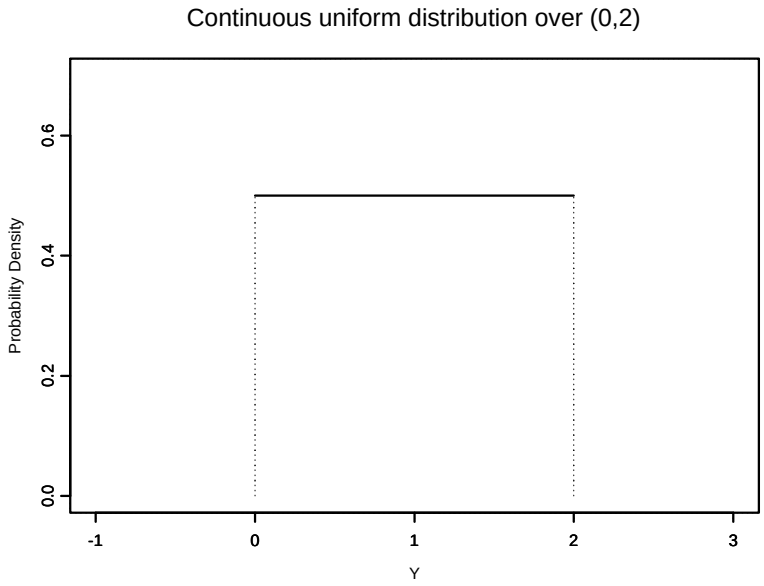
Chapter 5 Class Notes – Sampling Distributions

In the motivating in-class example (see handout), we sampled from the uniform (parent) distribution (over 0 to 2) graphed here.

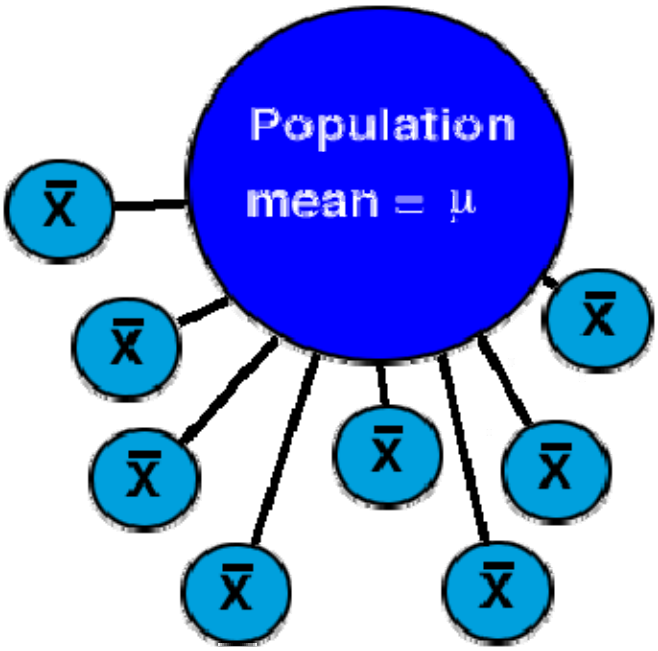


The parent population mean and variance are μ = 1.0 and σ^2 = ⅓ → $\sigma = 1/\sqrt{3}$ = 0.57735. (In this class, we're not interested in how the mean and variance values are obtained).

Now, consider taking all – i.e., an infinite number – samples of size n = 4 from this population. In practice, this is impossible, so we'll settle for taking just B = 10,000 samples of size n = 4. The 10,000 sample means are called "avg04", plotted in the top left graph and summarized in the table below. The mean (1.0029) is ≈ μ = 1.0, the SD (0.2878) is ≈ $\frac{\sigma}{\sqrt{4}}$ = 0.2887, and the shape of the means ("sampling distribution") is approximately Normal (this is quite surprising!).

Next, we shift to B = 10,000 samples from the parent population of size n = 9; the means ("avg09") plotted on the top right have mean

(1.0028) is ≈ μ = 1.0, the SD (0.1925) is ≈ $\frac{\sigma}{\sqrt{9}}$ = 0.19245, and the shape of the means ("sampling distribution") is again approximately Normal. This process continues for "avg16" and "avg25" with SD's of approximately $\frac{\sigma}{\sqrt{n}}$ and with shapes getting closer and closer to the Normal curve. A schematic representation of this meta-experiment is on p.150 and as follows (in our text, we use $\bar{y}$ instead of $\bar{x}$):



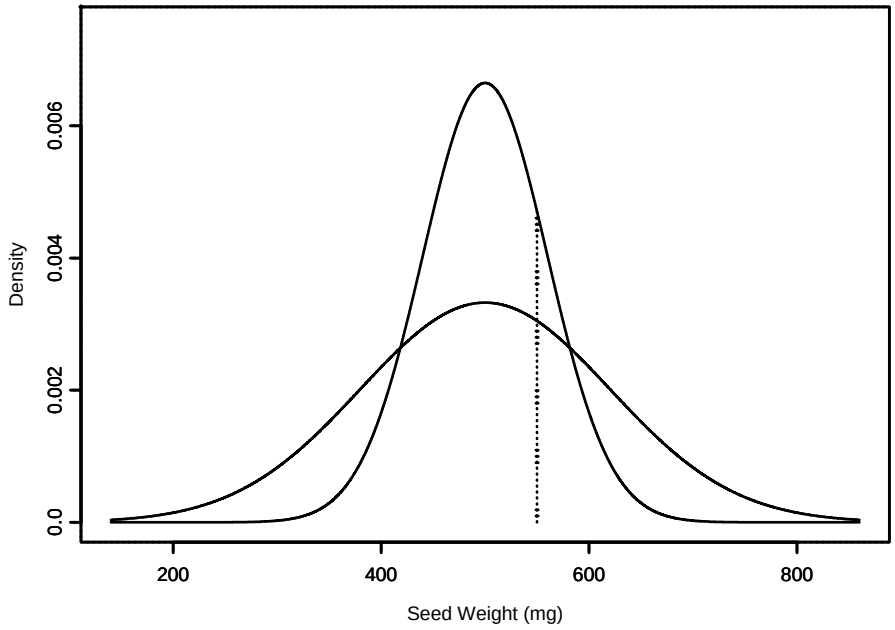
In Theorem 5.2.1, the underlying questions are (1) how close to μ to we expect $\bar{y}$ to be, and (2) what is the shape of the $\bar{y}$ distribution.

Result:

1. The MEAN of the sampling distribution of $\bar{Y}$ is equal to the population mean (i.e., $\mu_{\bar{Y}} = \mu$).
2. The *theoretical* STANDARD ERROR of the MEAN of the sampling distribution of $\bar{Y}$ is equal to $\frac{\sigma}{\sqrt{n}}$ (i.e., $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$).
3. Shape or Distribution:
   (a) If the DISTRIBUTION of the *parent population* is Normal then so too is the shape of the sampling distribution of $\bar{Y}$

(b) [This is the Central Limit Theorem, CLT] if n is large (n ≥ 25)*, then the sampling distribution of $\overline{Y}$ is *approximately Normal* even though the population distribution is not Normal.

*** For the in-class [Uniform] example, n = 16 is sufficient; for Ex. 5.3.1 (p.159), n = 32 is sufficient; for a mixture of populations such as in Ex. 5.3.2 (pp.160-1), we need n ≥ 32.**



Let's illustrate using Ex. 5.2.2 (pp.152-3) – the Weights of Seeds (Y), where Y ~ N(500mg,120mg):

(a) Find the probability of randomly choosing *a seed* with weight in excess of 550mg

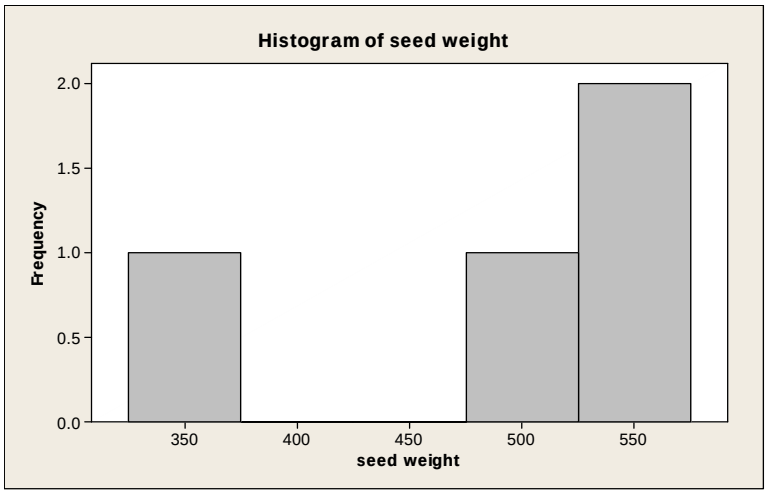
(b) Take a random sample of size n = 4: what is the probability that *the sample mean* exceeds 550mg.

(a) Y ~ N(500,120) → Pr{Y ≥ 550} = Pr{$\frac{Y-\mu}{\sigma} \geq \frac{550-500}{120}$} = Pr{Z ≥ 0.42} = 1 – Pr{Z < 0.42} = 1 – 0.6628 = 0.3372 (33.72%)

(b) $\overline{Y}$ ~ N(500,60) → Pr{$\overline{Y} \geq 550$} = Pr{$\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \geq \frac{550-500}{120/\sqrt{4}}$} =

Pr{Z ≥ 0.83} = 1 – Pr{Z < 0.83} = 1 – 0.7967 = 0.2033 (20.33%)

It's important to distinguish SE's from SD's and parent populations from sampling distributions! The Result and CLT focus on the distribution of the sample means. Note the distinctions given in Ex. 5.2.4 on p.155 and Ex. 5.2.5 on p.156. A sample of four seed weights may look like:



What's important instead is the distribution of the means.

Another illustration: Ex.5.S.1 on p.168 concerns wheat yield per plot (Y in pounds) where Y ~ N(88lbs.,7lbs.). In a sample of size n=5, find the probability the *sample mean* exceeds 90lbs. It's easy to show the answer is Pr{$\overline{Y} > 90$} = Pr{$\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \geq \frac{90-88}{7/\sqrt{5}}$ } = Pr{Z > 0.64} = 1 – 0.7389 = 0.2611 (26.11%). What if a new fertilizer treatment yields a sample (of n = 5) mean of $\overline{y}$ = 95; is this odd? Yes, since Pr{$\overline{Y} > 95$} = Pr{$\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \geq \frac{95-88}{7/\sqrt{5}}$ = Pr{Z > 2.24} = 1 – 0.9875 = 0.0125 (1.25%). More to come on this: is this proof of an improvement in wheat yields?

Section 5.4, The *Normal Approximation to the Binomial Distribution* – with the continuity correction – is very important. To illustrate the CLT applied to dichotomous variables, consider a population with p = 0.40 (40%) mutants (see, for example p.25). We'll take a sample of size n = 5 from this population *with replacement* and calculate $\hat{p}$, the sample proportion. The relevant question is again how close to the true p = 0.40 is $\hat{p}$ likely to be? To answer this question, we again need to think of a meta-experiment wherein all samples of size n = 5 from this population are chosen (with replacement) and the $\hat{p}$'s calculated and tallied. *This is actually easier for dichotomous variables than for continuous variables!* For example, what is the probability in our sample of size n = 5 with replacement that $\hat{p}$ = 0.50? How about that $\hat{p}$ = 0.60?

That's easy: $\hat{p}$ = 0.60 means Y = 3 where Y ~ BIN(n=5,p=0.40), so Pr{$\hat{p}$ = 0.60} = Pr{Y = 3} = $_5C_3(0.40)^3(0.60)^2$ = 0.2304. Thus, what's important here is Pr{Y = k} = $_5C_k(0.40)^k(0.60)^{5-k}$. Here is the sampling distribution for $\hat{p}$ with n = 5 and p = 0.40:

$\hat{p}$	Y	Probabilities
0	0	0.07776
0.20	1	0.25920
0.40	2	0.34560
0.60	3	0.23040
0.80	4	0.07680
1	5	0.01024

Note: Pr{0.20 ≤ $\hat{p}$ ≤ 0.60} = 0.25920 + 0.34560 + 0.23040 = 0.83520; how would this change if the sample size was n = 20 or n = 40?

Dependence on Sample Size – p = 0.40 & sampling w/replacement:

- **If n = 5, Pr{0.20 ≤ $\hat{p}$ ≤ 0.60} = Pr{1 ≤ Y ≤ 3} = 0.83520**
- **If n = 20, Pr{0.20 ≤ $\hat{p}$ ≤ 0.60} = Pr{4 ≤ Y ≤ 12} = 0.96301**
- **If n = 40, Pr{0.20 ≤ $\hat{p}$ ≤ 0.60} = Pr{8 ≤ Y ≤ 24} = 0.99460**

As to the shape of the sampling distribution, it is linked to the Binomial Distribution, but it turns out that in some cases *we can use the Normal Distribution to approximate this Binomial Distribution*. This is very important but we need to remember to use the CC (continuity correction).

Theorem 5.4.1 (p.163):

(a) Provided n is large, then the binomial distribution (Y) can be approximated by a N(μ, σ) distribution with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

(b) Provided n is large, then the sampling distribution of $\hat{p}$ can be approximated by a N(μ, σ) distribution with $\mu = p$ and $\sigma = \sqrt{\frac{p(1-p)}{n}}$.

The term "large" n here means that both of the following conditions must be met: np ≥ 5 and n(1-p) ≥ 5. Thus for the above mutant exercise, we could not use the N/A (Normal Approximation) when n = 5 (since np = 2 < 5), but we could use the N/A when n = 20 since both np = 8 and n(1-p) = 12 are ≥ 5.

To illustrate the *wrong* approach – without the CC – for this case (n = 20, p = 0.40), note that μ = np = 8 and $\sigma = \sqrt{np(1-p)}$ =

2.1909, so Pr{0.20 ≤ $\hat{p}$ ≤ 0.60} = Pr{4 ≤ Y ≤ 12} ≈ Pr{$\frac{4-8}{2.1909}$ ≤ Z ≤ $\frac{12-8}{2.1909}$} = Pr{-1.83 ≤ Z ≤ 1.83} = 0.9664 – 0.0336 = 0.9328 (i.e., 93.28%). To illustrate the *correct* approach – that is, with the CC – for this case, Pr{0.20 ≤ $\hat{p}$ ≤ 0.60} = Pr{4 ≤ Y ≤ 12} ≈ Pr{$\frac{3.5-8}{2.1909}$ ≤ Z ≤ $\frac{12.5-8}{2.1909}$} = Pr{-2.05 ≤ Z ≤ 2.05} = 0.9798 – 0.0202 = 0.9596 (i.e., 95.96%). Note that we should always use the CC (continuity correction) when we use the N/A (Normal approximation). Also, as pointed out above, the N/A is valid here since np = 8 and n(1-p) = 12 both are ≥ 5; *this always needs to be verified and underscored and pointed out on exams/quizzes*.

Note how much easier it is to solve this exercise in the above manner versus the following approach (again using the N/A with CC but with part (b) of the above Theorem): μ = p = 0.40 and $\sigma = \sqrt{\frac{0.40 \times 0.60}{20}}$ = 0.1095 and we wish to find Pr{0.20 ≤ $\hat{p}$ ≤ 0.60}. But how do we perform the CC here? It turns out – and this is the challenging part – that the CC here translates into adding and subtracting $\frac{1}{2n} = \frac{1}{40}$ = 0.025. Thus,

$$\Pr\{0.20 \le \hat{p} \le 0.60\} \approx \Pr\left\{\frac{0.175-0.40}{0.1095} \le Z \le \frac{0.625-0.40}{0.1095}\right\}$$
$$= \Pr\{-2.05 \le Z \le 2.05\}$$

Obviously, the rest of the story (i.e., the answer) will then be the same, but the point is to highlight that it is tougher to find the CC when dealing with sample proportions (as in part (b) of the Theorem) but it is easy to find the CC when dealing with the Binomial RV as in part (a) of the Theorem.