Ch. 18 - Sampling Distributions

Expanded def'n: A parameter is: - a numerical value describing some aspect of a pop'n
- usually regarded as constant
- usually unknown

A statistic is: - a numerical value describing some aspect of a sample
- regarded as random before sample is selected
- observed after sample is selected

The observed value depends on the particular sample selected from the population; typically, it varies from sample to sample. This variability is called sampling variability. The distribution of all the values of a statistic is called its sampling distribution.

Def'n: $\hat{p}$ = proportion of ppl with a specific characteristic in a random sample of size n
p = population proportion of ppl with a specific characteristic

The estimate of the standard deviation of a sampling distribution is called a standard error.

General Properties of the Sampling Distribution of $\hat{p}$:

Let $\hat{p}$ and p be as above. Also, $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$ are the mean and standard deviation for the distribution of $\hat{p}$. Then the following rules hold:

Rule 1: $\mu_{\hat{p}} = p$. (Textbook uses $\mu(\hat{p})$)

Rule 2: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}}$. (standard error → $\hat{\sigma}_{\hat{p}}$)

Ex18.1) Suppose the population proportion is 0.5.

a) What is the standard deviation of $\hat{p}$ for a sample size of 4?

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5(1-0.5)}{4}} = 0.25$$

b) How large must n (sample size) be so that the sample proportion has a standard deviation of at most 0.125?

$$\sqrt{n} = \frac{\sqrt{p(1-p)}}{\sigma_{\hat{p}}} \rightarrow n = \frac{p(1-p)}{\sigma_{\hat{p}}^2} = \frac{0.5(1-0.5)}{0.125^2} = 16$$

Rule 3: When n is large and p is not too near 0 or 1, the sampling distribution of $\hat{p}$ is approximately normal. The farther from $p = 0.5$, the larger n must be for accurate normal approximation of $\hat{p}$. Thus, if np and $n(1-p)$ are both sufficiently large (≥ 15), then it is safe to use a normal approximation.

Further assumptions: the sample should always be random and, if sampling without replacement, the sample should be less than 10% of the population.

Using all 3 rules, the distribution of $\hat{p}$ is approximately normal.

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \stackrel{\cdot}{\sim} N(0,1)$$

Ex18.2) Suppose that the true proportion of people who have heard of Sidney Crosby is 0.87 and that a new sample consists of 158 people.

a) Find the mean and standard deviation of $\hat{p}$.

$$\mu_{\hat{p}} = 0.870 \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.87(1-0.87)}{158}} = 0.0268$$

b) What can you say about the distribution of $\hat{p}$?

$np = 158(0.87) = 137.46$ $\qquad n(1-p) = 158(1-0.87) = 20.54$

Since both values are > 15, the distribution of $\hat{p}$ should be well approximated by a normal curve.

c) What is the probability of getting a sample proportion greater than 0.94?

$$P(\hat{p} > 0.94) \rightarrow P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} > \frac{0.94 - 0.87}{0.0268}\right) = P(Z > 2.62)$$

$$= 1 - P(Z < 2.62) = 1 - 0.9956 = 0.0044$$

Sampling Distribution of Mean

How does the sampling distribution of the sample mean compare with the distribution of a single observation (which comes from a population)?

Ex18.3) An epically gigantic jar contains a large number of balls, each labeled 1, 2, or 3, with the same proportion for each value.

Let Y be the label on a randomly selected ball. Find μ_Y and σ_Y.

y	$P(Y = y)$
1	1/3
2	1/3
3	1/3

$$\mu_Y = \sum y_i P(Y = y_i) = 1(1/3) + 2(1/3) + 3(1/3) = 2$$

$$\sigma_Y^2 = \sum (y - \mu_Y)^2 P(Y = y) = (1-2)^2(1/3) + (2-2)^2(1/3) + (3-2)^2(1/3) = 2/3$$

$$\sigma_Y = \sqrt{\frac{2}{3}}$$

Let $\{Y_1, Y_2\}$ be a random sample of size $n = 2$. Find the sampling distribution of the sample mean $\bar{Y}$. Calculate $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$.

There are 9 possible samples:
1, 1   1, 2   1, 3   2, 1   2, 2   2, 3   3, 1   3, 2   3, 3

$\bar{y}$	1	1.5	2	2.5	3
$P(\bar{Y} = \bar{y})$	1/9	2/9	3/9	2/9	1/9

$$\mu_{\bar{y}} = \sum \bar{y}_i P(\bar{Y} = \bar{y}_i) = 1(1/9) + 1.5(2/9) + 2(3/9) + 2.5(2/9) + 3(1/9) = 2$$

$$\sigma^2_{\bar{y}} = \sum (\bar{y} - \mu_{\bar{y}})^2 P(\bar{Y} = \bar{y}) = (1 - 2)^2(1/9) + (1.5 - 2)^2(2/9) + (2 - 2)^2(3/9) + (2.5 - 2)^2(2/9) + (3 - 2)^2(1/9) = 1/3$$

$$\sigma_{\bar{y}} = \sqrt{\frac{1}{3}}$$

Notice that $\frac{\sigma_Y}{\sqrt{n}} = \frac{\sqrt{2/3}}{\sqrt{2}} = \sqrt{\frac{1}{3}} = \sigma_{\bar{y}}$. This relation speaks to the following properties.

General Properties of the Sampling Distribution of $\bar{y}$ (or $\bar{x}$):
Let $\bar{y}$ denote the mean of the observations in a random sample of size n from a population having mean μ and standard deviation σ. Also, $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$ are the mean and standard deviation for the distribution of $\bar{y}$. Then the following rules hold:

Rule 1: $\mu_{\bar{y}} = \mu$.

Rule 2: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$.

Note also that:
1. The spread of the sampling dist'n of $\bar{y}$ is smaller than the spread of the pop'n dist'n.
2. As n increases, $\sigma_{\bar{y}}$ decreases.

Ex18.4) Suppose the population standard deviation is 10.
a) What is the std. dev. of the sample mean for each of the following sample sizes? n = 1, 2, 4, 9, 16, 25, 100

$\sigma_{\bar{y}} = 10/\sqrt{1} = 10$ $\sigma_{\bar{y}} = 10/\sqrt{2} = 7.071$ … $\sigma_{\bar{y}} = 10/\sqrt{100} = 1$

b) How large must n (sample size) be so that the sample mean has a standard deviation of at most 2?

$$\sqrt{n} = \frac{\sigma}{\sigma_{\bar{y}}} \rightarrow n = \left(\frac{\sigma}{\sigma_{\bar{y}}}\right)^2 = \left(\frac{10}{2}\right)^2 = 25$$

Rule 3: When the population distribution is normal, the sampling distribution of $\bar{y}$ is also normal for any sample size n.

Combining the 3 rules, if the population distribution is $N(\mu, \sigma)$, then $\bar{Y}$ is $N(\mu, \sigma/\sqrt{n})$.

Rule 4 (**Central Limit Theorem**): When n is sufficiently large, the sampling distribution of $\bar{y}$ is well approximated by a normal curve, even when the population distribution is not itself normal. The Central Limit Theorem can safely be applied if n exceeds 30.

Using all 4 rules, if n is large and/or the population is normal, then the sampling distribution of $\bar{Y}$ is approximately normal.

$$Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \,\dot{\sim}\, N(0,1)$$

Ex18.5) Suppose the mean length of all episodes of a (formerly) hilarious series is 20.834 minutes, whereas the standard deviation is 0.593 minutes. Let $\bar{Y}$ be the average length for a random sample of 100 episodes.

a) Find the mean and standard deviation of $\bar{Y}$.

$\mu_{\bar{y}} = 20.834$ min $\qquad \sigma_{\bar{y}} = 0.593/\sqrt{100} = 0.0593$

b) What can you say about the distribution of $\bar{Y}$?

Since n is fairly large, the distribution of $\bar{Y}$ should be well approximated by a normal curve.

c) What is the probability of getting a sample mean between 20.7 and 21 minutes?

$$P(20.7 \le \bar{Y} \le 21) \rightarrow P\left(\frac{20.7 - 20.834}{0.0593} \le \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \le \frac{21 - 20.834}{0.0593}\right) = P(-2.26 \le Z \le 2.80)$$

$= P(Z \le 2.80) - P(Z \le -2.26) = 0.9974 - 0.0119 = 0.9855$

d) Can you find $P(20.7 \le Y \le 21)$, where Y is the length of a single randomly selected episode? How would this value compare with the one in part c)?

No, we can't find it, unless we knew Y is normal (not good to just assume). If you were told it was normal, though,

$$P(20.7 \le Y \le 21) \rightarrow P\left(\frac{20.7 - 20.834}{0.593} \le \frac{Y - \mu}{\sigma} \le \frac{21 - 20.834}{0.593}\right) = P(-0.23 \le Z \le 0.28)$$

$= P(Z \le 0.28) - P(Z \le -0.23) = 0.6103 - 0.4090 = 0.2013$

$\rightarrow$ Considering individual point, not the average.