Categorical Data Analysis
Lecture Note 2

Instructor: Seokho Lee

Hankuk University of Foreign Studies

2. Categorical Data and Goodness of Fit

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

- Binomial distribution
- Poisson distribution
- Multinomial distribution

2.1. Binomial Probability Distribution

A **binomial experiment** satisfies the following 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: $\Pr(S) = p$

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

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

n = fixed number of trials
p = probability of success

$$f(y : n, p) = \Pr(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y = 0, 1, 2, \cdots, n$$

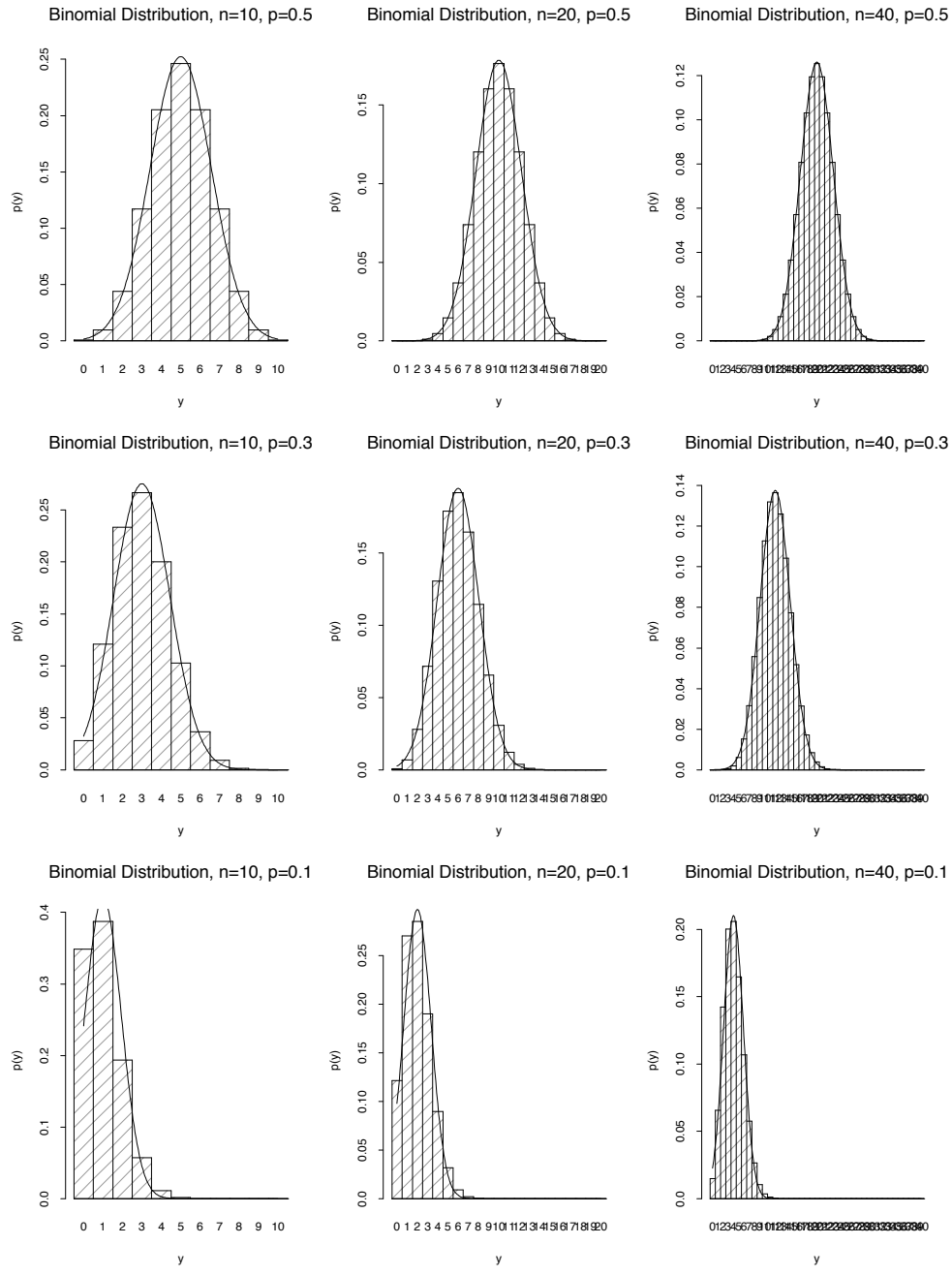
where

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

$$\mu = E(Y) = np$$

$$\sigma^2 = Var(Y) = np(1-p)$$

Some Binomial Distributions



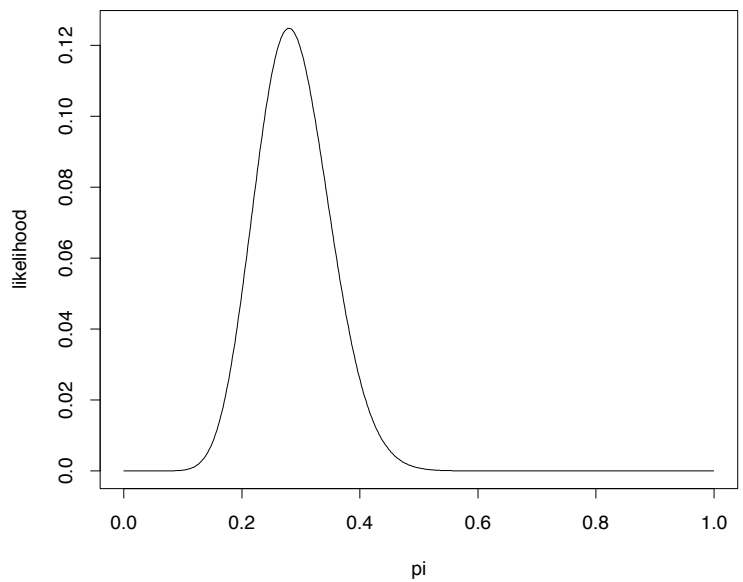
2.1.1. Review of Inference for a Single Proportion

Suppose we observe $Y \sim \text{Binomial}(n, p)$. Our goal is to estimate p.

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.

$$f(14 : 50, p) \;=\; \Pr(Y = 14) \;=\; \binom{50}{14} p^{14}(1-p)^{50-14}$$



Suppose $L(p) = \log f(p)$. The value of p that maximizes $L(p)$ is the maximum likelihood estimation (m.l.e.), $\hat{p} = \frac{14}{50} = 0.28$. We can verify that this value maximizes the likelihood using calculus. We take the derivative of the natural logarithm of the likelihood:

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

We set this equal to zero and solve for p:

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

Properties of M.L.E

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

Fact: If $Y \sim \text{Binomial}(n, p)$ and the probability histogram is not too skewed ($np > 5$, $n(1-p) > 5$), then Y is approximately normally distributed. Similarly, the m.l.e. $\hat{p} = \frac{Y}{n}$ is approximately normally distributed.

2.1.2. A Large-Sample Interval for p (Population Proportion)

We want to construct a confidence interval for the population proportion p where

$$Y \sim \text{Binomial}(n, p) \qquad \mu = np \qquad \sigma = \sqrt{np(1-p)}$$

We standardize

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

Note that the standard deviation involves the unknown parameter p so we substitute our estimate of p into the equation. Thus,

$$\Pr\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

Solving for our parameter p yields (approximately):

$$\Pr\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right) \approx 1 - \alpha$$

Thus, a 95% confidential interval for p is

$$\hat{p} \pm 1.96\widehat{s.e.}(\hat{p}) \quad \text{where } \widehat{s.e.}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$$

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 p.

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

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

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 K mutually exclusive categories.
3. Trials are independent.
4. The probability of the kth category (p_k) is the same for each trial, where $p_1 + \cdots + p_K = 1$.

The observable random variables are $(n_1, n_2, \cdots, n_K)$ where n_k =number of trials resulting in the kth category, where $n_1 + \cdots + n_K = n$. The random variables $\mathbf{n} = (n_1, \cdots, n_K)$ are said to have a multinomial distribution, Mult$(n, \mathbf{p})$ with probability function

$$f(\mathbf{n}; n, \mathbf{p}) = \frac{n!}{n_1! n_2! \cdots n_K!} p_1^{n_1} p_2^{n_2} \cdots p_K^{n_K} = \frac{n!}{\prod_{i=1}^{K} n_i!} \prod_{i=1}^{K} p_i^{n_i}$$

Expected value : $E(n_k) = np_k = E_k$

Variance : $Var(n_k) = np_k(1 - p_k)$

Covariance : $cov(n_i, n_j) = -np_i p_j$

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:

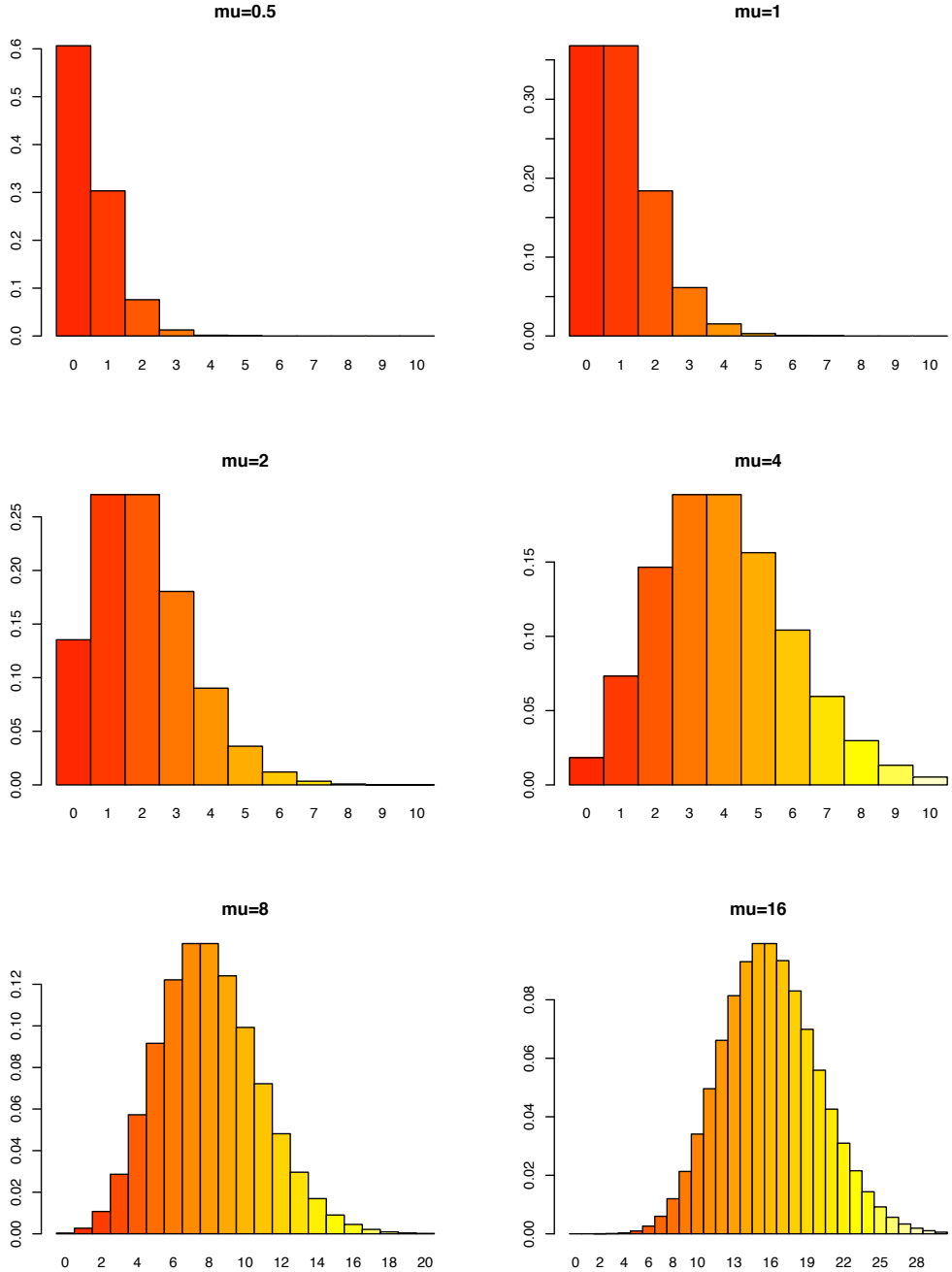
$$Y \sim \text{Poisson}(\mu)$$

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

$$f(y;\mu) = \Pr(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}, \qquad \mu > 0,\ y = 0, 1, 2, \cdots$$

$$E[Y] = \mu, \qquad Var[Y] = \mu$$

Some Poisson Distributions



2.4. 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 $\mathbf{n} = (n_1, \cdots, n_K)$ have a multinomial distribution, Mult$(n, \mathbf{p})$. We wish to test the hypothesis,

$$H_0 : \mathbf{p} = \mathbf{p}_0 \quad \text{versus} \quad H_a : \mathbf{p} \neq \mathbf{p}_0$$

When H_0 is true, $E(\mathbf{n}) = \mathbf{e} = n\mathbf{p}_0$ where the expected counts are $\mathbf{e} = (e_1, \cdots, e_K)$ and $e_i = np_i$. We base our test on the differences between the observed and predicted cell frequencies using *Pearson's test statistic*:

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

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

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

$$\Lambda = \frac{\text{Maximized likelihood under } H_0}{\text{Maximized likelihood for unrestricted } \mathbf{p}}$$

When the null hypothesis is true, the restricted likelihood under the null hypothesis will be similar in value to the unrestricted likelihood. However, if the null hypothesis is false, the restricted likelihood can be much smaller. Thus, we will reject H_0 when Λ is small, or equivalently, when $G^2 = -2 \log \Lambda$ is large. Here the LR statistic equals

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

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

Example: A study is run to see whether the public favors the construction of a new dam. It is thought that 40% favor dam construction, 30% are neutral, 20% oppose the dam, and the rest have not thought about it. A random sample of 150 individuals are interviewed resulting in 42 in favor, 61 neutral, 33 opposed, and the rest having not thought about it. Do the data indicate that the stated proportions are incorrect? Use $\alpha = 0.01$.

$$H_0 \ : \ p_1 = 0.4, \ p_2 = 0.3, \ p_3 = 0.2, \ p_4 = 0.1$$

$$H_a \ : \ \text{At least one probability is not as specified.}$$

	Favor	Neutral	Oppose	Unaware	Total
n_i	42	61	33	14	150
p_{i0}	0.4	0.3	0.2	0.1	1
e_i	60	45	30	15	150

Rejection region: X^2 or $G^2 > \chi^2_{0.01}(4-1) = 11.34$

$$X^2 = \frac{(42-60)^2}{60} + \frac{(61-45)^2}{45} + \frac{(33-30)^2}{30} + \frac{(14-15)^2}{15} = 11.46$$

$$G^2 = 2\left\{42\log\left(\frac{42}{60}\right) + 61\log\left(\frac{61}{45}\right) + 33\log\left(\frac{33}{30}\right) + 14\log\left(\frac{14}{15}\right)\right\} = 11.51$$

Both of X^2 and G^2 reject H_0 and conclude that the proportions of the public having the various views differs from that hypothesized.