**STA 371G, Statistics and Modeling**

# Sampling Distribution of A Sample Proportion

*Professor: Mingyuan Zhou*



**Optional text**: *Text in this color is optional to read.*

## 4.1 Sampling Distribution of A Sample Proportion

In our previous lecture note, we have discussed that if we survey $n$ people randomly selected from a large population with a "Yes/No" question, where the population size is considerably larger than $n$ and the true population proportion of answering "Yes" is $p$, then the number of "Yes" from a random sample of $n$ people can be considered as a binomial random variable

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

The mean and variance of the binomial random variable $X \sim \text{Binomial}(n, p)$ can be expressed as

$$\mathbb{E}[X] = np, \ \text{Var}[X] = np(1-p).$$

When the sample size $n$ is not too small and neither $p$ nor $1-p$ are too close to 0, then we may approximate $X \sim \text{Binomial}(n, p)$ with a normal random variable as

$$X \sim \mathcal{N}\big(np, np(1-p)\big).$$

Thus if the population proportion $p$ is known, then for a random sample of size $n$, we are 95% confident that

$$P(np - 2\sqrt{np(1-p)} < X < np + 2\sqrt{np(1-p)}).$$

For example, it is known that 13% of 4,438 Texas BBA students in Fall 2013 are non-Texas residents. If we randomly ask $n = 50$ Texas BBA students, since $P(1.7 < X < 11.3) \approx 0.95$, we are 95% confident that the number of non-Texas residents among them is between 2 to 11.

In practice, we usually don't know the population proportion $p$ and we are trying to infer its "True" value based on a random sample of size $n$. For example, if we randomly survey $n = 50$ Texas BBA students and we find 40 Texas residents, a point estimate of the true proportion of Texas residents would be $\hat{p} = 40/50 = 80\%$. But how accurate would this point estimate be? Below we intend to answer this question.

Suppose the number of "Yes" among the sample of size $n$ is $X$, then we define the sample proportion as

$$\hat{p} = \frac{X}{n}.$$

If the sample size $n$ is the same as the population size, then it is clear $\hat{p} = p$. However, the sample size is usually much smaller than the population size, and we have to find an approach to measure how accurate the sample proportion $\hat{p} = \frac{X}{n}$ is as an estimate of the true population proportion $p$.

Since if $X \sim \mathcal{N}\big(np, np(1-p)\big)$, then the sample proportion $\hat{p} = \frac{X}{n}$ is still normal distributed with

$$\mathbb{E}[\hat{p}] = \frac{\mathbb{E}[X]}{n} = p, \ \mathrm{Var}[\hat{p}] = \frac{\mathrm{Var}[X]}{n^2} = \frac{p(1-p)}{n}.$$

We usually refer $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ as the sampling distribution of the sample proportion. As $p$ is unknown, we further substitute $p$ with $\hat{p}$ to calculate the variance as $\frac{\hat{p}(1-\hat{p})}{n}$, leading to the following approximation:

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

Thus we are 95% confident that

$$\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

For example, if we randomly survey $n = 50$ Texas BBA students and find $X = 40$ Texas residents, then $\hat{p} = 40/50 = 0.80$ and we are 95% confident that

$$\left(68.7\% = 0.80 - 2\sqrt{\frac{0.80 \times (1-0.80)}{50}}\right) < p < \left(0.80 - 2\sqrt{\frac{0.80 \times (1-0.80)}{50}} = 91.3\%\right).$$

Understanding the sampling distribution of a sample proportion would be very useful for us to read surveys/polls. For example, a national poll conducted by Anzalone Liszt Grove Research shows that: "Americans oppose leaving the NSA's current surveillance programs in place as is by a 27-point margin (32% support / 59% oppose)... The poll of N=803 adults was conducted... The margin of error for the poll is plus or minus 3.5 percentage points at the 95% level of confidence."

After reading this note, given the number of samples, one shall now be able to calculate the margin of error of the sample proportion at the 95% level of confidence. For example, at the 95% level of confidence, if we increase the sample size from $n = 100$ to $n = 1000$, the error of the poll decreases from around $(-10\%, 10\%)$ to $(-3.2\%, 3.2\%)$, and if we further increase the sample size from $n = 1000$ to $n = 10,000$, the error of the poll decreases from around $(-3.2\%, 3.2\%)$ to $(-1.0\%, 1.0\%)$ (decreasing very slowly!).

Read this document: Americans Strongly Support Reining in NSA Surveillance Program