval for π (Population Proportion)

To construct a confidence interval for the population proportion π based on the m.l.e, $\hat{\pi}$, we write

$$P\left(-Z_{\alpha/2} < \frac{\hat{\pi} - \pi}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} < Z_{\alpha/2}\right) \approx 1 - \alpha$$

Solving for our parameter π yields (approximately):

$$P\big(\hat{\pi} - Z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n} < \pi < \hat{\pi} + Z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}\big) \approx 1 - \alpha$$

Thus, a 95% confidence interval for π is

$$\hat{\pi} \pm 1.96\, SE, \quad \text{where the estimated standard error } SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

Remark: The above confidence interval is known as the *Wald interval* since it is based on the Wald statistic for testing $H_0 : \pi = \pi_0$:

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

The Wilson interval is an approximate confidence interval based on the corresponding score statistic:

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

We use the approximate probability statement

$$P\left(-Z_{\alpha/2} < \frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} < Z_{\alpha/2}\right) \approx 1 - \alpha$$

to obtain the inequality

$$\frac{(\hat{\pi} - \pi)^2}{\pi(1-\pi)/n} < Z^2_{\alpha/2}$$

which results in a quadratic equation for π. The solutions to the quadratic equation provide the endpoints for the confidence interval. The resulting confidence interval has endpoints

$$\frac{\hat{\pi} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + (z^2_{\alpha/2})/n}.$$

The score interval has been shown to have better coverage properties than the Wald interval.

Remark: Agresti and Coull derived a simple approximation to the level α score interval. Form the Wald interval based on $n^* = n + Z^2_{\alpha/2}$ observations and $\tilde{\pi} = (Y + Z^2_{\alpha/2}/2)/(n + Z^2_{\alpha/2})$. This is equivalent to adding $Z^2_{\alpha/2}$ observations, half of which are successes and half are failures.

Example: In a survey of 277 randomly selected adult shoppers, 69 stated that if an advertised item is unavailable they request a rain check, construct a 95% CI for π.

$$\begin{aligned} \hat{\pi} &\pm Z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n} \\ .249 &\pm 1.96\sqrt{.249(1-.249)/277} \\ .249 &\pm .051 \end{aligned}$$

The 95% Wald interval is $(.198, .300)$.

We can also compute the more accurate confidence interval:

$$\frac{0.249 + \frac{1.96^2}{2(277)} \pm 1.96\sqrt{\frac{0.249 \cdot 0.751}{277} + \frac{1.96^2}{4(277)^2}}}{1 + (1.96^2)/277}$$

We obtain the 95% score confidence interval for π: $(0.202, 0.303)$.

For the 95% Agresti-Coull interval, we add two successes and two failures to obtain $\tilde{\pi} = \frac{71}{281} = 0.253$ with resulting 95% confidence interval:

$0.253 \pm 1.96\sqrt{\frac{(0.253)(0.747)}{281}}$ or $(0.202, 0.303)$.

3.1.3 Small Sample Inference for a Binomial Proportion

We have presented methods for inference concerning a binomial proportion that are based upon the normal approximation to the binomial distribution. This approximation does not apply when the sample size is small. Here we will use the binomial distribution directly to obtain $P-$values for a test.

Consider testing $H_0 : \pi = 0.6$ versus $H_a : \pi > 0.6$ for a binomial distribution with $n = 10$. If we observe $y = 9$ successes, the exact $P-$value equals

$$P(Y \geq 9) = \binom{10}{9} 0.6^9 0.4^1 + 0.6^{10} = 0.0403 + 0.0060 = 0.0463.$$

Using this $P-$value to decide whether or not to reject at a given level of significance α will lead to a conservative test. In this example, the probability that the $P-$value ≤ 0.05 when H_0 is true equals 0.0463, which is less than 0.05. This contrasts with the situation for tests with continuous test statistics, where the $P-$value under H_0 has a uniform distribution on the interval $(0, 1)$. An approach to providing a less conservative test is to use the mid $P-$value. We replace $P(y)$ by $P(y)/2$ in the computation of $P(Y \geq y)$. If we observe $y = 9$ successes, the mid $P-$value equals

$$\text{mid } P-\text{value} = \frac{1}{2}\binom{10}{9} 0.6^9 0.4^1 + 0.6^{10} = \frac{1}{2}(0.0403) + 0.0060 = 0.0262.$$

The following table provides the $P-$values and mid $P-$values for the one-sided test of $H_0 : \pi = 0.6$ versus $H_a : \pi > 0.6$:

y	$P(y)$	$P-$value	Mid $P-$value
0	0.0001	1.0000	0.9999
1	0.0016	0.9999	0.9991
2	0.0106	0.9983	0.9930
3	0.0425	0.9877	0.9665
4	0.1115	0.9452	0.8895
5	0.2007	0.8338	0.7334
6	0.2508	0.6331	0.5077
7	0.2150	0.3823	0.2748
8	0.1209	0.1673	0.1068
9	0.0403	0.0464	0.0262
10	0.0060	0.0060	0.0030

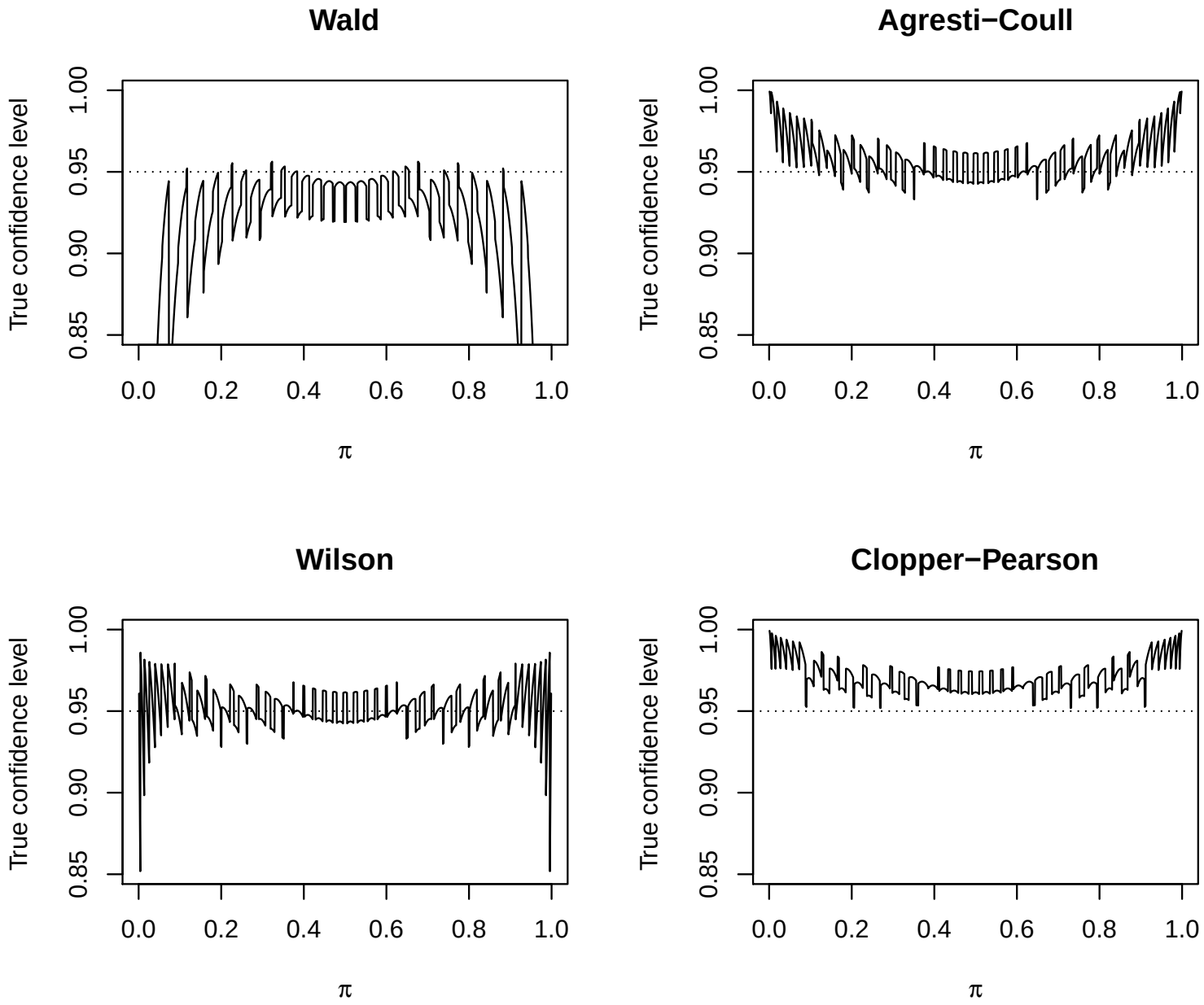
A small-sample confidence interval for the binomial proportion was developed by Clopper and Pearson (1934). The interval consists of alll the values of π_0 for which each one-sided binomial $P-$ value exceeds $\alpha/2$. The computation of the intervals is facilitated by the relationship of the binomial sum and the cumulative distribution function of the beta or F distributions. The interval is given by

$$\left[1+\frac{n-y+1}{yF_{2y,2(n-y+1)}(1-\alpha/2)}\right]^{-1}<\pi<\left[1+\frac{n-y+1}{(y+1)F_{2(y+1),2(n-y)}(\alpha/2)}\right]^{-1},$$

where $F_{a,b}(c)$ denotes the $1-c$ quantile of the F distribution with $df=(a,b)$. Due to the discreteness of the binomial distribution, the actual coverage probability of the Clopper-Pearson interval usually is greater than $1-\alpha$.

3.1.4 True Levels for Binomial Confidence Intervals

The various confidence level methods do not necessarily achieve the stated confidence level. The following plot shows the true confidence levels for the Wald, Wilson, Agresti-Coull, and Clopper-Pearson intervals when $n=40$ and $1-\alpha=0.95$ as a function of π.



3.2 Inference for the Poisson Mean

Suppose that $Y_1, \ldots, Y_n$ is a random sample from a Poisson distribution with mean μ. We wish to obtain the maximum likelihood estimator for μ and to form a confidence interval for μ.

The likelihood function is

$$\ell(\mu) = \prod_{i=1}^{n} P(y_i) = \prod_{i=1}^{n} \left[\frac{e^{-\mu} \mu^{y_i}}{y_i!} \right] = \left[\prod_{i=1}^{n} \frac{1}{y_i!} \right] e^{-n\mu} \mu^{\sum y_i}$$

To maximize the likelihood, we take the derivative of the log-likelihood:

$$\frac{\partial L(\mu)}{\partial \mu} = \frac{\partial}{\partial \mu} \left\{ \log \left[\prod_{i=1}^{n} \frac{1}{y_i!} \right] - n\mu + \log \mu \sum y_i \right\} = -n + \frac{\sum y_i}{\mu}$$

Set this equal to zero and solve for μ to get the m.l.e.:

$$\hat{\mu} = \frac{\sum y_i}{n} = \overline{y}$$

Properties of the M.L.E.

1. $E(\hat{\mu}) = \mu$
2. $\mathrm{Var}(\hat{\mu}) = \frac{\mu}{n}$
3. $SE = \sqrt{\frac{\mu}{n}}$
4. The distribution of $\hat{\mu}$ is approximately normal with the above mean and variance.

3.2.1 Confidence Intervals for μ

As we did for the binomial distribution, we can obtain confidence intervals for μ based on either the Wald statistic or the score statistic. The Wald statistic for testing $H_0 : \mu = \mu_0$ is

$$Z = \frac{\overline{Y} - \mu_0}{\sqrt{\overline{Y}/n}}$$

The resulting level $1 - \alpha$ confidence interval for μ is given by

$$\overline{Y} \pm Z_{\alpha/2}\sqrt{\overline{Y}/n}.$$

The score statistic for testing $H_0 : \mu = \mu_0$ is

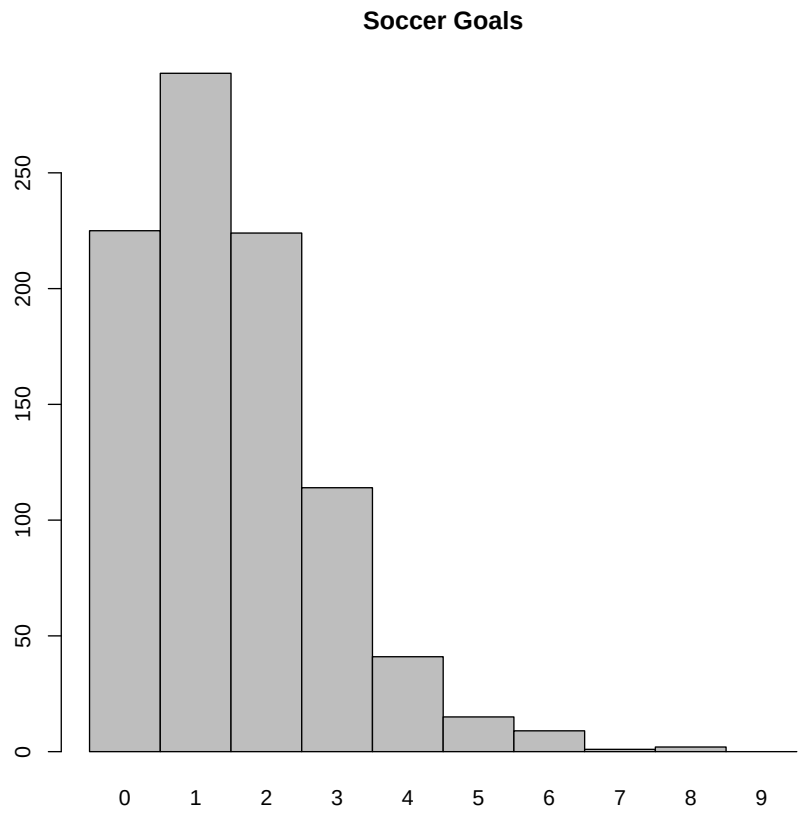
$$Z = \frac{\overline{Y} - \mu_0}{\sqrt{\mu_0/n}}$$

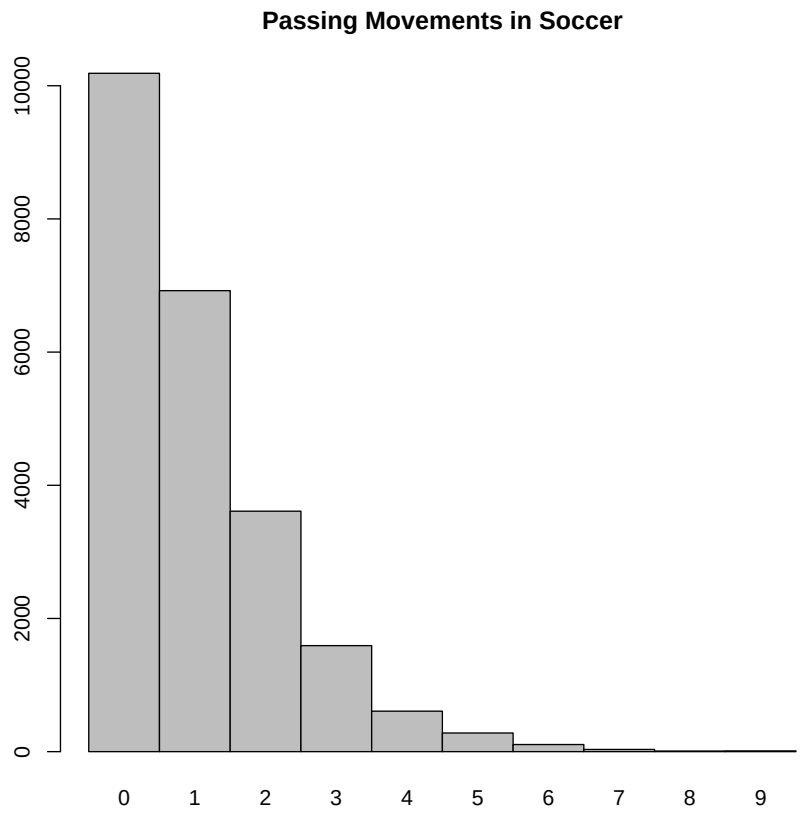
The resulting level $1 - \alpha$ confidence interval for μ is given by

$$\overline{Y} + \frac{Z^2_{\alpha/2}}{2n} \pm \frac{Z_{\alpha/2}}{\sqrt{n}} \sqrt{\overline{Y} + Z^2_{\alpha/2}/(4n)}.$$

Example Goals Scored in a Soccer Match

We consider a data set that consists of the soccer goals scored by each team in 462 games one season in an English soccer league. We also consider the number of passes in passing movements in 42 English soccer matches. Histograms and summary statistics for these data follow:





The sample mean, variance, and standard deviation for the goals and passes are

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 1.514,\ s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 1.745,\ s_X = \sqrt{1.745} = 1.328.$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 1.019,\ s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 1.503,\ s_Y = \sqrt{1.745} = 1.226.$$

Obtain a 95% Wald confidence interval for the mean number of goals:

$$\begin{aligned} \bar{x} &\pm Z_{\alpha/2}\sqrt{\bar{x}/n} \\ 1.514 &\pm 1.96\sqrt{1.514/924} \\ 1.514 &\pm 0.079 \end{aligned}$$

The 95% Wald interval is $(1.435, 1.593)$.

We can also compute the more accurate confidence interval:

$$1.514 + \frac{1.96^2}{2(924)} \pm \frac{1.96}{\sqrt{924}}\sqrt{1.514 + \frac{1.96^2}{4(924)}}$$

The 95% score interval is $(1.433, 1.596)$.

Using proc genmod in SAS, the 95% likelihood ratio interval is $(1.436, 1.595)$.

We can also use proc genmod to fit a negative binomial distribution to the data. In this case the 95% likelihood ratio interval for μ is $(1.430, 1.601)$, which is similar to that for the Poisson distribution. However, the confidence interval for the dispersion parameter D is $(0.038, 0.181)$. Since this interval does not include zero, there is some evidence of overdispersion in the data.

3.3 Inference for the Multinomial Distribution

The multinomial likelihood has the form

$$\ell(\pi_1, \ldots, \pi_c) = \frac{n!}{n_1! n_2! \cdots n_c!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}.$$

Using a derivation analogous to that for the mle of the binomial proportion, the mle for the multinomial probabilities $\pi_1, \ldots, \pi_c$ are

$$\hat{\pi}_1 = \frac{n_1}{n}, \cdots, \hat{\pi}_c = \frac{n_c}{n}.$$

Properties of the MLEs

Expected Value: $E(\hat{\pi}_i) = \pi_i$

Variance: $\mathrm{Var}(\hat{\pi}_i) = \pi_i(1 - \pi_i)/n$

Correlation: $\mathrm{Corr}(\hat{\pi}_i, \hat{\pi}_j) = -\sqrt{\frac{\pi_i \pi_j}{(1-\pi_i)(1-\pi_j)}}$

3.3.1 Testing Goodness of Fit for the Multinomial Distribution

Goodness-of-fit tests are used in categorical data to determine the adequacy of the assumed model. When the assumed model does not hold, statistical procedures designed for that model are often not useful.

Categorical data typically consists of counts of events in various categories. Goodness-of-fit tests compare the observed counts with the expected counts under a hypothesized model. Large differences indicate that the hypothesized model is not adequate.

Suppose that the random variables $\boldsymbol{N} = (N_1, \ldots, N_c)$ have a multinomial distribution, Mult$(n, \pi_1, \ldots, \pi_c)$. We wish to test the null hypothesis of specified cell probabilities:

$$H_0 : \pi_1 = \pi_{10}, \ldots, \pi_c = \pi_{c0} \qquad \text{versus} \qquad H_a : \text{some } \pi_i \neq \pi_{i0}.$$

When H_0 is true, the expected counts are $\mu_i = E(N_i) = n\pi_{i0}$.

We base our test on the differences between the observed and predicted cell frequencies using Pearson's chi-squared test statistic:

$$X^2 = \sum_{i=1}^{c} \frac{(n_i - \mu_i)^2}{\mu_i}$$

When H_0 is true, X^2 has approximately a chi-squared distribution with $c - 1$ degrees of freedom. We reject H_0 for large values of X^2.

An alternative statistic is the *likelihood ratio statistic*. The general form of the likelihood ratio statistic is

$$\Lambda = \frac{\text{Maximized likelihood under } H_0}{\text{Maximized likelihood for unrestricted } \boldsymbol{\pi}}.$$

- When the null hypothesis is true, the maximized restricted likelihood under the null hypothesis will be similar in value to the maximized unrestricted likelihood.
- However, if the null hypothesis is false, the maximized unrestricted likelihood can be much larger.

Thus, we will reject H_0 when Λ is small, or equivalently, when $G^2 = -2\log(\Lambda)$ is large.

For testing goodness of fit for the multinomial distribution, the LR statistic equals

$$G^2 = 2\sum_{i=1}^{c} n_i \log\left(\frac{n_i}{\mu_i}\right).$$

When H_0 is true, G^2 has approximately a chi-squared distribution with $c-1$ degrees of freedom. We reject H_0 for large values of G^2.

Remarks:

1. Pearson's statistic and the LR statistic will have similar values, particularly for large samples.
2. Both statistics will equal zero if $n_i = \mu_i$ for all i.
3. The approximation to the distribution of these statistics using the chi-squared distribution is valid when n is large enough so that all the μ_i values are large.
4. Generally both statistics will lead to the same inference for any given set of data. However, if the values of μ_i are small, the two statistics can differ quite a bit in value.

Example: Case Study from *Outliers*

Malcolm Gladwell in the book *Outliers* examined the background and careers of various highly successful people to try to identify possible causes for their success. One of his studies involved junior hockey players from Canada. Here is the player roster of the 2007 Medicine Hat Tigers. Take a close look and see if you can spot anything strange about it.

Number	Name	Position	L/R	Height	Weight	Birth Date	Hometown
9	Brennan Bosch	C	R	5' 8"	173	February 14, 1988	Martensville, SK
11	Scott Wasden	C	R	6' 1"	188	January 4, 1988	Westbank, BC
12	Colton Grant	LW	L	5' 9"	177	March 20, 1989	Standard, AB
14	Darren Helm	LW	L	6'	182	January 21, 1987	St. Andrews, MB
15	Derek Dorsett	RW	L	5' 11"	178	December 20, 1986	Kindersley, SK
16	Daine Todd	C	R	5' 10"	173	January 10, 1987	Red Deer, AB
17	Tyler Swynstun	RW	R	5' 11"	185	January 15, 1988	Cochrane, AB
19	Matt Lowry	C	R	6'	186	March 2, 1988	Neepawa, MB
20	Kevin Undershute	LW	L	6'	178	April 12, 1987	Medicine Hat, AB
21	Jerrid Sauer	RW	R	5' 10"	196	September 12, 1987	Medicine Hat, AB
22	Tyler Ennis	C	L	5' 9"	160	October 6, 1989	Edmonton, AB
23	Jordan Hickmott	C	R	6'	183	April 11, 1990	Mission, BC
25	Jakub Rumpel	RW	R	5' 8"	166	January 27, 1987	Hrnciarovce, SLO
28	Bretton Cameron	C	R	5' 11"	168	January 26, 1989	Didsbury, AB
36	Chris Stevens	LW	L	5' 10"	197	August 20, 1986	Dawson Creek, BC
3	Gord Baldwin	D	L	6' 5"	205	March 1, 1987	Winnipeg, MB
4	David Schlemko	D	L	6' 1"	195	May 7, 1987	Edmonton, AB
5	Trever Glass	D	L	6'	190	January 22, 1988	Cochrane, AB
10	Kris Russell	D	L	5' 10"	177	May 2, 1987	Caroline, AB
18	Michael Sauer	D	R	6' 3"	205	August 7, 1987	Sartell, MN
24	Mark Isherwood	D	R	6'	183	January 31, 1989	Abbotsford, BC
27	Shayne Brown	D	L	6' 1"	198	February 20, 1989	Stony Plain, AB
29	Jordan Benfield	D	R	6' 3"	230	February 9, 1988	Leduc, AB
31	Ryan Holfeld	G	L	5' 11"	166	June 29, 1989	LeRoy, SK
33	Matt Keetley	G	R	6' 2"	189	April 27, 1986	Medicine Hat, AB

	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec	Total
n_i	14	6	3	2	25
π_{i0}	0.25	0.25	0.25	0.25	1
μ_i	6.25	6.25	6.25	6.25	25

$H_0 : \pi_1 = 0.25,\ \pi_2 = 0.25,\ \pi_3 = 0.25,\ \pi_4 = 0.25$

H_a : At least one probability is not as specified.

Test statistic: X^2 **Rejection region:** $X^2 > 11.34$

$$X^2 = \frac{(14-6.25)^2}{6.25} + \frac{(6-6.25)^2}{6.25} + \frac{(3-6.25)^2}{6.25} + \frac{(2-6.25)^2}{6.25} = 14.2$$

Reject H_0 and conclude that the proportions of the hockey players being born the various quarters of the year differ.

The likelihood ratio statistic is

$$G^2 = 2\left(14\log\left(\frac{14}{6.25}\right) + 6\log\left(\frac{6}{6.25}\right) + 3\log\left(\frac{3}{6.25}\right) + 2\log\left(\frac{2}{6.25}\right)\right) = 13.13,$$

resulting in the same conclusion.

3.3.2 Testing Goodness of Fit with Estimated Parameters

In many situations the cell probabilities depend on a small number of parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$, where the value of $\boldsymbol{\theta}$ is not specified. We wish to test whether the cell probabilities are specified by $\pi_i = \pi_i(\boldsymbol{\theta})$, $i = 1, \ldots, c$. We need to estimate $\boldsymbol{\theta}$ to estimate the expected cell counts under H_0.

$H_0 : \pi_1 = \pi_1(\boldsymbol{\theta}), \ldots, \pi_c = \pi_c(\boldsymbol{\theta})$, some $\boldsymbol{\theta}$

$H_a :$ H_0 is not true

When H_0 is true, the joint pmf of $(n_1, \ldots, n_c)$ is

$$P(n_1, \ldots, n_c) = \frac{n!}{n_1! \cdots n_c!} \pi_1(\boldsymbol{\theta})^{n_1} \cdots \pi_c(\boldsymbol{\theta})^{n_c}$$

Use **maximum likelihood** to estimate $\boldsymbol{\theta}$:

Choose $\hat{\boldsymbol{\theta}}$ that maximizes $L(\boldsymbol{\theta}) = \log P(n_1, \ldots, n_k)$.

Test statistic:

$$X^2 = \sum_{i=1}^{c} \frac{\left(n_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i}$$

where $\hat{\mu}_i = n\pi_i(\hat{\boldsymbol{\theta}})$.

This is Pearson's goodness-of-fit statistic.

Rejection region: We lose one d.f. for each estimated parameter. The rejection region is $X^2 > \chi^2_\alpha$ where χ^2 is the chi-squared distribution with $c - 1 - q$ degrees of freedom.

Basic Guide: Want $\hat{\mu}_i \geq 5$ for all cells.

A common use of this statistic is testing goodness of fit of a discrete distribution to a set of data. We first estimate the parameters of the distribution to get estimated cell probabilities. These are used to compute Pearson's statistic.

Example: Eighty litters, each containing 3 rabbits, were observed and the number Y of male rabbits in the litter was recorded, resulting in the following frequency distribution:

# of males in litter	0	1	2	3	Total
# of litters	19	32	22	7	80

If the assumption of Bernoulli trials is correct for the sex of rabbits, the distribution of Y should be binomial with $n = 3$ and $\theta =$ probability of a male birth.

$H_0 : \ \pi_i(\theta) = \binom{3}{i}\theta^i(1-\theta)^{3-i}, \ i = 0, 1, 2, 3$

$H_a :$ At least one probability is not as specified.

1. Estimate θ. We need to maximize:

$$\ell(\theta) = \pi_0(\theta)^{n_0}\pi_1(\theta)^{n_1}\pi_2(\theta)^{n_2}\pi_3(\theta)^{n_3}$$

Equivalently, maximize

$$L(\theta) = \sum_{i=0}^{3} n_i \log(\pi_i(\theta))$$

For this problem,

$$L(\theta) = \sum_{i=0}^{3} n_i \left[\log\left(\binom{3}{i} \right) + i \log(\theta) + (3-i)\log(1-\theta) \right]$$

Take the derivative and set equal to zero:

$$L'(\theta) = \sum_{i=0}^{3} \left[\frac{n_i i}{\theta} - \frac{n_i(3-i)}{1-\theta} \right] = 0$$

After some algebra,

$$(1-\theta)\sum_{i=0}^{3} n_i i = \left(3\sum_{i=0}^{3} n_i - \sum_{i=0}^{3} n_i i \right) \theta$$

Solve for θ:

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{i=0}^{3} n_i i}{3\sum_{i=0}^{3} n_i} \\ &= \frac{\text{Number of Male Rabbits}}{\text{Total Number of Rabbits}} \end{aligned}$$

Here

$$\hat{\theta} = \frac{0 + 32 + 44 + 21}{3 \times 80} = \frac{97}{240} = 0.404$$

2. Test goodness of fit to the binomial distribution with $n = 3$ and $\pi = 0.404$.

# of males in litter	0	1	2	3	Total
observed	19	32	22	7	80
Binomial prob.	0.2117	0.4305	0.2918	0.0659	1
$\hat{\mu}_i$	16.94	34.44	23.35	5.28	80
$(O-E)^2/E$	0.25	0.17	0.08	0.46	1.07

Test statistic:

$$X^2 = \sum \frac{(O-E)^2}{E}$$

Rejection region: $X^2 > \chi^2_{4-1-1,.05} = 5.99$

Do not reject H_0 and conclude that the binomial distribution is consistent with the number of male rabbits per litter.

Remarks:

1. Failure to reject H_0 in a goodness-of-fit test does not imply that the hypothesized distribution is the true distribution. It is merely one that is consistent with the data.

2. Failure to reject H_0 can occur because the chi-squared test may not be powerful for certain alternatives. When the categories have some ordering, Pearson's chi-squared statistic can be partitioned to obtain tests with higher power for certain alternatives.

3. We often need to group cells so that the expected cell frequencies are all at least 5. If so, we should estimate the parameters using the grouped data likelihood.

4. The use of the chi-squared distribution is based on having a large sample resulting in large (≥ 5) expected cell counts. If some of the expected cell counts are small, then one can use exact methods for carrying out the test.