Chapter 2
Single Probability

Abstract The Binomial distribution and its properties are discussed in detail including maximum likelihood estimation of the probability p. Exact and approximate hypothesis tests and confidence intervals are provided for p. Inverse sampling and the Negative Binomial Distribution are also considered.

Keywords Bernoulli trials · Maximum likelihood estimate · Likelihood-ratio test · Inverse sampling · Negative-Binomial distribution · Exact hypothesis test for a probability · Exact and approximate confidence intervals for a probability · Poisson approximation to the Binomial distribution

2.1 Binomial Distribution

An important model involving a probability is the Binomial distribution. It was mentioned in Chap. 1 as an approximation for the Hypergeometric distribution. It is, however, an important distribution in its own right as it arises when we have a fixed number n of Bernoulli experiments or "trials." Such trials satisfy the following three assumptions:

1. The trials are mutually independent.
2. Each trial has only two outcomes, which we can label "success" or "failure."
3. The probability p $(= 1 - q)$ of success is constant from trial to trial.

2.1.1 Estimation

If X is the number of successes from a fixed number n of Bernoulli trials, then X has the Binomial probability function

$$f_1(x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, \ldots, n. \tag{2.1}$$

G. A. F. Seber, *Statistical Models for Proportions and Probabilities*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-39041-8_2,

Using the Binomial Theorem we note that

$$\sum_{x=0}^{n} f_1(x) = (p+q)^n = 1.$$

Ignoring constants, the likelihood function is $L(p) = p^x q^{n-x}$ so that the maximum likelihood estimator of p is obtained by setting

$$\frac{\partial \log L(p)}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p} = 0,$$

namely $\widehat{P} = x/n$. It is a maximum as the second derivative is negative. Also

$$\begin{aligned} -\mathrm{E}\left[\frac{\partial^2 \log L(p)}{\partial p^2}\right] &= \mathrm{E}\left[\frac{X}{p^2} + \frac{n-X}{(1-p)^2}\right] \\ &= \frac{n}{p} + \frac{n}{1-p} \\ &= \frac{n}{pq}, \end{aligned}$$

which is the inverse of the Cramér-Rao lower bound. As $\widehat{P} = X/n$ is unbiased and has variance pq/n, it is the minimum variance unbiased estimator of p.

2.1.2 Likelihood-Ratio Test

To test the hypothesis $H_0 : p = p_0$ versus the alternative $H_a : p \neq p_0$ we can use the likelihood-ratio test

$$\Lambda_n = \frac{L(p_0)}{\sup_p L(\widehat{p})} = \frac{p_0^x (1-p_0)^{n-x}}{\widehat{p}^x (1-\widehat{p})^{n-x}}.$$

When H_0 is true, $-2\log\Lambda_n$ is asymptotically distributed as χ_1^2, the Chi-square distribution with one degree of freedom. We reject H_0 at the α level of significance if the observed value of $-2\log\Lambda_n$ exceeds $\chi_1^2(\alpha)$, the upper α tail value.

2.1.3 Some Properties of the Binomial Distribution

Moments of the Binomial distribution can be found by differentiating or expanding its moment generating function

$$M_x(t) = \mathrm{E}(e^{tx}) = (q + e^t p)^n.$$

For example, $\mathrm{E}(X) = M_x'(0) = np$ and $\mathrm{E}(X^2) = \frac{1}{2!}M_X''(0) = n(n-1)p^2 + np$. Factorial moments are sometimes useful such as (setting $y = x - r$)

$$\begin{aligned}
\mathrm{E}[X(X-1)\cdots(X-r+1)] &= \sum_{x=r}^{n} x(x-1)\cdots(x-r+1)\frac{n!}{x!(n-x)!}p^x q^{n-x} \\
&= p^r n\cdots(n-r+1)\sum_{y=0}^{n-r}\binom{n-r}{y}p^y q^{n-r-y} \\
&= p^r n(n-1)\cdots(n-r+1)(p+q)^{n-r} \\
&= p^r n(n-1)\cdots(n-r+1).
\end{aligned}$$

For example, setting $r = 2$, $\mathrm{E}[X(X-1)] = p^2 n(n-1)$.

One other result that has been found useful is in the situation of studying variables like $\widehat{P}^{-1} = n/X$ as an estimate of p^{-1}. This raises problems as we can have $X = 0$. A useful idea is to modify the variable and consider

$$\begin{aligned}
\mathrm{E}\left(\frac{n+1}{X+1}\right) &= \frac{1}{p}\sum_{i=0}^{n}\frac{n+1!}{(x+1)!(n+1-x-1)!}p^{x+1}q^{n+1-(x+1)} \\
&= \frac{1}{p}\sum_{y=1}^{n+1}\binom{n+1}{y}p^y q^{n+1-y} \quad (y = x+1) \\
&= \frac{1}{p}[(p+q)^{n+1} - q^{n+1}] \\
&= \frac{1}{p}(1 - q^{n+1}).
\end{aligned}$$

For large n, $(n+1)/(X+1)$ is an approximately unbiased estimate of $1/p$. This technique works well with a number of other discrete distributions as we saw in Chap. 1.

When $n = 1$, X becomes an indicator variable J, say, where $J = 1$ with probability p, and $J = 0$ with probability q. Then $\mathrm{E}(J) = p$ and

$$\mathrm{var}(J) = \mathrm{E}(J^2) - (\mathrm{E}(J))^2 = p^2 - p = pq.$$

If J_i is the indicator variable associated with the ith trial, we can now write $X = \sum_{i=1}^n J_i$ with mean np and variance $\sum_{i=1}^n \mathrm{var}(J_i) = npq$. Also, $\widehat{P} = X/n = \bar{J}$ so that by the Central Limit Theorem $\widehat{P}$ is asymptotically $N(p, pq/n)$.

2.1.4 Poisson Approximation

If we let $p \to 0$ and $n \to \infty$ such that $\lambda_n = np \to \lambda$, where λ is a constant, then the Binomial moment generating function is given by

$$(q + pe^t)^n = \left[1 - \frac{\lambda_n}{n}(1 - e^t)\right]^n \to e^{-\lambda(1-e^t)},$$

the moment generating function of the Poisson distribution, Poisson(λ), with mean λ. We see then that the Binomial distribution can be approximated by the Poisson distribution with mean np when p is small and n is large. Quantiles of the Poisson distribution can be obtained from the Chi-square distribution using the result

$$\text{pr}(Y \leq x) = \text{pr}(\chi^2_{2(1+x)} \leq 2\lambda),$$

where $Y \sim$ Poisson(λ). We could use this result to construct an approximate confidence interval for np and hence for p.

2.2 Inverse Sampling

Suppose we have a sequence of Bernoulli trials that continues until we have r successes. If W is the number of failures, then the sample size $Y = W + r$ is random with the last trial being a success. Hence, W has a Negative-Binomial distribution with probability function

$$f_2(w) = \binom{w + r - 1}{r - 1} p^{r-1} q^w \cdot p = \binom{w + r - 1}{r - 1} p^r q^w, \quad w = 0, 1, \ldots.$$

This probability function can be expressed in a number of different ways (Johnson et al. 2005, Chap. 5). The moment generating function of W is $M_w(t) = (Q_0 - P_0 e^t)^{-r}$, where

$$P_0 = \frac{1 - p}{p} \quad \text{and} \quad Q_0 = \frac{1}{p}.$$

Differentiating $M_w(t)$ leads to

$$\text{E}(W) = rP_0 \quad \text{and} \quad \text{var}(W) = rP_0 Q_0.$$

We now wish to find an unbiased estimator of p and an unbiased estimate of its variance. We do this using estimators due to Murthy (1957) that were shown by Salehi and Seber (2001) to apply to inverse sampling. Since W is a complete sufficient statistic for p, it can be shown that the minimum variance unbiased estimator for p is

$$\widehat{P}_{in} = \frac{r-1}{r+W-1} = \frac{r-1}{Y-1},$$

which is the same as for sampling without replacement (Sect. 1.2). This is perhaps not surprising as the same equality occurs with simple random sampling with or without replacement. An unbiased variance estimator of $\mathrm{var}(\widehat{P}_{in})$ is (Salehi and Seber 2001)

$$\widehat{\mathrm{var}}(\widehat{P}_{in}) = \frac{\widehat{P}_{in}(1-\widehat{P}_{in})}{Y-2}.$$

Unbiasedness can also be proved directly using the methods of Sect. 1.2.

2.3 Inference for a Probability

There is a considerable literature on confidence intervals for the Binomial distribution. We begin by considering so-called "exact" confidence intervals mentioned in Sect. 1.4, which are confidence intervals based on the exact Binomial distribution and not on an approximation for it. This also leads to an exact hypothesis test. Because of the discreteness of the distribution we cannot normally obtain a confidence interval with an exact prescribed confidence of $(1-\alpha)\,\%$ but rather we aim for a (conservative) confidence level of at least $100(1-\alpha)\,\%$. After considering exact intervals we will then derive some approximate intervals and tests based on approximations for the Binomial distribution.

2.3.1 Exact Intervals

Given an observed value $X = x$ for a Binomial distribution, we can follow the method described in Sect. 1.4 to obtain an exact confidence interval. We want to find probabilities p_L and p_U such that, for a two-sided confidence interval with confidence $100(1-\alpha)\,\%$,

$$\mathrm{pr}(X \geq x \mid p = p_L) = \sum_{i=x}^{n} \binom{n}{i} p_L^i (1-p_L)^{n-i} = \frac{\alpha}{2},$$

and

$$\mathrm{pr}(X \leq x \mid p = p_U) = \sum_{i=0}^{x} \binom{n}{i} p_U^i (1-p_U)^{n-i} = \frac{\alpha}{2}.$$

The interval (p_L, p_U) is known as the Clopper-Pearson confidence interval (Clopper and Pearson 1934). The tail of the Binomial distribution can be related to the tail of the F-distribution through the relationship (Jowett 1963)

$$\sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i} = \text{pr}\left\{ Y \leq \frac{(1-p)(x+1)}{p(n-x)} \right\}, \tag{2.2}$$

where Y has the F-distribution $F(2(n-x), 2(x+1))$. If $F_{1-\frac{\alpha}{2}}(\cdot,\cdot)$ denotes the $100(1-\alpha/2)$th percentile of the F-distribution, then[1]

$$p_L = \frac{x}{x + (n-x+1)F_{1-\frac{\alpha}{2}}(2(n-x+1), 2x)}$$

and

$$p_U = \frac{(x+1)F_{1-\frac{\alpha}{2}}(2(x+1), 2(n-x))}{n-x+(x+1)F_{1-\frac{\alpha}{2}}(2(x+1), 2(n-x))}.$$

The percentiles of the F-distribution are provided by a number of statistical software packages. The above interval is very conservative and the coverage probability often substantially exceeds $1-\alpha$. One way of dealing with this is to use the so-called "mid p-value" where only half of the probability of the observed result is added to the probability of more extreme results (Agresti 2007, pp. 15–16). This method is particularly useful for very discrete distributions (i.e., with few well-spaced observed values). One theoretical justification for its use is given by Hwang and Yang (2001). Further comments about the method are made by Berry and Armitage (1995).

2.3.2 Exact Hypothesis Test

There is some controversy as to how to carry out an "exact" test of the hypothesis $H_0 : p = p_0$ versus the two-sided alternative $H_a : p \neq p_0$. One method is as follows. If $\widehat{p} < p_0$, we evaluate

$$g(p_0) = \sum_{i=0}^{x} \binom{n}{i} p_0^i (1-p_0)^{n-i} = \gamma,$$

where 2γ is the p-value of the test. If $\widehat{p} > p_0$, we evaluate $1 - g(p_0) + \text{pr}(\mathrm{X} = \mathrm{x}) = \delta$, where 2δ is the p-value of the test. We can use (2.2) to evaluate the p-value exactly or use F-tables if the software is not available. This method is referred to as the TST or Twice the Smaller Tail method by Hirji (2006, p. 59). As p-values tend to be too large, some statisticians prefer to use the mid p-value, as mentioned above. It involves halving the probability of getting the observed value x under the assumption of H_0 being true. Hirji (2006, pp. 70–73) also defines three other methods for carrying out the test,[2] and discusses exact tests in general.

[1] See, for example, http://www.ppsw.rug.nl/~boomsma/confbin.pdf.

[2] See also Fay (2010).

2.3.3 Approximate Confidence Intervals

We saw above that $\widehat{P}$ is asymptotically $N(p, pq/n)$, or more rigorously,

$$\frac{\widehat{P}-p}{\sqrt{pq/n}} \text{ is asymptotically } N(0,1).$$

An approximate two-sided $100(1-\alpha)\,\%$ confidence interval for p therefore has upper and lower limits that are solutions of the quadratic equation

$$(\widehat{p}-p)^2 = z(\alpha/2)^2\,\frac{p(1-p)}{n}, \tag{2.3}$$

where $z(\alpha/2)$ is the $\alpha/2$ tail value of standard Normal N(0,1) distribution. This confidence interval is usually referred to as the score confidence interval, though it is also called the Wilson interval introduced in 1927 as it inverts the test $H_0 : p = p_0$ obtained by substituting p for p_0 in (2.3). We note for later reference that $z(\alpha/2)^2 = \chi_1^2(\alpha)$, where the latter is the $1-\alpha$ quantile of the Chi-squared distribution with one degree of freedom.

An alternative method is based on the fact that

$$\frac{\widehat{P}-p}{\sqrt{\widehat{P}(1-\widehat{P})/n}} \text{ is also asymptotically } N(0,1)$$

yielding the confidence interval

$$\widehat{p} \pm z(\alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}. \tag{2.4}$$

This is usually referred as the Wald confidence interval for p, since it results from inverting the Wald test for p. It is the set of p_0 values having a p-value exceeding α in testing $H_0 : p = p_0$ versus $H_a : p \neq p_0$ using the test statistic $z = (\widehat{p} - p_0)/\sqrt{\widehat{p}(1-\widehat{p})/n}$.

Agresti and Coull (1998) compared the above two methods and the exact confidence interval and came to a number of conclusions. First, the score interval performed the best in having coverage probabilities close to the nominal confidence level. Second, they recommended its use with nearly all sample sizes and parameter values. Third, the exact interval remains quite conservative even for moderately large sample sizes when p tends to be 0 or 1. Fourth, the Wald interval fairs badly when p is near 0 or 1, one reason being that $\widehat{p}$ is used as the midpoint of the interval when the Binomial distribution is highly skewed. Finally, they provided an adaption of the Wald interval with $\widehat{p} = x/n$ replaced by $(x+2)/(n+4)$ in (2.4) (the "add two successes and two failures" rule) that also performs well even for small samples when $\alpha = 0.05$.

Brown et al. (2001) discussed the extreme volatility and oscillation of the Wald interval's behavior, due to the discreteness of the Binomial distribution, even when n is quite large and p is not near 0 or 1. They showed that the usual rules given in texts for when the Wald interval is satisfactory (e.g., $np, n(1-p)$ are ≥ 5 (or 10)) are somewhat defective and recommended three intervals: (1) the score (Wilson) interval, (2) an adjusted Wald interval they call the Agresti-Coull interval, and (3) an interval based on Jeffrey's prior distribution for p that they call Jeffrey's interval.

Solving the quadratic (2.3), the Wilson interval can be put in the form, after some algebra,

$$p \in \tilde{p} \pm \frac{\kappa n^{1/2}}{n+\kappa^2}(\widehat{pq} + \kappa^2/(4n))^{1/2},$$

where

$$\kappa = z(\alpha/2), \quad \text{and} \quad \tilde{p} = \frac{x+\kappa^2/2}{n+\kappa^2} = \frac{\tilde{x}}{\tilde{n}}, \text{ say.}$$

The Agresti-Coull interval takes the form

$$p \in \tilde{p} \pm \kappa\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}.$$

Note that $z(0.025) = 1.96 \approx 2$, which gives the above "add 2" rule. These two intervals have the same recentering that can increase coverage significantly for p away from 0 or 1 and eliminate systematic bias. Further simulation support for the "add 2" rule is given by Agresti and Caffo (2000).

Using a Beta prior distribution for p, say $\text{Beta}(a_1, a_2)$ (a conjugate prior for the Binomial distribution), it can be shown that the posterior distribution for p is $\text{Beta}(a_1 + x, a_2 + n - x)$. Using Jeffrey's prior ($a_1 = 1/2$, $a_2 = 1/2$), a $100(1-\alpha)$ % equal-tailed posterior confidence interval $(p_L(x), p_U(x))$ is given by

$$p_L(x) = B\left(\frac{\alpha}{2}; x+\frac{1}{2}, n-x+\frac{1}{2}\right) \text{ and } p_U(x) = B\left(1-\frac{\alpha}{2}; x+\frac{1}{2}, n-x+\frac{1}{2}\right).$$

where $p_L(0) = 0$, $p_U(n) = 1$, and $B(\gamma; m_1, m_2)$ is the γ quantile of the $\text{Beta}(m_1, m_2)$ distribution. Brown et al. (2001) considered further modifications when p is near 0 or 1. They recommended either the Wilson or Jeffrey's intervals for $n \leq 40$, while all three are fairly similar for $n > 40$, with the Agresti-Coull interval being easy to use and remember, though a little wider.

Brown et al. (2002) added as a contender the interval obtained by inverting the likelihood-ratio test that accepts the null hypothesis $H_0 : p = p_0$ of Sect. 2.1.2 at the α level of significance, though it requires some computation. It takes the form

$$\left\{ p : x \log p + (n - x) \log (1 - p) \geq x \log \widehat{p} + (n - x) \log (1 - p) - \chi_1^2 \left(\frac{\alpha}{2} \right) \right\}.$$

They also concluded that the Wilson, the likelihood ratio, and Jeffrey's intervals are comparable in both coverage and length, though the Jeffrey's interval is a bit shorter on average. Further comments are made by Newcombe (1998), who compared a number of intervals.

References

Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: Wiley Interscience.

Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, *54*(4), 280–288.

Agresti, A., & Coull, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126.

Berry, G., & Armitage, P. (1995). Mid-P confidence intervals: A brief review. *The Statistician*, *44*(4), 417–423.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation of a binomial proportion. *Statistical Science 16*(2), 101–133.

Brown, L. D., Cai, T. T., & DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, *30*(1), 160–210.

Clopper, C., & Pearson, E. S. (1934). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, *18*, 379–390.

Fay, M. P. (2010). Two-sided exact tests and matching confidence intervals for discrete data. *The R Journal 2*, 53–58. (See also http://journal.rproject.org/archive/2010-1/RJournal_2010-1_Fay.pdf)

Hirji, K. F. (2006). *Exact analysis of discrete data*. Boca Raton, FL: Taylor and Francis.

Hwang, J. T. G., & Yang, M.-C. (2001). An optimality theory for mid p-values in 2×2 contingency tables. *Statistica Sinica*, *11*, 807–826.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (3rd ed.). New York: Wiley.

Jowett, D. H. (1963). The relationship between the binomial and F distributions. *The Statistician*, *3*(1), 55–57.

Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, *18*, 379–390.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, *17*(8), 857–872.

Salehi, M. M., & Seber, G. A. F. (2001). A new proof of Murthy's estimator which applies to sequential sampling. *Australian and New Zealand Journal of Statistics*, *43*(3), 901–906.

http://www.springer.com/978-3-642-39040-1

Statistical Models for Proportions and Probabilities
Seber, G.A.F.
2013, VIII, 69 p., Softcover
ISBN: 978-3-642-39040-1