Introduction to Statistics - Young Researchers Fellowship Program

Lecture 4 - Statistical Inference Foundations

Daniel Sánchez Pazmiño

Laboratorio de Investigación para el Desarrollo del Ecuador

September 2024

Statistical Inference Foundations

- **Statistical Inference**: The process of making an estimate, prediction, or decision about a population based on sample data.
- Before, we've used the laws of probability to describe the behavior of random variables. Now, we will use these laws to make inferences about the population from which the sample was drawn.
- For this, we need to briefly review sampling and understand the concept of a **sampling distribution**.

Sampling

- **Population**: The entire group of individuals or instances about whom we want to draw conclusions.
- **Sample**: A subset of the population that we actually observe.
- **Sampling Frame**: A list of all the individuals in the population from which the sample is drawn.
- **Sampling Design**: The method used to select the sample from the population.

Simple Random Sampling (SRS)

- The most basic form of sampling is simple random sampling.
- In simple random sampling, each individual in the population has an equal chance of being selected.
- This is the most basic form of sampling, but it is also the most powerful.
- It is the basis for many other sampling methods.

Sampling with R - Simple Random Sampling

- Multiple ways to sample in R.
- The `dplyr` package provides a simple way to sample data.
- The `sample_n()` function can be used to sample a fixed number of observations.
- The `sample_frac()` function can be used to sample a fixed fraction of observations.

Sampling with R - Simple Random Sampling

- A base R function `sample()` can also be used to sample data.
- The `sample()` function can be used to sample a fixed number of observations.
- We may also sample from a vector of values.

Sampling with R - Simple Random Sampling

- `slice_sample()` function from the `dplyr` package can be used to sample a fixed number of observations.
- The `slice_sample()` function is useful when we want to sample a fixed number of observations from a data frame.
- All sample methods in R will require us to set a seed to ensure reproducibility.
  - This is like a random number generator that will always produce the same sequence of random numbers.
- If we don't set a seed, we will get different samples each time we run the code!
- Use `set.seed()` to set the seed for reproducibility.

Sampling with R - Simple Random Sampling

```r
# Load the dplyr package

# Sample 10 firms from the SUPERCIAS data.

set.seed(123)

supercias_sample <- supercias_raw %>%
  sample_n(10)

supercias_sample
```

```
# A tibble: 10 x 25
   no_fila expediente ruc        nombre situacion_legal fecha_co
     <dbl> <chr>      <chr>      <chr>  <chr>           <chr>
 1  182735 729286     099326627~ OHH D~ ACTIVA          25/06/20
 2  188942 735787     179313118~ REMOT~ ACTIVA          19/03/20
 3  134058 347755     099338173~ PROYE~ ACTIVA          03/05/20
```

With replacement or without replacement?

- In simple random sampling, we typically sample without replacement.
- This means that once an individual is selected, it is not replaced in the population.
- This is the most common form of sampling.
- Sampling with replacement is less common, but it is used in some cases.
  - For example, when we are sampling from a small population.
  - When we are sampling from a population that is changing.

Other types of sampling

- **Stratified Sampling**: The population is divided into subgroups, or strata, and a simple random sample is taken from each stratum.
  - This is useful when we want to ensure that each subgroup is represented in the sample.
  - May oversample some groups, and we must fix this in the analysis through **sample weights**.
  - Common with survey data.
- **Cluster Sampling**: The population is divided into clusters, and a simple random sample of clusters is taken.
  - Useful when the population is geographically dispersed.
  - Common with survey data too.
  - Must fix the oversampling of some groups through **sample weights** as well.

Other types of sampling

- **Systematic Sampling**: The population is ordered, and every nth individual is selected.
  - Useful when the population is ordered in some way.
  - May introduce bias if the population is ordered in a way that is related to the outcome.
  - Typically used in quality control (not too applicable in research!)
- **Convenience Sampling**: The sample is selected based on convenience.
  - This is the least reliable form of sampling. Never do this!
- **Snowball Sampling**: The sample is selected based on referrals from other participants.
  - Useful when the population is hard to reach.
  - Common in social network studies.

Concepts of the Sampling Distribution

- Consider a "thought experiment" where we purposely take several samples from the same population.
  - Not something we would do in practice, but it helps us understand the concept of the sampling distribution.
- Each sample will have its own sample mean, sample standard deviation, etc.
  - We know that they will vary from sample to sample.
- Because of the randomness in the sampling process, the sample statistics will be an RV.
  - And remember, RVs have probability distributions!

Notation for Sample Statistics vs. Population Parameters

- We use different notation for sample statistics and population parameters.
- Everything related to the sample will generally be lowercase.
  - Sample mean: $\bar{x}$
  - Sample standard deviation: s
- Everything related to the population will generally be greek letters.
  - Population mean: μ
  - Population standard deviation: σ
  - We never know these, remember! We're just giving them names for the sake of discussion.

Notation for Sample Statistics vs. Population Parameters

- Sometimes, we use a hat (circumflex) to denote that it is an estimate of the population parameter.
  - Sample mean estimate: $\hat{\mu}$
  - Sample standard deviation estimate: $\hat{\sigma}$
  - Sample proportion estimate: $\hat{p}$ (very common, because there is no greek letter for proportion!)
- We do observe the sample statistics, but we never observe the population parameters.

Sampling Distribution of the Sample Mean

- The **sampling distribution of the sample mean** is the distribution of the sample means from all possible samples of a given size from a population.
- The sampling distribution of the sample mean is a probability distribution that describes the behavior of the sample mean.
- The sampling distribution of the sample mean is a theoretical concept.
  - We will never observe it in practice.
  - But we will use it in theoretical discussions and derive important results from it.
  - We will make inferences about the population based on the sampling distribution of the sample mean.

It's sample time!

- Remember the SUPERCIAS data? We may assume it is a population.
- In this very, very particular case, we may actually *know* the population mean of `capital_suscrito`
  - There are of course ways to contest this, but play along for now.
- The statistics of the population are:

```
summary(supercias_raw$capital_suscrito)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
       0      400      800   136384     1200 900000000
```

It's sample time!

- Let's take a sample of 10 firms from the SUPERCIAS data.

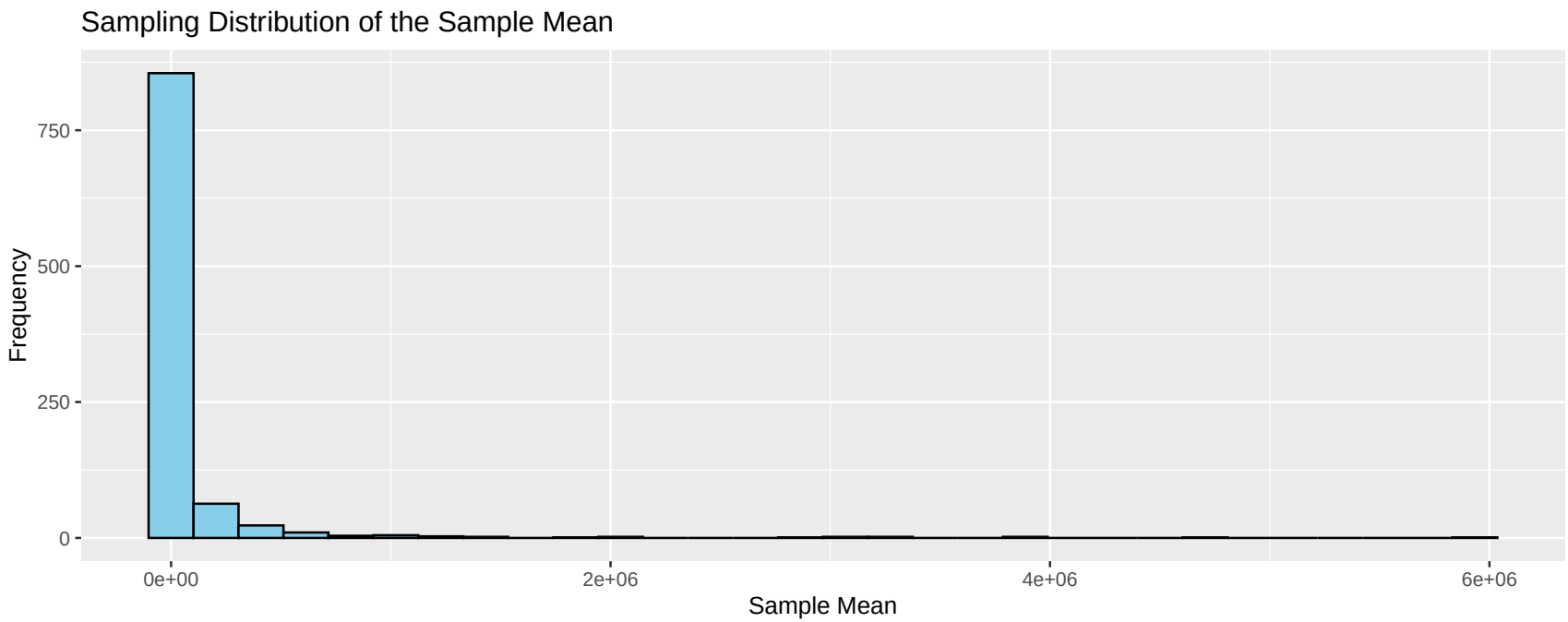
```
set.seed(123)

supercias_sample <- supercias_raw %>%
  sample_n(10)

summary(supercias_sample$capital_suscrito)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    150     575     800    2605     950   10800
```

It's sample time!

- Notice how my sample mean is different from the population mean.
  - This is expected, because the sample mean is a random variable!
- Let's repeat the process above 1000 times and plot the distribution of the sample means.
- We do this with a `for` loop in R, or with the `replicate()` function.

Sampling Distribution of the Sample Mean for `capital_suscrito`



Sampling Distribution of the Sample Mean for `capital_suscrito`

- Impressive! The distribution of the sample means is not too great
  - It inherited the shape of the population distribution.
- However, this strange distribution is not the population distribution.
  - It is the sampling distribution of the sample mean!
  - We may use laws of probability to describe its behavior.

Expected value and standard error of the sample mean

- The expected value of the sample mean would be calculated in the same way as the expected value of any other random variable.
  - Because it is not observed, we use the notation $E(\bar{x})$.
- The standard deviation from the sampling distribution of the sample mean is called the **standard error**.
  - It is calculated as $\frac{\sigma}{\sqrt{n}}$.
  - Sometimes, a corrected version is used: $\frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$ when the sample is a significant fraction of the population.
- The standard error is a measure of the variability of the sample mean.

The Central Limit Theorem

- The Central Limit Theorem (CLT) is one of the most important results in statistics.
- The CLT states that the sampling distribution of the sample mean will be approximately normally distributed for large sample sizes.
- The CLT is a fundamental result in statistics because it allows us to make inferences about the population based on the sampling distribution of the sample mean.
- The CLT is the reason why the normal distribution is so important in statistics.

The Central Limit Theorem

- The "golden rule" of the CLT is that with samples of size $n \geq 30$, the sampling distribution of the sample mean will be approximately normally distributed.
- This is very powerful, because it allows us to make inferences about the population based on the normal distribution (std. normal distribution).
- The CLT is the reason why the normal distribution is so important in statistics.

The Central Limit Theorem

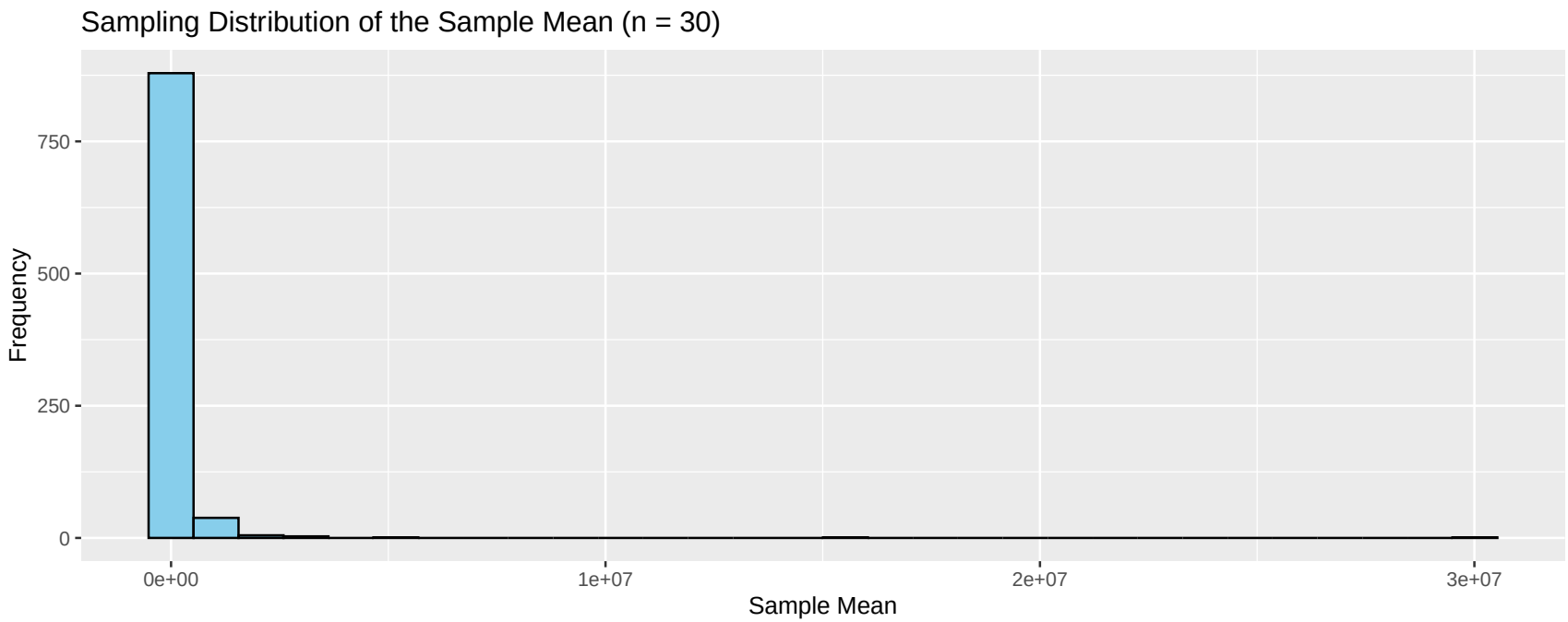
- Even if the underlying population distribution is not normal, the sampling distribution of the sample mean will be approximately normal for large sample sizes.
  - This is true for any population distribution, no matter how skewed or heavy-tailed.
- Do *NOT* confuse the population distribution with the sampling distribution of the sample mean.
  - The population distribution may be anything.
  - The sampling distribution of the sample mean will be approximately normal for large sample sizes.

Sample from SUPERCIAS with size 30

```r
set.seed(123)

# 1000 samples of size 30

sample_means_30 <- replicate(1000, {
  supercias_raw %>%
    sample_n(30) %>%
    summarise(mean_capital = mean(capital_suscrito)) %>%
    pull(mean_capital)
})

summary(sample_means_30)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
     625     6926    20083   161031    62647 30011533       72
```

Sampling Distribution of the Sample Mean for `capital_suscrito` with size 30



Stata implementations

- The `sample` command in Stata can be used to sample data.
  - `sample 10` will sample 10 percent of the data.
  - `sample 10, count` will sample 10 observations.
- Need to set the seed for reproducibility.
  - `set seed 123`
- Good to preserve the sample in a separate dataset.
  - `preserve` and `restore` commands can be used to do this.

Stata implementations

- An equivalent to `dnorm()` in Stata is the `normalden()` function.
  - `display normalden(1)` will give the density of the standard normal distribution at 1.
  - To modify the mean and standard deviation, use `normalden(mean, sd)`.
- For `pnorm()`, e.g., the CDF of the normal distribution, we can use the `normal()`, which automatically calculates the CDF for a standard normal distribution.
  - `display normal(1)` will give the CDF of the standard normal distribution at 1.
  - To modify the mean and standard deviation, get the Z-score and use `normal(Z)`

Stata implementations

- For `qnorm()`, e.g., the quantile function of the normal distribution, we can use the `invnormal()` function.
  - `display invnormal(0.95)` will give the 95th percentile of the standard normal distribution.
  - Transform the quantile to the Z-score to get the quantile for a normal distribution with a different mean and standard deviation.

Stata implementations

- To calculate a standard error in Stata, we can use the `egen` command.
  - `egen mean_capital = mean(capital_suscrito)`
  - `egen se_capital = std(capital_suscrito) / sqrt(_N)`
- Use the code above to calculate the standard error of the sample mean.