ST 516: Foundations of Data Analytics

Proportions

Proportions

- The Central Limit Theorem
- Exact Test
- Example

Special Case of the CLT

There is a special case of the CLT that deserves our attention, and that is when we obtain observations that fall into one of two categories (for example, success/failure, presence/absence, black/white).

- For the purpose of data analysis, we use a binary (or Bernoulli) random variable to represent data like this—each random variable takes the value 1 (for example, for a success, presence or black) or the value 0 (for a failure, absence, white).

- Then, when we take the mean of a sample of these random variables, we add up a bunch of 0s and 1s and divide by n, the sample size.

- This sample mean is the sample proportion, and we use it to make inference about the population proportion.

Binary Data

Suppose that $Y_1, Y_2, \ldots, Y_n$ are an independent sample of binary random variables, and let $\pi = \mathsf{P}(Y_i = 1)$ for $i = 1, 2, \ldots, n$.

Important note: Here, π represents an unknown population probability, not the real number $3.14159\ldots$.

You've already seen that the mean and variance of a binary random variable, Y, are: $E[Y] = \pi$ and $Var(Y) = \pi(1-\pi)$.

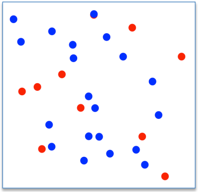
Therefore, using our independent sample if we calculate the sample prorportion,

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

then we have

$$\mathsf{E}[\hat{\pi}] = \pi \text{ and } \mathsf{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$$

Special Case of the CLT

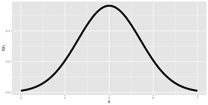


$$\longrightarrow \hat{\pi} \longleftarrow$$

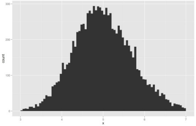
$$\downarrow$$

Inference about
population parameter

Sampling Distribution



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



Example

Suppose that an independent sample of $n = 57$ pine trees are evaluated for the presence or absence of pathogen that is very damaging to the tree. Of these 57 trees, 22 are found to have the pathogen. What is an estimate of the proportion of pine trees with this pathogen in the underlying population of pine trees from which this sample was obtained?

- First, $\hat{\pi} = 22/57 = 0.386$.
- Next, we calculate the standard error of $\hat{\pi}$:

$$SE_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = \sqrt{\frac{0.39(0.61)}{57}} = 0.064$$

Example (cont'd)

- Finally, we construct a 95% confidence interval:

$$0.386 \pm 1.96(0.064) = (0.261, 0.511)$$

Therefore, a plausible range of values for the underlying true population proportion runs from 0.261 to 0.511.

There is a function in R that will facilitate our confidence interval construction in this setting, and we'll take a look at that next.

A Hypothesis Test

Suppose we want to perform a test of the hypotheses

$$
\begin{aligned}
H_0: &\qquad \pi = \pi_0 \\
H_A: &\qquad \pi \neq \pi_0
\end{aligned}
$$

for some known quantity, π_0. Then the quantity

$$Z = \frac{\hat{\pi} - \pi_0}{SE_{\hat{\pi}}} \dot{\sim} N(0,1)$$

For constructing a confidence interval, we used

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

A Hypothesis Test

For a hypothesis test, we use a slightly different standard error:

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

Strictly speaking, we could use the other standard error for the hypothesis test also, but it turns out that when we do the test does not perform as we expect it to when $\hat{\pi}$ is close to zero or one.

- The test that uses $\hat{\pi}$ in the standard error calculation is called the Wald test. The test that uses π_0 in the standard error calculation is called the score test.

- If you use the different calculations for a test and a confidence interval, then there are some situations in which the two may not agree.

Using R

You can use R to make inference about a population proportion by using the function `prop.test()`:

1. The confidence interval is constructed by inverting the score test. This means that the confidence interval we get from R and the one we calculated above will not be precisely the same.

2. By default, the confidence interval is constructed using a "continuity correction." In cases where the continuity correction might make a difference, however, we instead recommend using what's known as an exact test.

Exact Test

The procedures we just discussed rely on the sample size being large (i.e., for the CLT to hold), so now we'll turn to a procedure for small sample sizes.

Remember that the number of successes in n independent trials can be described using a Binomial distribution, where

$$Pr(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

where π is the true probability of success, and $k = 0, 1, \ldots, n$.

A p-value for an exact test is calculated by adding up the probabilities for those values of k that result in estimates of π that are at least as far from the null hypothesized value as is $\hat{\pi}$.

Example

Suppose that in an independent sample of $n = 10$ pine trees, $x = 2$ of them have a pathogen present. If we want to test the hypotheses

$$\begin{aligned} H_0: &\qquad \pi = 0.5 \\ H_A: &\qquad \pi \neq 0.5 \end{aligned}$$

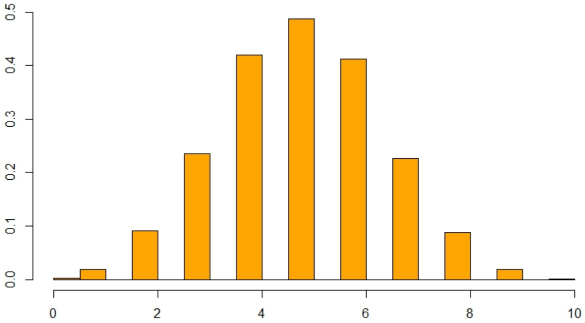
then we need to calculate the probability, assuming that $\pi = 0.5$ (i.e., under the null hypothesis):

$$\mathsf{P}(X = 0, 1, 2, 8, 9, \text{ or } 10).$$

And, we can do this *exactly* using the Biomial distribution.

In R, we can use the function `binom.test()`.

Example, continued



```
pval <- pbinom(2,10,0.5) + (1-pbinom(7,10,0.5))
```

Deciding Between Tests

Some authors use a “success-failure condition” to decide whether to use the proportion test based on the Central Limit Theorem or the exact test.

- This is essentially a condition about the size of the sample
- If $n\pi \geq 10$ and $n(1-\pi) \geq 10$, the Normal approximation works quite well
- Of course we don’t know π, but you should recognize that if π is close to zero or close to one, we will require a larger sample size for the Normal approximation to work well