Rashika Singh
IST 687

Lecture Notes- 4

Sampling:

Sampling distributions are conceptual key to statistical inference. Example of it are drawing marbles from a large jar which is sample of population. Forces of randomness drive uncertainty and multiple draws are used for increasing confidence.

In the process of drawing marbles from a jar:

The gum balls are the population which we want to sample and look at the distribution.

Example-

Sample() is function in R which draws random samples with a single call.

```
sample (USstatePops$april10census, size=16, replace=TRUE)
#first parameter-vector (where you want to sample from), number of samples, replace in the
#sample
Mean (sample(USstatePops$april10census, size=16, replace=TRUE)) #mean of the sample
```

Why is sampling from a population important?

We want to do sampling because it is not always possible to take population as it would take time, so we take a section of population called the sample instead.

What are some key things to think about when sampling?

Key things to think about while sampling are how variable the population is, the precision level required and how confident should the results be.

Replication:

Replication is repeating the sampling to get many samples, this is done by replication.

```
replicate (4,mean(sample(USstatePops$april10census, size=16, replace=TRUE)),
simplify=TRUE)
# 4 is the number of times we want to replicate and simplify=TRUE argument asks R to return
the results as a simple vector of means,
```

Example-

```
Mean(replicate(400, mean(sample(USstatePops$april10census, size=16,
replace=TRUE)),simplify=TRUE)
#Draws 400 samples of size 16 from the state population. Calculates the mean from each sample
#and keep in a list. Calculate the mean of 400 samples.
```

Law of large numbers:

It states that if you run a statistical process a large number of times, it will converge on a stable result.

Central limit theorem:

When we look at sample means and consider "law of large numbers", the distribution of sampling means creates a bell-shaped normal distribution and the center of that distribution, the mean of the sample means, gets close to the population mean.

Code:

```
SampleMeans<- replicate(10000,mean(sample(USstatePops$april10census, size=120,
replace=TRUE)), simplify=TRUE)
#stores the replication into a variable SampleMeans with size of the sample as 120 and
#replication to be done 10000 times
length(SampleMeans) #gives length of the sampleMeans
mean(SampleMeans) # gives mean of the sampleMeans
Hist(SampleMeans)  #creates histogram
summary( SampleMeans) #gives a summary such as quartiles, min, max
quantile(SampleMeans, prob=c(0.25, 0.50, 0.75))
#generates quantiles to the corresponding probabilities
```

Basis for most statistical inferences:

Construct a comparison distribution

Identify a zone of extreme values

Compare new sample of data to the distribution relative to the "extreme" zones

If new sample does fall in the “extreme zone”, you can conclude that new sample is obtained from other source than the comparison distribution
Quantify the distribution with sd()
sd(SampleMeans) #standard deviation for SampleMeans

Standard error of mean:
Standard deviation of distribution of sampling means known as standard error of mean
Sd(population)/Sqrt (# of samples)
Standard error of the mean is relative to calculating sd
sd(USstatePops$april10census)/sqrt(120)
Differences are due to randomness of the distribution

Two standard deviations down from mean is 5% cut point and two standard deviation s if the 95% cut point

```
StdError<-sd(USstatePops$april10census)/sqrt(120)
CutPoint5<-mean(USstatePops$april10census)-(2*StdError)
CutPoint95<-mean(USstatePops$april10census)+(2*StdError)
```

Why is it useful (or when it is useful) to compare two samples?
We need to compare two samples example in case of comparing climate for same month through different years. This helps to get an idea of whether August this year is hotter than last year. Similarly, it is used to compare old product to new products, in case you purchased a new dryer and want to compare the average amount of time to dry clothes you can use both the old and new dryer’s data to get an accurate result.

As a data scientist, example of new coke machine helps us to get analysis of different types of coke flavors preferred which they can make it available to normal supply or inventory management. It can be used for sales revenue count and have insights into the customers choices.

Rashika Singh
IST 687

Question:

Are there any different types of sample, as mentioned in the book statisticians mainly use sampling with replacement, are there any other ways to do it? What are its benefits?