

Handout 06 - Sampling Distribution

Jin, Ick Hoon

July 20th, 2009

- *Population Distribution* (of a variable)
 - The distribution of all the members of the population.
- *Sampling Distribution*
 - This is not the distribution of the sample.
 - The sampling distribution is the distribution of a statistic.
 - If we take many, many samples and get the statistic for each of those samples, the distribution of all those statistics is the sampling distribution.
 - We will most often be interested in the sampling distribution of the sample mean or the sample proportion.

- Until now, we have only talked about population distributions.
 - Example: Suppose the proportion of those who agree with a particular UN policy is 0.53.
- Suppose we randomly sample 1000 individuals and ask them if they agree with the UN policy.
 - What is the distribution of the sample proportion?
 - Does this distribution differ from the population distribution?

- Population Distribution
 - Example: Scores on an Intelligence Scale for the 20 to 34 age group are normally distributed with mean 110 and standard deviation 25.
- Suppose we sample 50 individuals between 20 and 34 and obtain the mean and standard deviation of that sample.
 - What is the sampling distribution of the sample mean?
 - Does this sampling distribution differ from the population distribution?

- The random variable, Y , is a count of the occurrences of some outcome in a fixed number of observations.
- We say the distribution of the counts, Y , of successes has a Binomial distribution.
- Rules for the binomial setting
 - There are a fixed number n of observations.
 - The n observations are all independent.
 - Each observation falls into one of just two categories, which for convenience we call “success” and “failure”.
 - The true probability of a success, π , is the population proportion of successes. It is the same for each observation.

- Let Y be a count of the occurrences, “successes”, of some outcome in a random sample of size n . For example, Y is the number of books in Spanish language in a sample of many books. Note that $0 \leq Y \leq n$.
- If the following holds:
 - Observations in a sample are independent.
 - Each observation falls into one of two categories: “success” and “failure”. For example if the randomly chosen book is Spanish we treat it as a “success”, $X = 1$. If it is not Spanish, we count it as a “failure”, $X = 0$. Note that $Y = \sum_{i=1}^n x_i$.
 - The true probability of “success”, which is the population proportion (parameter), is the same for each observation.

For $Y \sim \text{Bin}(n, \pi)$

$$P(Y = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k} = \frac{n!}{k!(n-k)!} \pi^k (1 - \pi)^{n-k}$$

$$\text{where } n! = \prod_{i=1}^n i = 1 \times 2 \times 3 \times \dots \times n$$

- The mean for Y is $\mu_Y = n\pi$ and
- The standard deviation for Y is $\sigma_Y = \sqrt{n\pi(1 - \pi)}$.
- For n is **very large**, Binomial distribution can be **approximated** by normal distribution

$$Y \sim \text{Bin}(n, \pi) \xrightarrow{n \text{ large}} Y \overset{\text{approx}}{\sim} N\left(n\pi, \left(\sqrt{n\pi(1 - \pi)}\right)^2\right)$$

- Typographic errors in a text are either nonword errors (as when “the” is typed as “teh”) or word errors that result in a real but incorrect word. Spellchecking software will catch nonword errors but not word errors.
- Suppose human proofreaders catch 70% of nonword errors. You ask a fellow student to proofread an essay in which you have deliberately made 29 nonword errors.

- What are we given? We know $\pi = 0.7$, $n = 29$

- What is the mean number of errors caught?

$$\mu_{\text{nonword}} = n\pi = 29(0.7) = 20.3$$

- What is the standard deviation of the number of errors caught?

$$\sigma_{\text{nonword}} = \sqrt{n\pi(1 - \pi)} = \sqrt{29(0.7)(0.3)} = 2.47$$

- We call the population proportion, π . (Your textbook uses p for the population proportion).
- There is only one population proportion, π .
- We call the sample proportion p . (Your textbook uses \hat{p}).
- p is calculated as x/n , where x is the count of successes and n is the total sample size

$$p = \frac{\text{number of successes}}{n}$$

- The distribution of the sample proportions, or **sampling distribution** of p is approximately

$$p \sim N \left(\pi, \left(\sqrt{\frac{\pi(1-\pi)}{n}} \right)^2 \right)$$

when n (the sample size) is large.

- **As a rule of thumb, use this approximation for values of n and p that satisfy $n\pi \leq 10$ and $n(1 - \pi) \leq 10$.**
- Why do we need to check conditions $n\pi = 10$ and $n(1 - \pi) = 10$?
These conditions say how large the sample size n should be to ensure that p is a reasonable estimate for π .
- Note that p which is close to 0.5 is easy to estimate. Indeed, when $p = 0.5$, sample size $n = 20$ is large enough, since $20(0.5) = 10$.
- When p is close to 0 or 1, the estimation problem becomes extremely difficult, since the conditions $np = 10$ and $n(1 - p) = 10$ can only be met for extremely large n , or $n = \infty$.

Bookstore example (revisited)

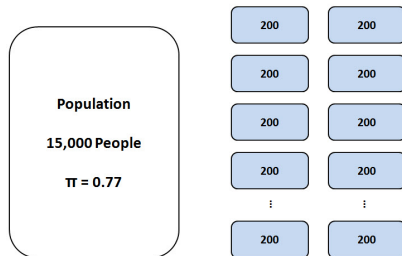
- Probability that a book selected at random in Spanish is $\pi = 0.04$.
- Suppose we selected $n = 10$ books at random from the store and no one of them is Spanish. In fact, we can show that probability of this is event (no Spanish book among 10 chosen books) is big, 0.66.
- Then we will find that $p = 0/10 = 0$ which is a poor estimate of $\pi = 0.04$. Note that $n\pi = 10(0.04) = 0.4 < 10$
- If we now take $n = 250$ books, the chance that our sample will not contain any Spanish book will be very low. In fact, we can show that probability of this is event (no Spanish book among 250 chosen books) is 0.000037. Our p will be better estimate of $\pi = 0.04$. Note in this case, $n\pi = 250(0.04) = 10$.

- Suppose a large department store chain is considering opening a new store in a town of 15,000 people.
- Further, suppose that 11,541 of the people in the town are willing to patronize the store, but this is unknown to the department store chain managers.
- Before making the decision to open the new store, a market survey is conducted.
- 200 people are randomly selected and interviewed. Of the 200 interviewed, 162 say they would patronize the new store.

- What is the population proportion π ? $11,541/15,000 = 0.77$
- What is the sample proportion p ? $162/200 = 0.81$
- What is the approximate sampling distribution (of the sample proportion)?

$$\begin{aligned} p &\sim N\left(\pi, \left(\sqrt{\frac{\pi(1-\pi)}{n}}\right)^2\right) = N\left(0.77, \left(\sqrt{\frac{0.77(1-0.77)}{200}}\right)^2\right) \\ &= N(0.77, 0.0297^2) \end{aligned}$$

- Suppose we take many, many samples (of size 200): Then we find the sample proportion for each sample.



- The sampling distribution of all those p 's (0.74, 0.81, 0.76, 0.77, 0.80, 0.71, 0.75, 0.75, 0.82, ...) is

$$p \sim N(0.77, 0.0297^2)$$

- The managers did not know the true proportion so they took a sample.
- As we have seen, the samples vary.
- However, because we know how the sampling distribution behaves, we can get a good idea of how close we are to the true proportion.
- This is why we have looked so much at the normal distribution.
- Mathematically, the normal distribution is the sampling distribution of the sample proportion, and, as we will see, the sampling distribution of the sample mean as well.

The lead level in a child's body is considered to be dangerously high if it exceeds 30 micrograms per deciliter. Children come into contact with lead from a variety of sources, but are particularly susceptible to exposure from eating paint from toys, furniture, and other objects.

A random sample of 1000 of 20,000 children living in public housing projects in a particular city revealed that 200 of them had dangerously high lead levels in their bodies.

(Adapted from *Intro. to Statistics*, Milton, McTeer and Corbet, 1997)

In 1987 over 3 million acres were reforested with 2 billion seedlings. A drought during the next growing season killed many of these seedlings. A sample of 1000 seedlings is obtained, and it is discovered that 300 are dead.

(Info. in Howard Burnett, “A Report on Our Stressed-Out Forests”, *American Forests*, April 1989, pp.21-25)

- We have already considered population means and sample means.
- The distribution of the sample mean, or **sampling distribution of the sample mean** is approximately

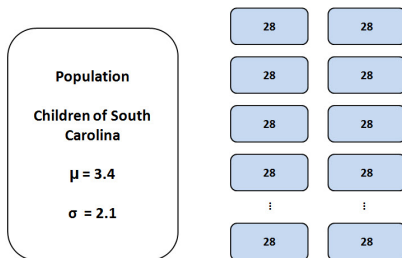
$$\bar{X} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$$

- There has been some concern that young children are spending too much time watching television.
- A study in Columbia, South Carolina recorded the number of cartoon shows watched per child from 7:00 a.m. to 1:00 p.m. on a particular Saturday morning by 28 different children.
- The results were as follows: 2, 2, 1, 3, 3, 5, 7, 5, 3, 8, 1, 4, 0, 4, 2, 0, 4, 2, 7, 3, 6, 1, 3, 5, 6, 4, 4, 4. (Adapted from **Intro. to Statistics**, Milton, McTeer and Corbet, 1997)
- Suppose the **true** average for all of South Carolina is 3.4 with a standard deviation of 2.1.

- What is the population mean? 3.4
- What is the sample mean? $99/28 = 3.535$
- What is the approximate sampling distribution (of the sample mean)?

$$\bar{X} \sim N\left(3.4, \left(\frac{2.1}{\sqrt{28}}\right)^2\right) = N(3.4, 0.4^2)$$

- Suppose we take many, many samples (each sample of size 28), then we find the sample mean for each sample.
- The sampling distribution of all those means (2.9, 3.4, 4.1, ...) is distributed $N(3.4, 0.4^2)$.



Some remarks:

- Similar to the example for sample proportions, the sampling distribution of the sample means follow a normal distribution.
- This allows us to determine with some certainty how likely our sample mean is to be near the true population mean.
- In reality, we don't have the luxury of obtaining many, many samples. We can only assume we do and say our sample is one of those many.

Suppose past studies indicate it takes an average of 6 minutes to memorize a short passage of 20 words.

A psychologist claims a new method of memorization will reduce the average time to 4.5 minutes.

A random sample of 40 people are to use the new method.

The average time required to memorize the passage will be found.

The accepted maximum exposure level to microwave radiation in the United States is 10 microwatts per square centimeter.

Citizens of a small town near a large television transmitting station feel the station is polluting the air with enough microwave radiation to push the surrounding levels above the standard exposure limit.

The people randomly select 25 days to measure the microwave radiation to obtain statistical evidence to back up their contention.

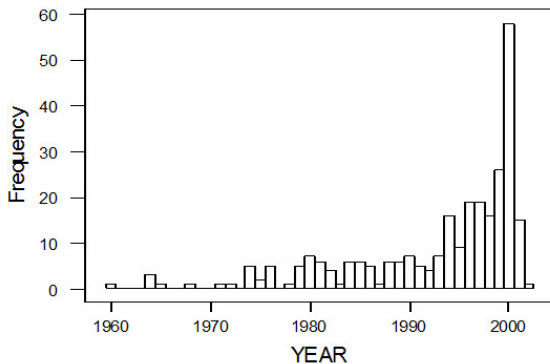
(Adapted from **Intro. to Statistics**, Milton, McTeer and Corbet, 1997).

- The Central Limit Theorem states that for **any** population with mean μ and standard deviation σ , the sampling distribution of the sample mean, \bar{x} , is approximately normal when n is large.

$$\bar{x} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$$

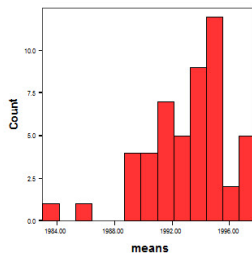
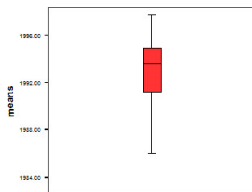
- The central limit theorem is a very powerful tool in statistics.
- Remember, the central limit theorem works for any distribution.
- Let us see how well it works for the years on pennies.

Penny Population Distribution (276)



- Note from the previous slide, the distribution is highly left skewed.
- The mean of the 276 pennies is 1992.9.
- The standard deviation of the 276 pennies is 8.7.
- Let us take 50 samples of size 10.
- According to the Central Limit Theorem, the sampling distribution of the sample means should be normal with mean 1992.9 and standard deviation $8.7/\sqrt{10} = 2.75$.

- That is, the sampling distribution should be a normal distribution
- Suppose we took 50 samples from these pennies.



- The distribution of the means of the 50 samples is

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
MEANS	50	1983.00	1997.80	1993.0820	2.9117
Valid N (listwise)	50				

- Notice the mean is close to 1992.9 and the standard deviation is not far from 2.75
- The previous slide shows the distribution of the means of the 50 samples is slightly skewed but closer to the normal distribution.
- A suggestion would be to take samples of sizes larger than 10

Suppose $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ and X and Y are independent.

- Then $\mu_{X \pm Y} = \mu_X \pm \mu_Y$ and $\sigma_{X \pm Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$. In particular,

$$X \pm Y \sim N\left(\mu_X \pm \mu_Y, \left(\sqrt{\sigma_X^2 + \sigma_Y^2}\right)^2\right)$$

- Then for any constants a and b the distribution of $W = aX \pm bY$ is normal: $W \sim N(\mu_W, \sigma_W^2)$. where

$$\mu_W = a\mu_X \pm b\mu_Y$$

$$\sigma_W = \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2}$$

For example:

Let \bar{X} and \bar{Y} denote the average scores for the class on the first and the second STAT test, respectively. It is known that

$$\bar{X} = N(80, 4^2)$$

$$\bar{Y} = N(70, 3^2)$$

Then

$$\bar{X} - \bar{Y} = N\left(-10, \left(\sqrt{4^2 + 3^2}\right)^2\right)$$