Wednesday, Feb 8

Sampling Distributions of $\bar{x}$ and $\hat{p}$

1. The mean of $\bar{x}$ equals μ_x (i.e., $\mu_{\bar{x}} = \mu_x$). The standard deviation of $\bar{x}$ is $\sigma_x/\sqrt{n}$ (i.e., $\sigma_{\bar{x}} = \sigma_x/\sqrt{n}$).
2. The mean of $\hat{p}$ equals p (i.e., $\mu_{\hat{p}} = p$). The standard deviation of $\hat{p}$ is $\sqrt{p(1-p)/n}$ (i.e., $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$).

Proportions are Means

Suppose we have a population distribution for successes and failures, but we define a random variable x so that $x = 1$ if we observe a success, and $x = 0$ if we observe a failure.

x	$P(x)$
1	p
0	$1-p$

The mean of x is

$$\mu_x = \sum xP(x) = 1 \times p + 0 \times (1-p) = p,$$

and the standard deviation of x is

$$\sigma_x = \sqrt{\sum (x-\mu)^2 P(x)} = \sqrt{(1-p)^2 \times p + (0-p)^2 \times (1-p)} = \sqrt{p(1-p)}.$$

The mean of a sample of observations of x (i.e., $\bar{x}$) *is a proportion*. For example, if our observations of x are 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, then the mean is

$$\bar{x} = \frac{1+1+0+1+0+0+0+1+1+1}{10} = 0.6,$$

which is also the *proportion of observations where we observe a success*. So $\hat{p} = \bar{x}$.

Now applying what we know about the sampling distribution of $\bar{x}$, we have

$$\mu_{\bar{x}} = \mu_x = p,$$
$$\sigma_{\bar{x}} = \sigma_x/\sqrt{n} = \sqrt{p(1-p)}/\sqrt{n} = \sqrt{p(1-p)/n}.$$

Central Limit Theorem

Central Limit Theorem: If $X_1, X_2, \dots, X_n$ are independently and identically distributed random variables so that $E(X_i) = \mu$ and $E(X_i - \mu)^2 = \sigma^2 < \infty$, then

$$\sqrt{n}(\bar{X} - \mu)/\sigma \xrightarrow{d} N(0,1),$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and where $\xrightarrow{d}$ denotes convergence in distribution.

Central Limit Theorem (Layperson's Version): As n increases, the *shape* of the sampling distribution of $\bar{x}$ "approaches" that of a normal distribution.

Example: Suppose we roll $n = 2$ fair 6-sided dice. What is the shape of the sampling distribution of the *mean* number of dots (i.e., $\bar{x}$)?

Note: Because each side has probability 1/6, the probability of each sample is $1/6 \times 1/6 = 1/36$.

Table 1: Population Distribution

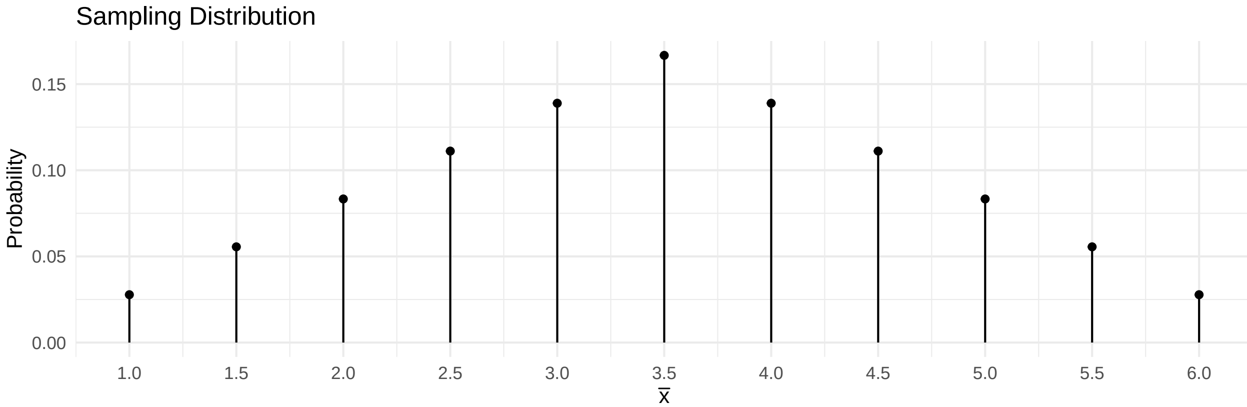
x	$P(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Table 2: Sample Space

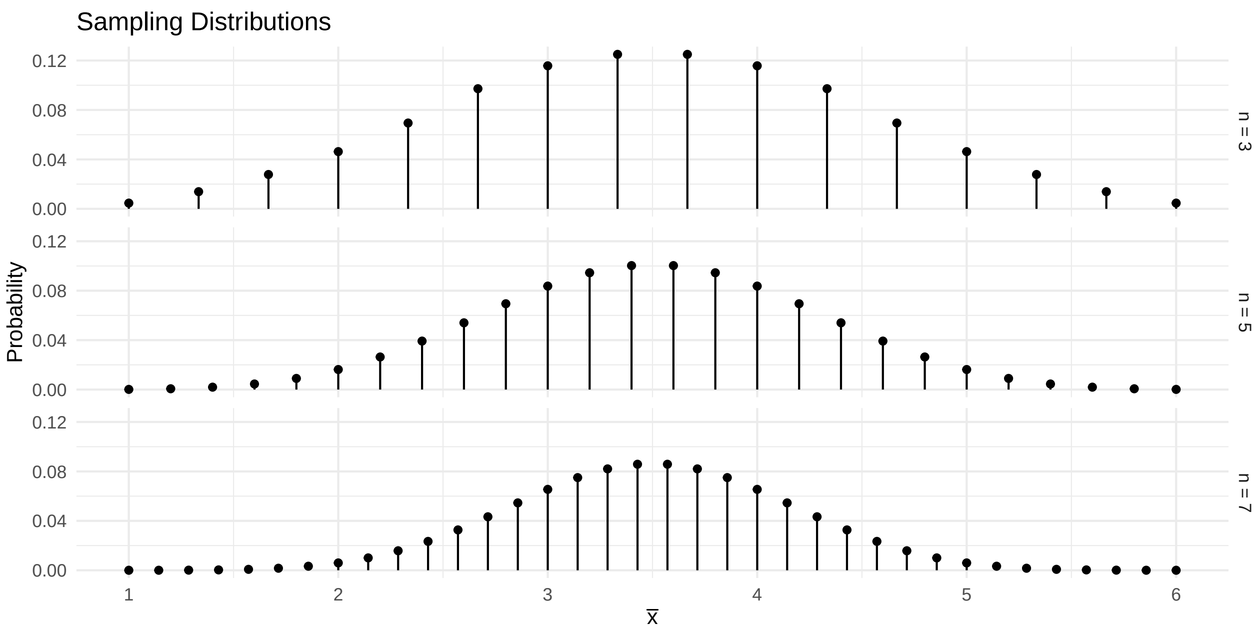
	Second Die					
First Die	1	2	3	4	5	6
1	1.0	1.5	2.0	2.5	3.0	3.5
2	1.5	2.0	2.5	3.0	3.5	4.0
3	2.0	2.5	3.0	3.5	4.0	4.5
4	2.5	3.0	3.5	4.0	4.5	5.0
5	3.0	3.5	4.0	4.5	5.0	5.5
6	3.5	4.0	4.5	5.0	5.5	6.0

Table 3: Sampling Distribution

$\bar{x}$	$P(\bar{x})$
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36



What if we roll $n = 3$, $n = 5$, or $n = 7$ dice?



Here is another demonstration that uses simulation to illustrate the central limit theorem.

The *practical* implication of the central limit theorem is that we can often assume that the shape of the sampling distribution of $\bar{x}$ (or $\hat{p}$) is approximately that of a normal distribution.

Applying the Central Limit Theorem

Recall that the empirical rule states that *approximately 95% of observations are within two standard deviations of the mean.* Adapting this to a *normal probability distribution*, we can say that *there is approximately a probability of 0.95 that the random variable will be within two standard deviations of the mean of the distribution.*

1. The probability that $\bar{x}$ will be between $\mu_x - 2\sigma_x/\sqrt{n}$ and $\mu_x + 2\sigma_x/\sqrt{n}$ is approximately 0.95.
2. The probability that $\hat{p}$ will be between $p - 2\sqrt{p(1-p)/n}$ and $p + 2\sqrt{p(1-p)/n}$ is approximately 0.95.

Example: Recall Darwin's experiment that compared cross-fertilization and self-fertilization.

	Fertilization		
Obs	Cross	Self	Difference
1	23.500	17.375	6.125
2	12.000	20.375	-8.375
3	21.000	20.000	1.000
4	22.000	20.000	2.000
5	19.125	18.375	0.750
⋮	⋮	⋮	⋮
15	12.000	18.000	-6.000

1. Let x be the *difference* in height for a given pair of seedlings. *Assume* that x has a mean of 3 and a standard deviation of 5. So $\mu_x = 3$, but Darwin didn't know that. So he'd use $\bar{x}$ to estimate μ_x. How close might it be? What are the mean and standard deviation of $\bar{x}$ based on a sample of $n = 15$ observations? What is the interval that has a probability of approximately 0.95 of containing $\bar{x}$?

2. Let x be which seedling is taller, with the following *population distribution*.

x	$P(x)$
cross	0.8
self	0.2

   So here the seedling produced by cross-fertilization is more likely to be taller than one produced by self-fertilization. This would be useful to know, but Darwin didn't know the value of p. But he could estimate it based on a sample of observations using $\hat{p}$, the *proportion* of pairs in a sample of observations in which the seedling produced by cross-fertilization is taller. What are the mean and standard deviation of $\hat{p}$ based on a sample of $n = 15$ observations? What is the interval that has a probability of approximately 0.95 of containing $\hat{p}$?

Estimation

Estimation is a kind of inference in which we use a statistic to estimate a parameter.

1. We can use $\bar{x}$ (i.e., the mean of a sample of n observations of x) to estimate the mean of a single observation (i.e., μ_x).

2. We can use $\hat{p}$ (i.e., the proportion of observations in a sample of n observations where we observed a "success") to estimate the probability of a "success" (i.e., p).

Note: Parameters like μ_x and p have a couple of interpretations here. One is that they are properties of the population distribution. But in a survey with a finite number of observations, μ_x is also the mean of *all* observations in the population, and p is a proportion based on *all* observations in the population. This is because in these cases the population distribution is both a probability distribution and also the distribution of all observations in the population.

The *sampling distributions* of $\bar{x}$ and $\hat{p}$ are what we use to determine how effective these statistics are at estimating the parameters μ_x and p, respectively.

1. Both $\bar{x}$ and $\hat{p}$ are **unbiased**, meaning that the mean of $\bar{x}$ equals μ_x, and the mean of $\hat{p}$ equals p.

2. A **standard error** is the standard deviation of a statistic. The standard error of $\bar{x}$ is $\sigma_x/\sqrt{n}$, and the standard error of $\hat{p}$ is $\sqrt{p(1-p)/n}$.

3. The **central limit theorem** implies that (unless n is very small) we can regard the *shape* of the sampling distributions of $\bar{x}$ and $\hat{p}$ as approximately that of a normal distribution for the purpose of computing probabilities concerning $\bar{x}$ or $\hat{p}$.

The above statements require certain technical assumptions about how the data are collected. We will discuss that in a later lecture.