BIOSTAT 701

Introduction to Statistical Theory and Methods I

Lynn Lin

Statistical inference

- In inferential statistics, we want to use characteristics of the sample to estimate the characteristics of the population.
- For example: one uses sample mean to estimate the population mean. In doing so, we need to know the properties of the sample mean. That is why need to study the sampling distribution of the sample mean.

Example

- A large tank of fish from a hatchery is being delivered to the lake. We want to know the average length of the fish in the tank. Instead of measuring all (tens of thousands) the fish, we randomly sample some of them and use the sample mean to estimate the population mean.
  - This is point estimation.

Remark

- Note that the sample mean $\bar{X}$ is random since its value depends on the sample we get. It is a statistic. The population mean is fixed and denoted as μ: $\bar{X} \neq \mu$
- Thus, the sampling distribution of the (sample) mean is also called the distribution of the variable $\bar{X}$.
- Usually, the sampling distribution of the sample mean is complicated except for very small sample size or for large sample size.
- In the following example, we illustrate the sampling distribution for a very small population. The sampling method is to sample without replacement.

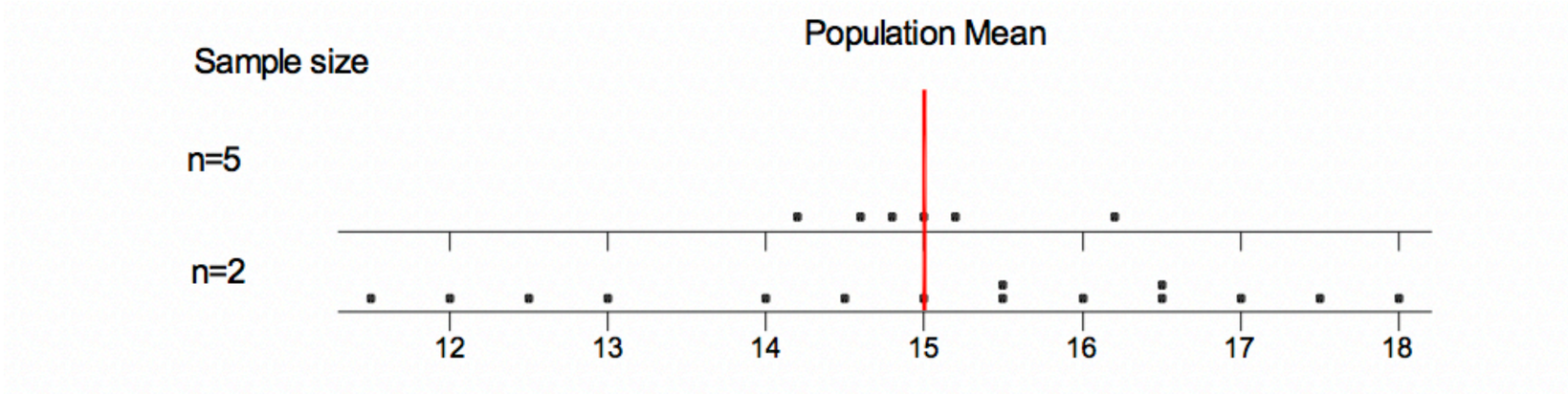
Example

- The population is the weight of 6 pumpkins in pounds displayed in a carnival “guess the weight” game stall. You are asked to guess the average weight of the 6 pumpkins by taking a random sample without replacement from the population.
- Suppose we estimate the population mean by randomly select 2 pumpkins. What is the distribution of the sample mean? What if we select 5 pumpkins?

Pumpkin	A	B	C	D	E	F
weight	19	14	15	9	16	17

Example

- The following dot plots show the distribution of the sample means corresponding to sample sizes of 2 and 5.



Definition

- Sampling error is the error resulting from using a sample to estimate a population characteristic.
- As the dotplots in the previous figure show, the possible sample means cluster more closely around the population mean as the sample size increases. Thus, possible sampling error decreases as sample size increases.

Formulas

- The mean of sample mean is the population mean: $E(\bar{X}) = \mu$
- When sampling with replacement, the standard deviation of the sample mean equal the population standard deviation divided by the square root of the sample size: $\sqrt{Var(\bar{X})} = \sigma/\sqrt{n}$

Estimation of the population mean

- In practice, the population distribution is usually unknown. We are often interested in population parameters, like the population mean.
- As all we know about the population is the sample, we can only use the sample to estimate the population parameter of interest, called statistic.
- Sample statistics vary from sample to sample.
- How close is the sample mean to the population mean?

Sampling distribution

- The probability distribution of a statistic is called the sampling distribution of the statistic.
- E.g., sampling distribution of the sample mean

Key fact 1

- The mean of sample mean is the population mean: $E(\bar{X}) = \mu$
- When sampling with replacement, the standard deviation of the sample mean equal the population standard deviation divided by the square root of the sample size: $\sqrt{Var(\bar{X})} = \sigma/\sqrt{n}$
- This holds true for all types of distribution for the population, and no distribution assumption on X is needed.

Key fact 2

- If the population is normally distributed, then the sampling distribution of the sample mean is also normally distributed no matter what the sample size is.

Key fact 3

- If the population is not normally distributed, then the distribution of $\bar{X}$ may be different from a normal distribution unless the sample size is large (say > 30).

Central limit theorem

- For a large sample size (n ≥ 30), $\bar{X}$ is approximately normally distributed, regardless of the distribution of the population one samples from.
- If the population has mean μ and standard deviation σ, then $\bar{X}$ has mean μ and standard deviation $\sigma/\sqrt{n}$: $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$
- What is the distribution of the sum n$\bar{X}$?

Example

- A housing survey was conducted to determine the price of a typical home in a city of CA. The mean price of a house was roughly $1.5 million with an SD of $0.8 million. There were no houses listed below $0.5million but a few houses above $4 million.
- Can we find an approximate probability that a randomly chosen house in that city costs more than $1.6 million using the normal distribution?

Example

- The CLT says you take larger and larger samples from a population, the histogram of the sample values looks more and more normal.
- True or false?

Example

- A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean $\mu = 205$ pounds and standard deviation $\sigma = 15$ pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

Example

- Suppose a certain movie has a distribution of ratings, in a 1 to 10 scale, of those watched the movie, 1/3 gave 9 points, 1/3 gave 2 points, and the remaining 1/3 gave 1 point.
- What is the population distribution?

Example

- In practice, since the population are hard to examine, we take a sample to learn about the population.
- We take a simple random sample of size n from the population.
- We will first examine the histogram of the movie ratings in the sample.

```
ratings = c(1,2,9)
n = 400
s = sample(ratings, size = n, replace=T, prob=c(1/3,1/3,1/3))
hist(s, breaks=0:10+.5, xlab="Ratings", main="Sample Size = 400")
```

Example

- What will the histogram of the sample means look like?
  - Take a random sample of size n, say n = 25, from the population, compute and record the sample mean, and the put the sample back.
  - Repeat the previous step 10000 times, and then obtain 10000 sample means.

Example

```
n = 25
samplemean = vector("numeric", 10000)
for(i in 1:10000){
    samplemean[i] = mean(sample(ratings, size = n, replace=T, prob=c(1/3,1/3,1/3)))
}
hist(samplemean, xlab="sample mean", main="Histogram of the Means of 10000 Samples of Size 25")
abline(v=4, col=2)
```

Example

```
n = 25
samplemean = vector("numeric", 10000)
for(i in 1:10000){
    samplemean[i] = mean(sample(ratings, size = n, replace=T, prob=c(1/3,1/3,1/3)))
}
hist(samplemean, xlab="sample mean", breaks=seq(1,8,by=0.04), main="Histogram of the Means of
10000 Samples of Size 25")
abline(v=4, col=2)
```

Example

```
n = 100
samplemean = vector("numeric", 10000)
for(i in 1:10000){
    samplemean[i] = mean(sample(ratings, size = n, replace=T, prob=c(1/3,1/3,1/3)))
}
hist(samplemean, xlab="sample mean", breaks=seq(1,8,by=0.04), main="Histogram of the Means of
10000 Samples of Size 100")
abline(v=4, col=2)
```