5.3 Statistics and Their Distributions:

When we get a sample of data $x_1, \ldots, x_n$, these data will usually change if we get another sample.

So, before we get the sample, the values are Random Variables $X_1, \ldots, X_n$.

The sample mean is also a RV: $\overline{X} = \dfrac{X_1 + \ldots + X_n}{n}$.

A statistic is any quantity whose value can be obtained from the sample data.

Before the sample is obtained, a statistic is a Random Variable, so we write it with a capital letter.

After the sample is obtained, a lowercase letter represents the particular value of the statistic.

The distribution of a statistic is sometimes called a sampling distribution.

Random Samples: The RVs $X_1, \ldots, X_n$ form a simple random sample (or just random sample) of size n if

1. The X_i's are independent RVs
2. Every X_i has the same probability distribution.

Another way to state conditions 1 and 2 is to say the X_i's are independent and identically distributed or abbreviated as iid.

In practice, we often sample without replacement from a finite population, so X_i are not exactly iid. But as long as we are sampling less than 5% of the population, this is approximately iid.

Deriving a Sampling Distribution:

Example 1: A certain brand of pen is sold in packs of 1, 2, or 4.

At a certain store, 20% of customers choose the 1-pack, 50% choose the 2-pack, and 30% choose the 4-pack. Let X be the number of pens a random customer purchases.

x	1	2	4
$p(x)$	.2	.5	.3

$E[X] = 2.4$, $Var(X) = 1.24$

On a certain day two customers buy pens. Let X_1 and X_2 be the number of pens sold to the two customers. Assume X_1 and X_2 are independent.

Let $\bar{X} = \dfrac{X_1 + X_2}{2}$.

Find the pmf (sampling distribution) for $\bar{X}$. Find $E[\bar{X}]$ and $Var(\bar{X})$.

$p(x_1, x_2)$	x_2: 1	2	4
x_1: 1	0.04	0.1	0.06
2	0.1	0.25	0.15
4	0.06	0.15	0.09

$P(X_1 = 1, X_2 = 1) = P(X_1 = 1)P(X_2 = 1) = 0.2(0.2) = 0.04$

$P(X_1 = 1, X_2 = 2) = 0.2(0.5) = 0.1$

$\vdots$

$\bar{x}$	x_2: 1	2	4
x_1: 1	1	1.5	2.5
2	1.5	2	3
4	2.5	3	4

pmf for $\bar{X}$

$P(\bar{X} = 1) = 0.04$

$P(\bar{X} = 1.5) = 0.1 + 0.1 = 0.2$

$P(\bar{X} = 2) = 0.25$

$P(\bar{X} = 2.5) = 0.06 + 0.06 = 0.12$

$P(\bar{X} = 3) = 0.15 + 0.15 = 0.3$

$P(\bar{X} = 4) = 0.09$

$$p(\bar{x}) = P(\bar{X} = \bar{x}) = \begin{cases} 0.04, & \bar{x} = 1 \\ 0.2, & \bar{x} = 1.5 \\ 0.25, & \bar{x} = 2 \\ 0.12, & \bar{x} = 2.5 \\ 0.3, & \bar{x} = 3 \\ 0.09, & \bar{x} = 4 \end{cases}$$

$$E[\bar{X}] = 1(0.04) + 1.5(0.2) + 2(0.25) + 2.5(0.12) + 3(0.3) + 4(0.09) = \boxed{2.4}$$

$$E[\bar{X}^2] = 1^2(0.04) + 1.5^2(0.2) + 2^2(0.25) + 2.5^2(0.12) + 3^2(0.3) + 4^2(0.09) = 6.38$$

$$Var(\bar{X}) = E[\bar{X}^2] - E[\bar{X}]^2 = 6.38 - 2.4^2 = \boxed{0.62}$$

We can also simulate a sampling distribution instead.

Example in R.

5.4 The Distribution of the Sample Mean

It turns out we can find the mean and variance of the sample mean without finding the whole sampling distribution.

Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and standard deviation σ. Let $\overline{X} = \dfrac{X_1 + \ldots + X_n}{n}$.

Then $E[\overline{X}] = \mu$

$$Var(\overline{X}) = \frac{\sigma^2}{n} \quad \text{and} \quad SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

often called the standard error of the mean

Similarly let $T_0 = X_1 + \ldots + X_n$ (the sample total). Then

$$E[T_0] = n\mu$$

$$Var(T_0) = n\sigma^2 \quad \text{and} \quad SD(T_0) = \sqrt{n} \cdot \sigma.$$

Example 1: A certain brand of battery lasts on average 10 days with a standard deviation of 4 days.

We pick a sample of 5 batteries. What are the mean and SD for the total lifetime of the 5 batteries? What are the mean and SD for the average lifetime of the 5 batteries?

$$E[T_0] = n \cdot \mu = 5 \cdot 10 = 50 \text{ days}$$

$$SD(T_0) = \sqrt{n} \cdot \sigma = \sqrt{5} \cdot 4 \approx 8.9 \text{ days}$$

$$E[\overline{X}] = \mu = 10 \text{ days}$$

$$SD(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{5}} \approx 1.8 \text{ days}$$

Normal Population Case:

If $X_1, \dots, X_n$ are taken from a normal distribution with mean μ and variance σ^2

then for any n, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ and $T_o \sim N(n\mu, n\sigma^2)$.

Example 2: Certain apples are normally distributed with mean 100g and SD 20g.

What is the probability the total mass of 6 random apples is more than 650g?

$T_o \sim N(6 \cdot 100g, 6 \cdot 20^2) = N(600, 2400)$. Want $P(T_o > 650)$.

In R: `1 - pnorm(650, mean = 600, sd = sqrt(2400))` $\approx \boxed{0.154}$

OR

`pnorm(650, mean = 600, sd = sqrt(2400), lower.tail = FALSE)` $\approx \boxed{0.154}$

(lower.tail = FALSE: complement)

Using Table: $P(T_o > 650) = P\left(\frac{T_o - 600}{\sqrt{2400}} > \frac{650-600}{\sqrt{2400}}\right) = \Phi\left(\frac{650-600}{\sqrt{2400}}\right) \approx \boxed{0.154}$

The General Population Case and Central Limit Theorem:

The Central Limit Theorem (CLT):

Let $X_1, \dots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. If n is sufficiently large (say $n > 30$)* then $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$ and $T_o \approx N(n\mu, n\sigma^2)$.

The larger n is, the better this approximation is.

*This depends on how skewed the population is. $n > 30$ covers most practical examples. For a uniform population, $n \geq 12$ is quite good. Very skewed distributions may need $n > 50$ or higher.

Example in R

Example 3: The amount of a particular impurity X in a batch of a certain chemical product is a RV with mean 4.0g and SD 1.5g. If 50 batches are independently prepared, what is the approximate probability that the sample average $\bar{X}$ is between 3.5 and 3.8 g?

$$\bar{X} \approx N\left(4.0, \frac{1.5^2}{50}\right)$$

$P(3.5 \leq \bar{X} \leq 3.8) =$ `pnorm(3.8, mean = 4.0, sd = sqrt(1.5^2/50))`

`- pnorm(3.5, mean = 4.0, sd = sqrt(1.5^2/50))`

The CLT also applies to discrete RVs. In particular it allows us to approximate the Binomial Distribution.

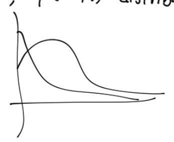
Say we have n independent trials with success probability p.

Let $X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ trial is a success} \\ 0 & \text{if the } i^{th} \text{ trial is a failure} \end{cases}$

The X_i are iid, and $X = X_1 + \ldots + X_n$ has the $Bin(n,p)$ distribution.

So the $Bin(n,p)$ distribution can be approximated by a $N(np, np(1-p))$ distribution.

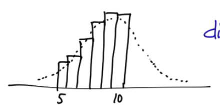
The approximation is good when $\underbrace{np}_{\text{expected \# of successes}} \geq 10$ and $\underbrace{n(1-p)}_{\text{expected \# of failures}} \geq 10$.



One change we need for handling discrete distributions is the

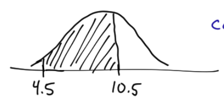
Continuity Correction:

Convert probabilities like $P(5 \leq B \leq 10)$



-0.5 $+0.5$

to $P(4.5 \leq B \leq 10.5)$



Example 4: Suppose at a large university, 30% of students are employed on campus. Let X be the number of students in a random sample of size 50 who work on campus. Approximate

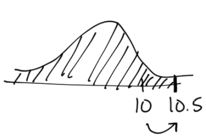
$$P(X \leq 10) \quad \text{and} \quad P(12 \leq X \leq 18).$$

We expect $n \cdot p = 50 \cdot 0.3 = 15 \geq 10$ to work on campus

$n(1-p) = 50 \cdot 0.7 = 35 \geq 10$ to not work on campus.

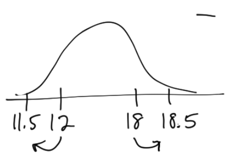
We can use the CLT. $X \approx N(n \cdot p, n \cdot p \cdot (1-p)) = N(15, 10.5)$

$P(X \leq 10) \approx$ pnorm (10.5, mean = 15, sd = sqrt (10.5))



$P(12 \leq X \leq 18) \approx$ pnorm(18.5, mean = 15, sd = sqrt (10.5))

$-$ pnorm (11.5, mean = 15, sd = sqrt (10.5))



5.5 The Distribution of a Linear Combination

The sample mean and sample total are special cases of a linear combination.

If $X_1, X_2, \ldots, X_n$ are RVs and $a_1, a_2, \ldots, a_n$ are constants, then

$a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$ is called a linear combination.

Ex:

$T_o = 1 \cdot X_1 + 1 \cdot X_2 + \ldots + 1 \cdot X_n$, $\quad X_1 + 2X_2 + 3X_3 + \ldots + nX_n$

$\bar{X} = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n} X_1 + \frac{1}{n} X_2 + \ldots + \frac{1}{n} X_n$, $\quad 0 \cdot x_1 + 0 \cdot x_2 + \ldots + 3X_n$

Proposition: Let $X_1, \ldots, X_n$ be RVs with mean values $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$ respectively.

1) Whether or not the X_i's are independent,

$$E[a_1 X_1 + a_2 X_2 + \ldots + a_n X_n] = a_1 E[X_1] + a_2 E[X_2] + \ldots + a_n E[X_n]$$
$$= a_1 \mu_1 + a_2 \mu_2 + \ldots + a_n \mu_n .$$

2) If $X_1, \ldots, X_n$ are independent,

$$Var(a_1 X_1 + a_2 X_2 + \ldots + a_n X_n) = a_1^2 Var(X_1) + a_2^2 Var(X_2) + \ldots + a_n^2 Var(X_n)$$
$$= a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \ldots + a_n^2 \sigma_n^2$$

$$SD(a_1 X_1 + a_2 X_2 + \ldots + a_n X_n) = \sqrt{a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \ldots + a_n^2 \sigma_n^2}$$

3) For any $X_1, \ldots, X_n$

$$Var(a_1 X_1 + \ldots + a_n X_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \, Cov(X_i, X_j) .$$

4) If $X_1, \ldots, X_n$ are normal RVs and are independent, then

$a_1 X_1 + \ldots + a_n X_n$ is also a normal RV.

Example 1: A shipping company accepts 3 sizes of boxes. Let X_1, X_2, X_3 be the number of each type shipped on a randomly selected day. The company knows

$E[X_1] = 100 \qquad E[X_2] = 200 \qquad E[X_3] = 50$

$Var(X_1) = 25 \qquad Var(X_2) = 100 \qquad Var(X_3) = 16$

Find

$$E[X_1 + X_2 + X_3] = E[X_1] + E[X_2] + E[X_3] = 100 + 200 + 50 = \boxed{350}$$

$$E[10X_1 + 30X_2 + 50X_3] = 10 \cdot E[X_1] + 30\, E[X_2] + 50\, E[X_3] = 10 \cdot 100 + 30 \cdot 200 + 50 \cdot 50 = \boxed{9500}$$

If X_1, X_2, X_3 are independent, find

$$Var(X_1 + X_2 + X_3) = Var(X_1) + Var(X_2) + Var(X_3) = 25 + 100 + 16 = \boxed{141}$$

$$Var(2X_1 - X_2) = 2^2 \, Var(X_1) + (-1)^2 \, Var(X_2) = 4\, Var(X_1) + Var(X_2) = 4 \cdot 25 + 100 = \boxed{200}$$