

Sampling Distributions

- A **population** may be modeled as a box with numbered or colored balls.
- We think of a **sample** of data as having been selected at random from this population.
- From each sample, we can calculate a **sample statistic** such as a sample mean.
- The **sampling distribution** of the sample mean is the collection of all possible sample means that could occur by random sampling (of a given sample size n).
- The textbook refers to the thought exercise of considering all the ways a sample could have turned out as a **meta-experiment**.

Larger Example

For the previous cross, what is the probability that exactly 15 of 20 offspring are dominant?

The number of dominant offspring will have a binomial distribution with $n = 20$ and $p = 0.75$.

```
> dbinom(15, 20, 0.75)
[1] 0.2023312
```

What is the probability that \hat{p} is within 0.05 of p ?

Translate the probability to a binomial question.

$$\begin{aligned}\Pr\{0.70 \leq \hat{p} \leq 0.80\} &= \Pr\{0.70 \leq Y/20 \leq 0.80\} \\ &= \Pr\{14 \leq Y \leq 16\}\end{aligned}$$

```
> sum(dbinom(14:16, 20, 0.75))
[1] 0.5606259
```

Sampling Distributions

Bret Larget

Department of Statistics

University of Wisconsin - Madison

October 8, 2003

Dichotomous Observations

Consider a cross of two heterozygotes, $Aa \times Aa$.

The probability distribution of the genotypes of the offspring is as follows.

cross	Offspring Genotype		
	AA	Aa	aa
$Aa \times Aa$	0.25	0.50	0.25

If Y is the number of dominant offspring (AA or Aa) in a sample of size $n = 2$, and if $\hat{p} = Y/n$ is the sample proportion, then the sampling distribution of \hat{p} is tabulated below. Probabilities are from the binomial distribution.

Y	\hat{p}	Prob.
0	0.0	0.0625
1	0.5	0.3750
2	1.0	0.5625

Quantitative Observations

- Now consider a population where each individual is associated with a quantitative variable.
- We can compute the sample mean from each sample.
- The [sampling distribution of \$\bar{Y}\$](#) is the collection of sample means from the meta-experiment of all possible samples of size n .

Example calculation

Suppose that the weights of seeds are approximately normal with a mean of 500 mg and a standard deviation of 150 mg. Find the probability that the sample mean is between 450 and 550 for a variety of sample sizes.

For $n = 4$, we have

$$\begin{aligned}\Pr\{450 \leq \bar{Y} \leq 550\} &= \Pr\left\{\frac{450 - 500}{150/\sqrt{4}} \leq \frac{\bar{Y} - 500}{150/\sqrt{4}} \leq \frac{550 - 500}{150/\sqrt{4}}\right\} \\ &= \Pr\{-0.67 \leq Z \leq 0.67\} \\ &= 0.5028\end{aligned}$$

from a normal table calculation.

A fancy R trick

Here is R code to do the previous calculation for a variety of sample sizes.

```
> N = 10 * c(2, 4, 8, 16, 32, 64)
> for (n in N) {
+   print(sum(dbinom(seq(0.7 * n, 0.8 * n, by = 1), n, 0.75)))
+ }
[1] 0.5606259
[1] 0.6389116
[1] 0.7553899
[1] 0.8799318
[1] 0.9670862
[1] 0.9970046
```

Sampling Distribution of \bar{Y}

- The mean of the sampling distribution of \bar{Y} , $\mu_{\bar{Y}}$, is the same as the population mean. In symbols,

$$\mu_{\bar{Y}} = \mu.$$

- The standard deviation of the sampling distribution of \bar{Y} , $\sigma_{\bar{Y}}$, is smaller than the population standard deviation by a factor of \sqrt{n} . In symbols,

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}.$$

- If the sample size n is sufficiently large, the shape of the sampling distribution of \bar{Y} will be approximately normal. This is the [Central Limit Theorem](#).
- If the population is normal, a sample size of 1 suffices.
- If the population is not normal, it depends on how the population differs from normality to determine if the normal approximation is reasonably accurate.

Exercise 5.18

Assume that height of corn plants are normally distributed with a mean 145 cm and a standard deviation of 22 cm.

What proportion of plants are between 135 and 155 cm?

```
> pnorm(155, 145, 22) - pnorm(135, 145, 22)
[1] 0.3505637
```

Find $\Pr\{135 \leq \bar{Y} \leq 155\}$ when $n = 16$.

```
> pnorm(155, 145, 22/sqrt(16)) - pnorm(135, 145, 22/sqrt(16))
[1] 0.9309637
```

Find $\Pr\{135 \leq \bar{Y} \leq 155\}$ when $n = 36$.

```
> pnorm(155, 145, 22/sqrt(36)) - pnorm(135, 145, 22/sqrt(36))
[1] 0.993614
```

Fancy R example

Here is sample R code to do this calculation for several n differently than the previous example.

```
> N = c(4, 8, 16, 32, 64)
> len = length(N)
> p = rep(0, len)
> for (i in 1:len) {
+   p[i] = pnorm(550, 500, 150/sqrt(N[i])) - pnorm(450, 500,
+     150/sqrt(N[i]))
+ }
> cbind(N, p)
      N      p
[1,]  4 0.4950149
[2,]  8 0.6542214
[3,] 16 0.8175776
[4,] 32 0.9406536
[5,] 64 0.9923392
```

The first number in this table disagrees with the previous calculation slightly because R did not round off the z score to the nearest hundredth.