Proportions and Normality

Grinnell College

March 31, 2025

Warm up

- What is the relationship between the normal and t-distribution?
- What distributional parameters does the normal distribution have? What about the t?
- How is the t-distribution used in the construction of confidence intervals?

Sample Means and Proportions

There is an interesting relationship between means and proportions

For example, consider taking a coin and flipping it 10 times. How many heads would you expect to see?

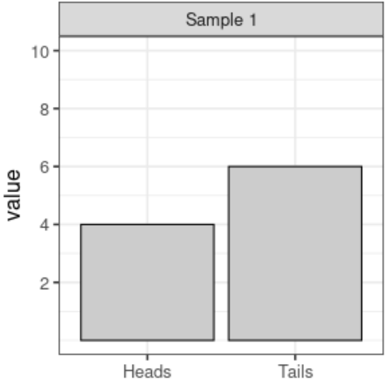
$$\mathcal{S} = \{H, H, T, T, H, T, H, T, T, T\}$$
$$X = \{1, 1, 0, 0, 1, 0, 1, 0, 0, 0\}$$

We can find the *proportion* of heads from our sample $\mathcal{S}$ by simply taking the total number of heads and dividing by the total number of flips, giving

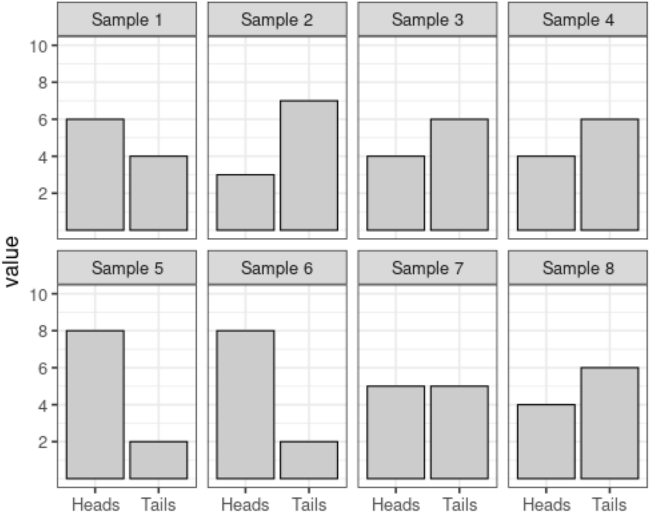
$$\hat{p} = \frac{4}{10}$$

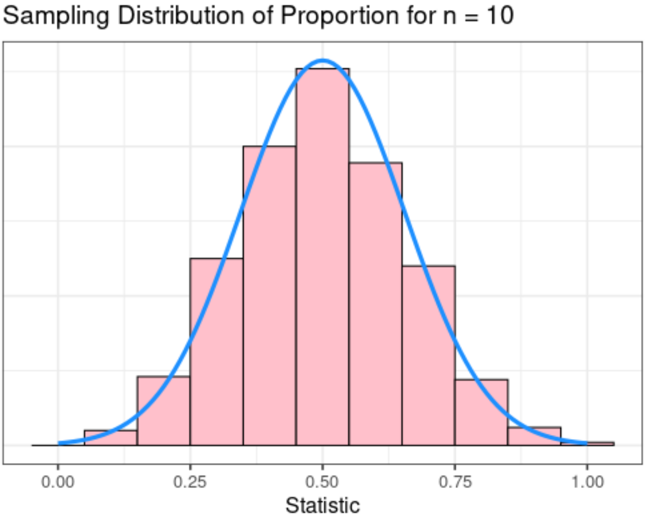
However, if we consider X, which defines H as 1 and T as 0, we can also find the sample mean:

$$\begin{aligned}\overline{x} &= \frac{1}{10}\sum_{i=1}^{n} x_i \\ &= 0.4\end{aligned}$$



Repeated Samples for $n = 10$





Sampling Distribution of Proportion for n = 10

Central Limit Theorem

For a sample with one proportion, the sampling distribution of our proportion statistic, $\hat{p}$ is approximately

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

There a few rules of thumb relating to the size and the proportion:

1. $n \times p \geq 10$
2. $n \times (1-p) \geq 10$

In particular, it is often difficult to estimate proportions precisely that are near the boundaries (0 and 1)

Example

In a study conducted by Johns Hopkins University researchers investigated the survival of babies born prematurely. They searched their hospital's medical records and found 39 babies born at 25 weeks gestation (15 weeks early), 31 of these babies went on to survive at least 6 months. With your group:

1. Use a normal approximation to construct a 80% confidence interval estimate for the true proportions of babies born at 25 weeks gestation that are expected to survive
2. An article on Wikipedia suggests that 70% of babies born at a gestation period of 25 weeks survive. Is this claim consistent with the Johns Hopkins study?

Example – Critical Values

As we are looking for an 90% confidence interval with $n = 39$, we need to use the `qt()` with $df = n - 1 = 38$

```
> qt(c(0.05, 0.95), df = 38)
[1] -1.686  1.686
```

We can also find our sample mean and standard error:

$$\hat{p} = \frac{31}{39} = 0.795$$

$$SE = \sqrt{\frac{0.795(1 - 0.795)}{39}} = 0.065$$

Example – Confidence Interval

1. Together, we find an 80% confidence interval of

$$0.795 \pm 1.686 \times 0.065 = (0.685, 0.905)$$

2. As 0.7 is contained within our constructed 90% CI, it is consistent with the results of the study by Johns Hopkins

Example – t-statistic

Note that in this problem we began with our data, collected from John's Hopkins, as well as a claim from Wikipedia. As such, we have everything we need to create a t-statistic showing us how many "standard deviations" our observed data is from our hypothesis:

$$t = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{0.795 - 0.7}{0.065} = 1.4615$$

We know from last week that this follows a t-distribution with $df = 38$, the 90% critical values of which are ± 1.686. Since our observed data falls within the critical values of this distribution, we can again say that our observed data is consistent with the hypothesis $p_0 = 0.7$

Differences in Proportion

A useful extension to the distribution of a sample proportion is the *difference in proportions* which, according to the CLT, also follows an approximately normal distribution:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2,\ \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

The t-statistic we use implicitly assumes that the true difference is equal to zero:

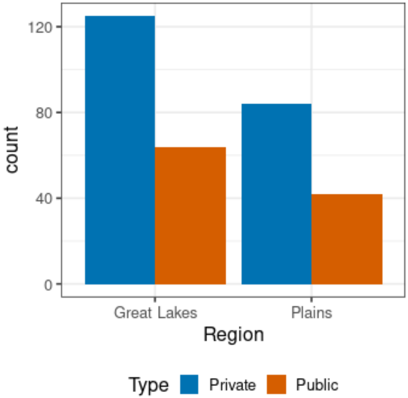
$$t = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

which follows a t-distribution with $df = n_1 + n_2 - 2$ degrees of freedom

Difference in Proportions

Suppose we are interested in determining if the composition of public and private schools is the same between the Plains region and the Great Lakes

	Private	Public
Great Lakes	125	64
Plains	84	42



Difference in Proportions

	Private	Public	Sum
Great Lakes	125	64	189
Plains	84	42	126

- $\hat{p}_1 = 0.661$
- $\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} = 0.0011$
- $\hat{p}_2 = 0.666$
- $\frac{\hat{p}_2(1-\hat{p}_2)}{n_2} = 0.0017$

Using $C = 1.649$ as our critical value for $df = 313$, we find a 90% CI of

$$\hat{p} \pm C \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = (-0.094, 0.084)$$

Computing t-statistics

	Private	Public	Sum
Great Lakes	125	64	189
Plains	84	42	126

- $\hat{p}_1 = 0.661$
- $\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} = 0.0011$
- $\hat{p}_2 = 0.666$
- $\frac{\hat{p}_2(1-\hat{p}_2)}{n_2} = 0.0017$

Together, this gives us a t-statistic of

$$t = \frac{0.661 - 0.666}{\sqrt{0.0011 + 0.0017}} = -0.09$$

Indicating that our observed data is very near what we would expect if these proportions were truly equal

Key Takeaways

- Proportions share the same properties as the mean
- CLT for proportion and difference of proportion
- Confidence intervals and test statistics computed the same way