Oct 29, 2009

We discuss the distribution of a sample proportion, $\hat{p}$, and how we can tie this information back in with the normal distribution.

Distribution of $\hat{p}$

The sample statistic $\hat{p}$ tends to approximate the true proportion, p. That is, $\hat{p}$ provides a *point estimate* of p. We also know – both intuitively and we could even prove it mathematically – that $\hat{p}$ tends to be closer to p in a big sample relative to a small sample. These properties provide qualitative descriptions of the center and variation of the statistic $\hat{p}$.

We also should describe the center and variation quantitatively. If we were able to sample many $\hat{p}$ (many samples and the proportion of each sample), we would find the mean of these sample proportions is the true proportion, p. The standard deviation of these sample proportions, if each had a sample size of n, would be

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

One problem: we rarely know p, which means we cannot compute $\sigma_{\hat{p}}$. However, $\hat{p}$ is about equal to p, so we sub $\hat{p}$ in to estimate the standard deviation:

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The "hat" on $\sigma_{\hat{p}}$ is just a reminder that this is only an estimate.

Thus far, we have quantified the mean and standard deviation of $\hat{p}$. Next, we describe its distribution. If np and $n(1-p)$ are each at least 10, then we have that $\hat{p}$ is approximately normally distributed. Again, we usually do not know p, so we can use $\hat{p}$ as an estimate to check this condition and for normality: check if $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.

If $n\hat{p}$ and $n(1-\hat{p})$ are both at least 10, then $\hat{p}$ is (roughly) from

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

As already mentioned, p is usually unknown, thus we estimate the uncertainty (variation) associated with $\hat{p}$ as

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Describing how close p is to $\hat{p}$

Suppose X is a normal random variable with mean μ and standard deviation σ. How often is X within 1 standard deviation of μ? Within 2 standard deviations?

About 45% of the students at UCLA are men. Suppose we randomly sample 35 students and look at the proportion who are men, $\hat{p}_{men}$. Can we use the normal model?

What is the mean and standard deviation of $\hat{p}_{men}$?

About how often will our estimated proportion from our sample, $\hat{p}_{men}$ be within 2 standard deviations of the mean?

U of Minnesota

Suppose you attend a required class at U of MN and find that 52% of the 47 people in the class are women. Since this is a required class for all students, we will suppose it is as good as a random sample of the students. We would like to provide not only the **point estimate** of the proportion of students at U of MN who are women but we would like to provide a confidence interval that we think will capture the true proportion.

Our point estimate is $\hat{p}_{women} = 0.52$. What is the mean and standard deviation of $\hat{p}_{women}$? Also, what distribution would our sample proportion follow? (Verify any assumptions necessary to answer this second question.)

If we want to be 95% sure that our interval captures the true proportion of women at U of MN, p_{women}, how can we do this?

Example

We will investigate this with Skittles. We sampled 315 Skittles and found 72 green ones. We computed our point estimate of the proportion of green Skittles, $\hat{p} = 0.23$, and then created a 95% confidence interval for the true proportion of Skittles that are green:

$$\begin{aligned}(\hat{p} - 1.96 * \sigma_{\hat{p}}, \hat{p} + 1.96 * \sigma_{\hat{p}}) \\ (0.229 - 0.046, 0.229 - 0.046) \\ (0.183, 0.275)\end{aligned}$$

We say we are 95% *confident* that the true proportion of Skittles that are green is between 0.183 and 0.275.