Aula 4: Inferência e Intervalos de Confiança

João Paulo Lazzarini Cyrino

Agosto de 2020

Inferência Estatística e sua Importância na Linguística

Uma das grandes coisas da estatística é nos permitir compreender propriedades de uma população a partir de apenas algumas observações de seus indivíduos: a **inferência estatística**. Por exemplo, podemos entrevistar algumas pessoas de uma cidade para saber preferências de todos os seus habitante (algo que ocorre em pesquisas eleitorais, por exemplo) ou testar medicamentos em algumas pessoas para saber o quão seguro (ou perigoso) ele será para toda a poupulação. Isso é muito bom porque, na maior parte das vezes, ou é muito caro obter os dados de toda a população ou isso é simplesmente impossível.

Como linguistas, quando estamos diante de dados linguísticos estamos sempre diante de um subconjunto dos dados da língua: dadas as propriedades combinatórias dos morfemas e sintagmas, as línguas possuem um quantidade infinita de dados. Por essa razão, na atividade de descrição linguística, sempre é necessário fazer algum tipo de inferência (mesmo que não estatística) sobre o comportamento da língua a partir dos dados que temos disponíveis. Dada também a dinamicidade das línguas, o mesmo ocorre com tipologia, já que as línguas sempre estão mudando a cada instante e nunca conseguiremos ter acesso a toda a população de línguas: sempre teremos que fazer um recorte, nem que seja todas as línguas do mundo em um determinado instante.

Para fazer inferência estatística precisamos, em primeiro lugar, de uma boa amostra. Essa amostra precisa ter um tamanho adequado e ser coletada de forma adequada. Amostras adequadas para inferência estatística são amostras de **probabilidade**, em que os dados são coletados aleatoriamente. Existem, basicamente, três tipos de amostras de probabilidade:

- **Amostra aleatória simples:** simplesmente coletar os dados aleatoriamente.
- **Amostra estratificada:** dividir a população em grupos (por exemplo, faixas etárias; ou famílias linguísticas) e coletar um número n de dados em cada grupo, aleatoriamente.
- **Amostra por conglomerado:** dividir a população em vários grupos e se coletam os grupos aleatoriamente.

Normalmente utilizaremos amostras aleatórias simples ou amostras estratificadas em nossos estudos.

A partir da amostra podemos estimar a média, porporção ou variância da população por meio da inferência estatística. Por exemplo, qual a média de sílabas nas palavras de uma língua? Ou, qual a proporção de objetos nulos que co-ocorrem com antecedentes animados? A partir dessas estimativas conseguimos ainda saber se essas médias e proporções são significativas para confirmar ou não hipóteses que levantamos a partir dos *testes de hipótese* estatísticos.

Nesta aula vamos aprender a fazer a estimativa de média, proporção e variância da população a partir de uma amostra. Fazemos isso com a técnica de **intervalos de confiança**. Antes de partir diretamente para os cálculos todos, vamos primeiro entender um pouco da teoria por trás da inferência estatística.

A teoria por trás da inferência estatística:

Existem diferentes formas de se fazer inferência estatística. Aqui aprendemos sobre inferência frequentista, que tem sido a mais comum dentro de estudos científicos (havendo também a inferência bayesiana, assunto

apra outro curso). A inferência frequentista se baseia fortemente em um teorema da teoria das probabilidades que se chama *Teorema do Limite Central*.

O Teorema do Limite Central prevê que, se retirarmos de uma população n amostras aleatórias com reposição, a média ou proporção dessas amostras tenderá à média/proporção da população conforme n tende ao infinito. Isso pode ser ilustrado com um exemplo: vamos criar uma população de 10000 bolinhas das quais 3000 são vermelhas e 7000 são azuis. Podemos fazer isso em R com a função `rep`:

```
pop <- c(rep("vermelha", 3000), rep("azul",7000))
mean(pop == "vermelha") # ver a proporção de bolinhas vermelhas na população
```

```
## [1] 0.3
```

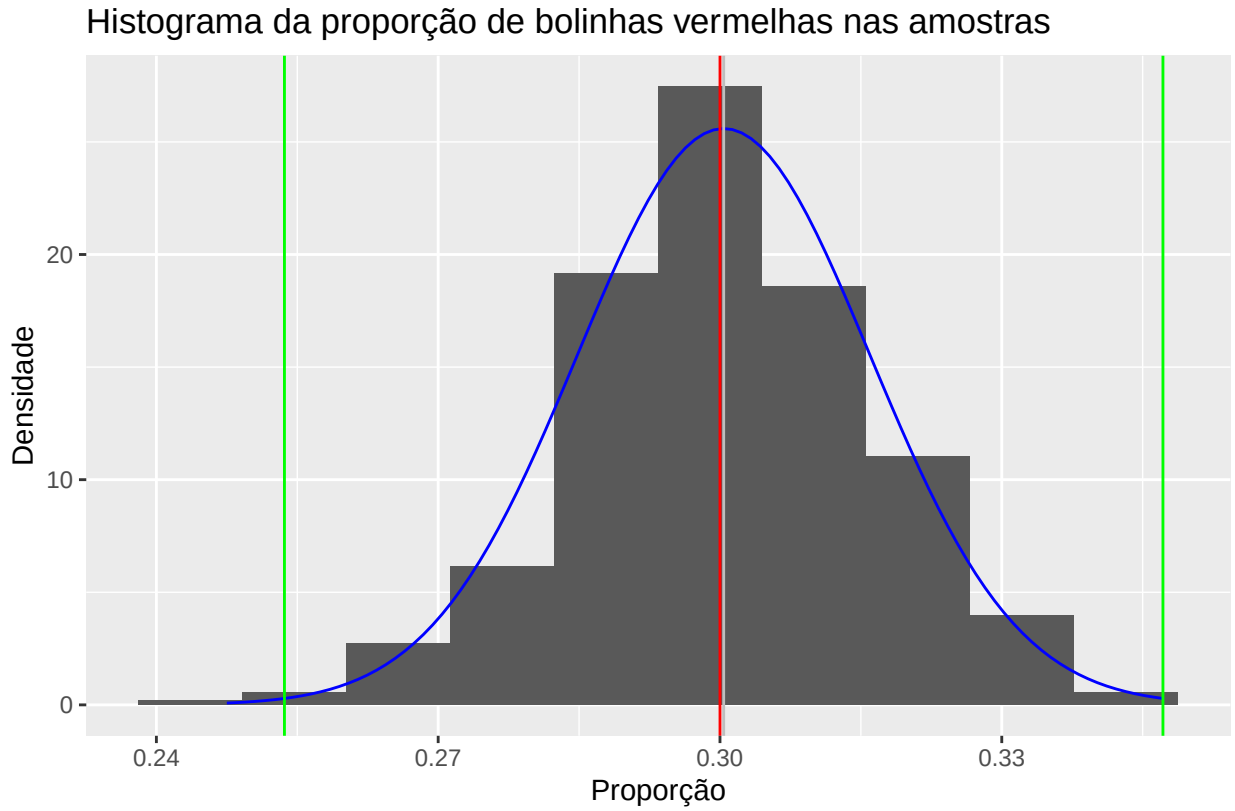
Criamos um vetor *pop* representativo de nossa população de bolinhas. Nele usamos a função `rep` para inserir 3000 itens com o valor "vermelha" e 7000 com o valor "azul". Medimos a proporção de bolinhas vermelhas em seguida.

Agora, vamos retirar amostras aleatórias dessa população. Essas amostras serão de 800 bolinhas, com reposição. Vamos anotar em um vetor a média de cada amostra. Para fazer isso em R usamos a função `sample`, que retira uma amostra aleatória a partir de uma população. Com a opção *replace=TRUE* essa amostra se derá com reposição. Aplicando a função `mean` à função `sample` temos a média/proporção da amostra. A função `replicate` replica uma função por n vezes e armazena os resultados em um vetor, vamos aplicá-la ao conjunto `mean(sample(...))` para obter nosso vetor de médias/proporções de amostras a partir da população. No caso trabalharemos com proporções, veja o código:

```
# Tomamos a proporção de bolinhas vermelhas em cada uma das 500 amostras de 800 indivíduos que
# retiramos, com reposição, a partir da população pop
amostras <- replicate(500, mean(sample(pop,800,replace=TRUE) == "vermelha"))
# Vejamos se a média dessas proporções se aproxima da proporção de bolinhas vermelhas na população (apr
mean(amostras)
```

```
## [1] 0.300395
```

Como vemos, o valor é bastante aproximado, o que mostra o Teorema do Limite Central em funcionamento! Uma outra previsão do Teorema do Limite Central é a de que as médias/proporções se distribuem em uma distribuição de probabilidades denominada **distribuição normal**. Percebemos isso olhando o histograma das proporções das amostras e vendo que seu formato se assemelha ao de um sino (em azul):



Essa curva em formato de sino é o que se chama em probabilidades de **distribuição normal**. Seguindo o Teorema do Limite Central, essa distribuição é característica das amostras: se tomarmos um grande número de amostras veremos que suas médias ou proporções seguirão essa distribuição. Como é uma distribuição muito conhecida, podemos utilizá-la para estimar o comportamento da população a partir de uma única amostra ao invés de um grande número de amostras.

Sabe-se que, na distribuição normal, cerca de 99,7% das médias/proporções de qualquer amostra estão dentro de uma faixa centrada na média/proporção da população (linha vermelha do gráfico) com 3 desvios padrões da população para baixo ou para cima (linhas verdes do gráfico). Ou seja, você tem 99,7% de chances de errar a média/proporção da população por três desvios padrões da população apenas. Se você achar que 3 desvios padrões é muito, pode considerar que terá 95% de chances de errar por 2 desvios padrões, ou 68% de chances de errar por apenas 1 desvio padrão. Essas chances e esses erros em função de desvios padrões são a base do que entendemos por intervalos de confiança: temos, por exemplo, um **nível de confiança** de 99% de que nossa média/proporção está numa **margem de erro** de 3 desvios padrões da média populacional.

O que fizemos até agora foi colher 500 amostras a partir de uma população de 10000 bolinhas e ver que, em boa parte das amostras, as proporções de bolinhas vermelhas em cada amostra fica bem próxima da proporção da população: por conta da distribuição normal, sabemos que 99% delas tem um erro de até 0.05 (3 vezes o desvio padrão de cerca de 0.016) em relação à proporção da população. Ou seja, se pegarmos uma amostra qualquer de 800 bolinhas temos 99% de chances de ter uma proporção de bolinhas vermelhas entre 0.25 e 0.35.

O **intervalo de confiança** é, portanto, composto pelo nível de confiança e pela margem de erro. Intervalos de confiança são muito importantes porque eles mostram o quanto nossas observações a partir de uma amostra são confiáveis. E, como o comportamento das amostras segue o Teorema do Limite Central, podemos fazer cálculos para estabelecer o intervalo de confiança usando uma única amostra.

Calculando o Intervalo de Confiança

Na seção anterior tivemos que levantar 500 amostras de 800 indivíduos para poder ter um intervalo de confiança sobre qualquer uma dessas 500 amostras. Fazer isso na prática não faz nenhum sentido, já que envolve coletar uma imensidão de dados o que é, muitas vezes, caro e impraticável.

Porém, como sabemos que as amostras aleatórias se comportam de acordo com o Teorema do Limite Central, podemos utilizar as propriedades matemáticas da distribuição normal para calcular o intervalo de confiança da média ou proporção de nossa amostra. Matematicamente há duas formas de se fazer esse cálculo, a depender do tamanho da amostra e se conhecemos a variância populacional.

Se conhecemos a amostra populacional ou temos 30 ou mais indivíduos em nossa amostra podemos construir o intervalo de confiança a partir da distribuição normal padrão. Em um curso de estatística ou em tempos mais antigos, teríamos que estabelecer o nível de confiança de nosso estudo (por exemplo, 95%) e buscar em uma tabela o escore-z para esse nível de confiança. Com esse escore-z calculamos a margem de erro de uma média da seguinte forma:

$$E = z\frac{\sigma}{\sqrt{n}}$$

A letra σ refere-se ao o desvio-padrão populacional, mas, se tivermos mais de 30 indivíduos em nossa amostra podemos substituir pelo desvio-padrão amostral s. n é o tamanho da amostra e z é o escore-z.

Podemos procurar o escore-z em R com a função `qnorm`. Para utilizar essa função precisamos entender que, quando queremos um nível de confiança de 95%, por exemplo, queremos descartar 5% da distribuição normal. Como descartamos valores dos dois lados da distribuição, queremos descartar 2,5% de cada lado. Dessa forma devemos pedir na função `qnorm` o valor $1 - 2,5$ ou $0,975$, que equivale a aproximadamente $1,95$:

```
qnorm(.975)
```

```
## [1] 1.959964
```

Supondo que temos uma amostra de 200 indivíduos, média 42 e desvio-padrão 8, temos a seguinte margem de erro para um nível de confiança de 95%:

```
E <- qnorm(.975)*(8/sqrt(200))
E
```

```
## [1] 1.108723
```

Para ter os valores do intervalo de confiança subtraímos a margem de erro E da média, para o valor menor, e adicionamos para o valor maior:

```
ic <- c(42-E, 42+E)
ic
```

```
## [1] 40.89128 43.10872
```

Para proporções a fórmula é um tanto diferente:

$$E = z\sqrt{\frac{p(1-p)}{n}}$$

Nesse caso, p é a proporção. Vamos supor que, em uma amostra de 80 indivíduos tenhamos uma proporção de 0,7 de alguma coisa que queremos medir. Nesse caso, calculamos o intervalo de confiança (considerando o nível de confiança de 95%) da seguinte forma:

```
E <- qnorm(.975)*sqrt((0.7*0.3)/80)
E
```

```
## [1] 0.1004183
```

O intervalo de confiança será então:

```
ic <- c(0.7-E, 0.7+E)
ic
```

```
## [1] 0.5995817 0.8004183
```

Quando nossa amostra é menor que 30 indivíduos, as coisas são um pouco diferentes. Utilizamos, ao invés da distribuição normal, a distribuição-t. A distribuição-t é semelhante à distribuição normal quando temos mais de 30 indivíduos, porém ela tende a produzir intervalos de confiança mais largos quando a amostra é menor que 30, garantindo mais precisão. Os cálculos são como os cálculos da distribuição normal, mas, ao invés de termos que buscar o escore-z, buscamos o escore-t. O escore-t depende, além do nível de confiança desejado, dos graus de liberdade da amostra (tamanho da amostra menos 1). Dessa forma, para uma amostra de 25 indivíduos precisamos buscar um escore-t para 24 graus de liberdade. A função `qt` em R produz escores-t para nós: basta fornecer o nível de confiança e os graus de liberdade, nesta ordem:

```
# Escore-t para 25 indivíduos e 90% de confiança:
t <- qt(.95, 24)
t
```

```
## [1] 1.710882
```

Para calcular o intervalo de confiança para 90% com uma amostra de 25 indivíduos, média 12 e desvio-padrão 1 temos:

```
# Cálculo da margem de erro
E <- t*1/sqrt(25)
# Cálculo do intervalo de confiança:
ic <- c(12-E, 12+E)
ic
```

```
## [1] 11.65782 12.34218
```

Muito bem, isso é como as coisas são feitas na teoria. No dia-a-dia de nossas pesquisas as coisas são mais práticas: R possui duas funções que tornam bastante prático o cálculo dos intervalos de confiança. Introduziremos essas funções adiante. No entanto alerto para a necessidade de se conhecer a teoria em duas situações:

- Quando queremos calcular o tamanho da amostra que precisamos coletar para obter uma determinada margem de erro.
- Quando apenas temos disponível um valor de média e desvio-padrão, mas não os dados propriamente.

O cálculo do tamanho da amostra será ilustrado a seguir. Pede-se bastante atenção pois é algo extremamente útil quando estamos planejando nossa pesquisa.

Determinando o tamanho da amostra

É importante que saibamos o tamanho da amostra que devemos coletar para ter uma margem de erro desejável. Para isso basta que tenhamos uma amostra menor em que possamos determinar a variância/desvio-padrão ou proporção. A partir disso calculamos o tamanho da nova amostra com as seguintes fórmulas.

Para médias:

$$n = (\frac{z\sigma}{E})^2$$

Para proporções:

$$n = p(1-p)(\frac{z}{E})^2$$

Supondo que você esteja estudando um fenômeno que ocorre numa proporção de 0,3 da sua mostra. Nesse caso, para ter uma margem de erro de 5% (0,05) em um nível de confiança de 95%, calculamos o tamanho da amostra da seguinte forma:

```
n <- 0.3*0.7*(qnorm(.975)/.05)^2
n
```

```
## [1] 322.6825
```

Precisaremos, portanto, de no mínimo 323 observações para ter uma margem de erro de, ao menos, 5%.

Nossa amostra será maior conforme:

- A variância dos dados ou a diferença de proporções for maior.
- O nível de confiança desejado for maior.
- A margem de erro desejada for menor.

Vamos calcular o tamanho da amostra supondoo que quiséssemos um nível de confiança de 99% e uma margem de erro de 1%:

```
n <- 0.3*0.7*(qnorm(.995)/.01)^2
n
```

```
## [1] 13933.28
```

Veja como o número de indivíduos já salta para 13934 para obtermos um intervalo de confiança extremamente preciso.

De forma geral, é possível ter bons resultados com trabalhos quantitativos em linguística tendo níveis de confiança de 95% e margens de erro de 5%.

A função `t.test`:

Na pasta deste curso temos o objeto R *silabas* (*silabas.rds*). Este objeto é um vetor com a quantidade de sílabas de cada palavra encontrada em um conto na língua portuguesa. Ele servirá para termos uma noção da media de sílabas que encontramos nas palavras do português. É uma amostra de 878 indivíduos/observações. Abaixo calculamos a média e o desvio padrão desses dados:

```
# carregar o arquivo silabas.rds no objeto silabas
silabas <- readRDS("silabas.rds")
# mostrar um vetor com a média e desvio padrão de silabas, respectivamente:
c(mean(silabas), sd(silabas))
```

```
## [1] 1.9271071 0.9169224
```

Agora vamos calcular a margem de erro e o intervalo de confiança que conseguimos obter com essa amostra:

```
# Margem de Erro:
E <- qnorm(.975)*sd(silabas)/sqrt(length(silabas))
# Intervalo de confiança:
ic <- c(mean(silabas)-E, mean(silabas)+E)
ic
```

```
## [1] 1.866457 1.987757
```

Uma outra forma de obter o intervalo de confiança é simplesmente utilizando a função `t.test`. Essa função fará o cálculo com base em uma distribuição-t. Porém, como dissemos anteriormente, a ditribuição-t tende a ter o mesmo comportamento da distribuição normal para graus de liberdade acima de 29. Dessa forma, obter o intervalo de confiança pela distribuição-t é uma ótima aproximação. Para fazê-lo simplesmente fornecemos um vetor de dados para a função. No caso, o vetor é *silabas*:

```r
t.test(silabas)
```

```
##
##  One Sample t-test
##
## data:  silabas
## t = 62.276, df = 877, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.866373 1.987841
## sample estimates:
## mean of x
##  1.927107
```

Por padrão o nível de confiança é 95%. Podemos alterá-lo para 99%, por exemplo, da seguinte forma:

```r
t.test(silabas, conf.level=.99)
```

```
##
##  One Sample t-test
##
## data:  silabas
## t = 62.276, df = 877, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  1.847225 2.006989
## sample estimates:
## mean of x
##  1.927107
```

Se quisermos salvar em uma variável o intervalo de confiança:

```r
ic <- t.test(silabas)$conf.int
ic
```

```
## [1] 1.866373 1.987841
## attr(,"conf.level")
## [1] 0.95
```

A função `prop.test`:

Além da função `t.test` para calcular intervalos de confiança de médias, temos a função `prop.test` para intervalos de confiança de proporções. Ela funciona de forma um pouco diferente da função `t.test`. Aqui fornecemos à função o número de sucessos e o tamanho da amostra. Por exemplo, em *silabas* temos 358 palavras com duas sílabas e nossa amostra tem 878 observações. Dessa forma:

```r
prop.test(358,878)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  358 out of 878, null probability 0.5
## X-squared = 29.523, df = 1, p-value = 5.526e-08
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3751502 0.4411573
## sample estimates:
```

```
##         p
## 0.4077449
```

De resto, a função é muito semelhante a `t.test`:

```
# salvar o intervalo de confiança em um nível de 99% na variável ic
ic <- prop.test(358,878,conf.level=.99)$conf.int
# visualizar ic
ic
```

```
## [1] 0.3653169 0.4515750
## attr(,"conf.level")
## [1] 0.99
```