

## Math 373 Lecture 10

One point extra credit for finding errors in handouts. Half a point for grammatical errors.

Read the last third of Lecture 9.

- Suppose there are 100 rats in cages numbered 1- 100. The rats in the earlier cages are older than later cages. Half of the rats are female. You are to test the safety of a drug on the rats. You test the drug on 25 rats and use another 25 rats as controls. How should you pick the rats?

### Population, sample and sampling distributions

- You have 4 coins in your pocket: 2 pennies, 1 nickel, 1 dime.

- EXPERIMENT A. Randomly pick 1 coin and note its value.

Population:  $\{1, 1, 5, 10\}$ .

Population mean:

$$\mu = (1+1+5+10)/4 = 17/4 = 4.25\text{¢}$$

Std. dev.

$$\sigma = 3.70\text{¢} \quad \text{This is a population distribution.}$$

If  $S$  is a 3-element sample chosen from the population, then  $S$  will have its own sample mean  $\bar{x}$  and std. dev.  $s$ .

Suppose  $S = \{1, 1, 10\}$ .

- EXPERIMENT B. Randomly pick 1 coin from the sample  $S = \{1, 1, 10\}$ .

Population:  $\{1, 1, 10\}$

Sample mean

$$\bar{x} = (1+1+10)/3 = 12/3 = 4\text{¢},$$

Std. dev.

$$s = 5.20\text{¢} \quad \text{This is a sample distribution.}$$

Now consider the process of drawing a 3-element sample from the population  $\{1, 1, 5, 10\}$  and measuring the sample mean.

This is also a statistical experiment and we can calculate the mean and std. dev. of these sample means.

- EXPERIMENT C. Pick a 3-element sample from  $\{1, 1, 5, 10\}$  and measure its mean.

Experimental units:  $\{\{1, 1, 5\}, \{1, 1, 10\}, \{1, 5, 10\}, \{1, 5, 10\}\}$

Population:  $\{7/3, 12/3, 16/3, 16/3\} = \{2.33, 4, 5.33, 5.33\}$

Probability distribution:

$\bar{x}$	2.33	4	5.33
$P(x)$	0.25	0.25	0.5

This is the 3-element sampling distribution for  $\{1, 1, 5, 10\}$

Mean  $E(\bar{x}) = 2.33(.25) + 4(.25) + 5.33(.5) = 4.25\text{¢}$ .

The standard error,  $S.E. = \text{std. dev. of } \bar{x} =$

$$\sqrt{(2.33 - 4.25)^2 .25 + (4 - 4.25)^2 .25 + (5.33 - 4.25)^2 .50} = 1.23\text{¢}.$$

This is not the distribution of items in a particular sample; it is the distribution of 3-element samples in the set of all 3-element samples.

Some of the sample means  $\{2.33, 4, 5.33, 5.33\}$  are less than the population mean of 4.25, some are more. But the **average of the sample means is exactly the population mean**. However, the std. dev. of the sample means is smaller than the std. dev. of the population.

This isn't an accident.

CENTRAL LIMIT THEOREM. Consider the experiment of drawing an  $n$ -element sample from an  $N$ -element population. Suppose:

- the sample is small compared the population:  $n \leq .05N$ ,
- the population has mean  $\mu$  and std. dev.  $\sigma$ ,
- either the original population is normal or  $n \geq 30$ .

Then, the sampling distribution of  $\bar{x}$  is approximately a normal distribution with

$$\text{mean } E(\bar{x}) = \mu_{\bar{x}} = \bar{x}, \text{ and S.E.} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

In the above example, we had  $n=3$ ,  $N=4$ ,  $\mu=4.25$  and  $\sigma=3.70$ . The mean  $\mu_{\bar{x}}$  of the sample means  $\bar{x}$  of the 3 coin samples will be exactly the mean  $\mu=4.25$  of the population. The std. dev.  $\sigma_{\bar{x}}$  of the sample means  $\bar{x}$  will be approximately  $3.70/\sqrt{3} = 2.14$ . In this case the approximation isn't very good since  $n=3$  is too small.

Since  $n > .05N$  we need the hypergeometric formula to get

$$\text{the exact answer: } \sigma \frac{1}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = 3.70 \frac{1}{\sqrt{3}} \sqrt{\frac{4-3}{4-1}} = 1.23.$$

Every sample statistic (mean, std. dev. median, range,  $Q_1$ ,  $Q_3$ , IQR) is a random variable for the experiment of drawing a sample and thus each statistic has its own probability distribution called its *sampling distribution*.  $\mu_{\bar{x}}$  and  $\sigma_{\bar{x}}$  are the mean and std. dev. of the  $n$ -element *sampling distribution* for the sample mean  $\bar{x}$ . For the sample median statistic, the sampling mean is the mean of the dataset of the medians of all possible  $n$ -element samples.

Note: When  $n=1$ , Experiment C (picking a 1-element sample) is the same as Experiment A (Randomly picking an element of the original population). When  $n=N$ , Experiment B is the same as Experiment A.

A sample size is *small compared to the population* if  $n \leq .05N$ . A sample is *large* if  $n \geq 30$ ; it is *small* if  $n < 30$ .

Thus the Central Limit Theorem says that for large samples which are small compared to the population, probabilities dealing with the sample mean can be calculated using the normal distribution. For small samples, Student's t-distribution is used instead.