STAT 210
Applied Statistics and Data Analysis
Comparing Proportions

Joaquin Ortega

Fall 2020

Comparing Proportions

Comparing proportions

Frequently, we are interested in proportions for statistical analysis: the proportion of people who smoke, the proportion of individuals with a given trait in a population, the proportion of defective items produced in a factory, or the proportion of birds that are recaptured after marking.

The analysis of proportions is linked to the binomial distribution and its approximation by the normal distribution.

The binomial distribution

The binomial distribution

Suppose we draw an independent sample of size n with replacement from a population and we are interested in estimating the proportion p of individuals that have a certain characteristic. Let's say that an individual is of type A if it has the characteristic in question.

The distribution of the number of individuals of type A, n_A, in the sample is binomial with parameters n and p:

$$P(n_A = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The expected value and variance for this distribution are given by

$$E(n_A) = np, \qquad Var(n_A) = np(1-p).$$

The binomial distribution

A natural estimator for the (unknown) proportion p is the observed proportion of individuals of type A in the sample:

$$\pi = \frac{n_A}{n}.$$

The expected value of π is

$$E(\pi) = E\left(\frac{n_A}{n}\right) = \frac{1}{n}E(n_A) = p$$

We say that π is an *unbiased* estimator for p. Its variance is

$$Var(\pi) = \frac{1}{n^2}Var(n_A) = \frac{p(1-p)}{n}.$$

The normal approximation

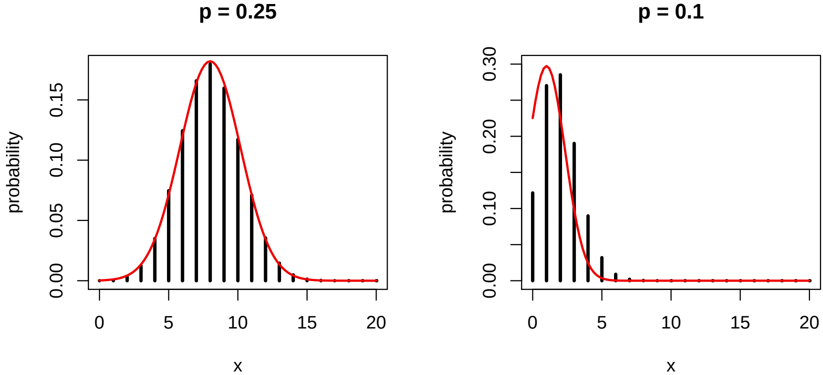
By the Central Limit Theorem, for n large, the binomial distribution can be approximated by a normal distribution.

Rule of Thumb
If n and p are such that $np \geq 5$ and $n(1-p) \geq 5$ the binomial distribution can be approximated by the normal distribution.

The normal approximation

```
par(mfrow=c(1,2))
plot(0:20, dbinom(0:20,20,0.4), type='h', lwd=3,
     ylab='probability',xlab='x',main='p=0.25')
curve(dnorm(x,8,sqrt(20*0.4*0.6)),0,20, col='red2',
      add = TRUE, lwd=2)
plot(0:20,dbinom(0:20,20,0.1), type='h', lwd=3,
     ylim = c(0,0.3),ylab='probability',xlab='x',
     main='p=0.1')
curve(dnorm(x,1,sqrt(20*0.1*0.9)),0,20, col='red2',
      add = TRUE, lwd=2)
```

The normal approximation



$$np = 8, n(1-p) = 12; \qquad np = 2, n(1-p) = 18$$

The normal approximation

Thus, if $np \geq 5$ and $n(1-p) \geq 5$ the sampling density for the (sample) proportion π can be approximated by a normal distribution with parameters

$$p \qquad \text{and} \qquad \frac{p(1-p)}{n}.$$

The closer p is to $1/2$, the faster the convergence to the normal distribution.

A continuity correction due to Yates, can improve the approximation by the normal distribution.

Example

The following data come from Kaye, D.H., *Statistical Evidence of Discrimination*, JASA (1982).

In a case about discrimination against blacks in grand jury selection in Alabama, the plaintiff argued that of the 1050 individuals called to serve as jurors, only 177 were black.

At the time, 25% of those eligible to serve were blacks.

Do the data support the claim of discrimination?

Example

We want to test

$$H_0 : p = 0.25 \quad vs \quad p < 0.25$$

and choose a level $\alpha = 0.01$.

We have

- $n = 1050$,
- $n_A = 177$,
- $\pi = 177/1050 \approx 0.169$,
- $n \times p = 1050 \times 0.25 = 262.5 > 5$ and
- $n \times (1 - p) = 1050 \times 0.55 = 787.5 > 5$.

Example

We calculate the *p*-value using the normal approximation.

```
n <- 1050 ; n.A <- 177
p_0 <- 0.25 ; alpha <- 0.01
SE <- sqrt(p_0 * (1 - p_0) / n)
(pi <- 177/1050)
```

```
## [1] 0.1685714
```

```
pnorm(pi,p_0,SE)
```

```
## [1] 5.521473e-10
```

Example

```r
prop.test(n.A,n,p_0)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  n.A out of n, null probability p_0
## X-squared = 36.698, df = 1, p-value = 1.379e-09
## alternative hypothesis: true p is not equal to 0.25
## 95 percent confidence interval:
##  0.1466952 0.1929145
## sample estimates:
##         p
## 0.1685714
```

Example

```r
binom.test(n.A,n,p_0)
```

```
##
##  Exact binomial test
##
## data:  n.A and n
## number of successes = 177, number of trials = 1050, p-va
## alternative hypothesis: true probability of success is n
## 95 percent confidence interval:
##  0.1464049 0.1926129
## sample estimates:
## probability of success
##              0.1685714
```

Two independent proportions

Two independent proportions

Assume now that we have two samples of sizes n_1 and n_2, respectively, with number of successes m_1 and m_2. The corresponding proportions are

$$\pi_i = \frac{m_i}{n_i}$$

for $i = 1, 2$, and we want to compare these two values.

Their difference $d = \pi_1 - \pi_2$ is approximately normally distributed with mean 0 and variance

$$Var_p(d) = p(1-p)\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)$$

as long as both samples have the same probability of success p and the sample sizes are large.

Two independent proportions

To test the hypothesis that $\pi_1 = \pi_2$ use the normal approximation with the pooled estimate for the proportion

$$\pi = \frac{m_1 + m_2}{n_1 + n_2}$$

in place of p in the formula for the variance.

The test statistic

$$u = \frac{d}{\sqrt{Var_\pi(d)}}$$

has approximately a standard normal distribution.

There is also a continuity correction in this case.

Two independent proportions

The normal approximation requires

$$\begin{aligned} n_i \times \pi_i &\geq 5 \\ n_i \times (1 - \pi_i) &\geq 5 \end{aligned}$$

for $i = 1, 2$.

The test can be carried out using `prop.test`.

Example

The following matrix corresponds to the number of patients involved in car accidents that survived or died. The use of seat belts is also reported in the data, which were registered at a hospital in North Carolina.

```
car.accidents <- data.frame(survived = c(1781,1443),
                            died=c(135,47))
rownames(car.accidents) <- c('nsb','sb')
car.accidents
```

```
##     survived died
## nsb     1781  135
## sb      1443   47
```

Example

To test whether the use of seat belts affected the rates of survival we compare the proportions using the function prop.test

```
prop.test(as.matrix(car.accidents))
```

```
##
##  2-sample test for equality of proportions with continui
##
## data:  as.matrix(car.accidents)
## X-squared = 24.333, df = 1, p-value = 8.105e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.05400606 -0.02382527
## sample estimates:
##    prop 1    prop 2
## 0.9295407 0.9684564
```

Example

Another syntax:

```
prop.test(c(1781,1443),c(1916,1490))
```

```
##
##  2-sample test for equality of proportions with continu
##
## data:  c(1781, 1443) out of c(1916, 1490)
## X-squared = 24.333, df = 1, p-value = 8.105e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.05400606 -0.02382527
## sample estimates:
##    prop 1    prop 2
## 0.9295407 0.9684564
```