Slides Semana 7

Inferência Estatística

Introdução à Inferência Estatística

A Estatística é uma ciência que tem como objetivo a tomada de decisão em situações de incerteza. Esta ciência divide-se basicamente em duas partes. A primeira parte é conhecida como **Estatística Descritiva** que trata da coleta, organização e descrição de dados. A segunda é a **Estatística Inferencial** que se preocupa em fazer afirmações e/ou testar hipóteses sobre características numéricas em situações de incerteza.

Para iniciar o estudo da Estatística Inferencial é necessário compreender os seguintes conceitos básicos:

- **População:** A população é um conjunto formado por todos os elementos cujas características desejamos conhecer.
- **Amostra:** A Amostra é apenas uma parte da população, ou seja, é um subconjunto da população cujas características serão medidas.

Dois outros conceitos estreitamente relacionados com os de **População** e **Amostra** são os de **Parâmetro** e **Estatística**.

- **Parâmetro:** É uma característica numérica da população.
- **Estatística:** É uma característica numérica da amostra que será usada para extrair informações sobre a população.

Exemplo

- *População*: Os eleitores da cidade de Salvador
- *Amostra*: 650 eleitores escolhidos aleatoriamente (ao acaso)
- *Característica de interesse*: percentual de eleitores que planejam votar num candidato A nas próximas eleições.

Exemplo

- *População*: população acima de 15 anos na cidade de Salvador.
- *Amostra*: 200 pessoas com mais de 15 anos.
- *Características de interesse*:
- percentual de bebedores de cerveja.
- dentre os bebedores de cerveja, quantos são homens?
- dentre os bebedores de cerveja, quantos preferem Brahma?
- dentre os bebedores de Brahma, quantas cervejas eles tomam por semana e a que classe social eles pertencem?
- Existe alguma relação entre as variáveis Marca de Cerveja consumida e Classe Social?

Existe uma infinidade de características populacionais de interesse ao realizar uma pesquisa.

O principal objetivo da Inferência Estatística é fazer afirmações e/ou testar hipóteses sobre essas características, tomando como base as informações de dados amostrais e levando-se em consideração uma margem de erro a um nível de confiança estabelecido.

Características numéricas como **média**, **variância** e **proporção** são consideradas parâmetros se obtidas pelo uso de dados populacionais e não apresentam incerteza sobre seu real valor.

Quando estas características são baseadas em dados amostrais (dados de uma parte da população) tem-se as estatísticas, as quais apresentam diferentes valores (*apresenta variabilidade*) se obtidas a partir de diversas amostras.

Em resumo, é a partir de uma amostra que coletamos informações que nos permitirão aprender alguma coisa interessante sobre a população de interesse.

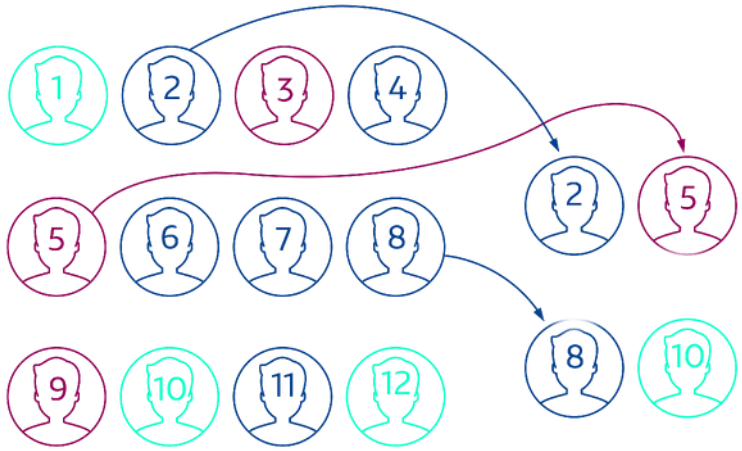
Parâmetro	Estatística
Total de elementos na população:	Total de elementos na amostra:
N	n
Média Populacional:	Média Amostral:
$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$	$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
Variância Populacional:	Variância Amostral:
$S^2 = \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$	$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$
Proporção Populacional:	Proporção Amostral:
$P = \frac{A}{N}$	$\overline{p} = \frac{a}{n}$
em que A é o $n^{\underline{o}}$ de elementos da população que possuem uma certa característica de interesse.	em que a é o $n^{\underline{o}}$ de elementos da amostra que possuem uma certa característica de interesse.

Como Selecionar uma Amostra: Vários motivos levam a necessidade de se observar apenas uma parte da população (amostra), como, por exemplo: a falta de tempo, de recursos financeiros e/ou humanos ou, ainda, a inacessibilidade a toda população.

"A amostra deve ser obtida através de procedimentos científicos que permitam fazer inferências adequadas sobre a população. A maneira de se obter a amostra é tão importante, e existem tantos modos de fazê-lo, que esses procedimentos constituem especialidades dentro da Estatística, sendo Amostragem e Planejamento de Experimentos as duas mais conhecidas."(Estatística Básica - Morettin, P.A. e Bussab W.O, - Ed. Saraiva.)

Os procedimentos científicos de obtenção de dados amostrais podem ser divididos em três grandes grupos:

Levantamentos Amostrais: Neste tipo de procedimento a amostra é obtida a partir de uma população bem definida, por meio de processos bem protocolados e controlados pelo pesquisador.



Basicamente, existem dois tipos de levantamentos amostrais: Probabilísticos e Não-Probabilísticos.

- **Amostragem Probabilística:** Neste tipo de amostragem a probabilidade de cada elemento pertencer a amostra é conhecida e diferente de zero.

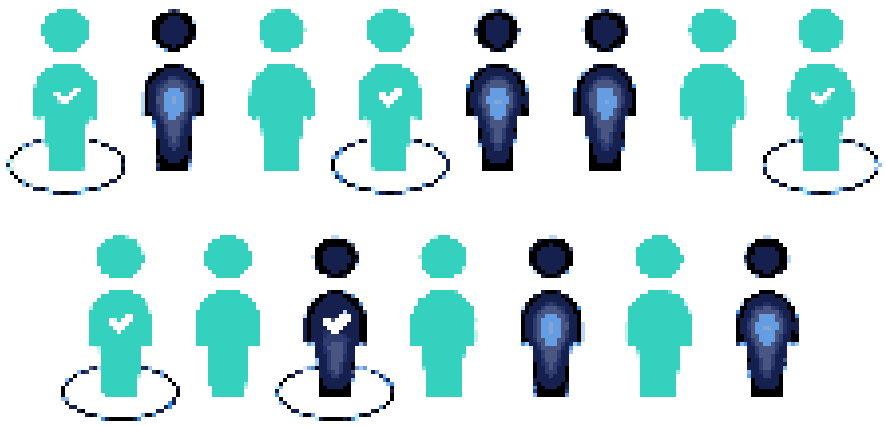
A amostragem probabilística implica em sorteio com regras bem determinadas, cuja realização só será possível se a população for finita e totalmente acessível.

A utilização de uma amostragem probabilística é a melhor recomendação que se deve fazer no sentido de se garantir a representatividade da amostra, pois o acaso ou a aleatoriedade será o(a) único(a) responsável por eventuais discrepâncias entre as características da população e da amostra, o que é levado em consideração pelos métodos de análise da Estatística Inferencial.

Os principais tipos de amostragem probabilística são:

1 - **Amostragem Aleatória Simples (AAS)**: A amostragem aleatória simples é uma técnica de coleta de dados amostrais que equivale a um sorteio lotérico. Nela, todos os elementos da população *finita* têm igual probabilidade de pertencer à amostra, e todas as possíveis amostras têm igual probabilidade de ocorrer.

O processo da amostragem aleatória simples exige que se atribuam números consecutivos às unidades da população e proceda-se a um sorteio.

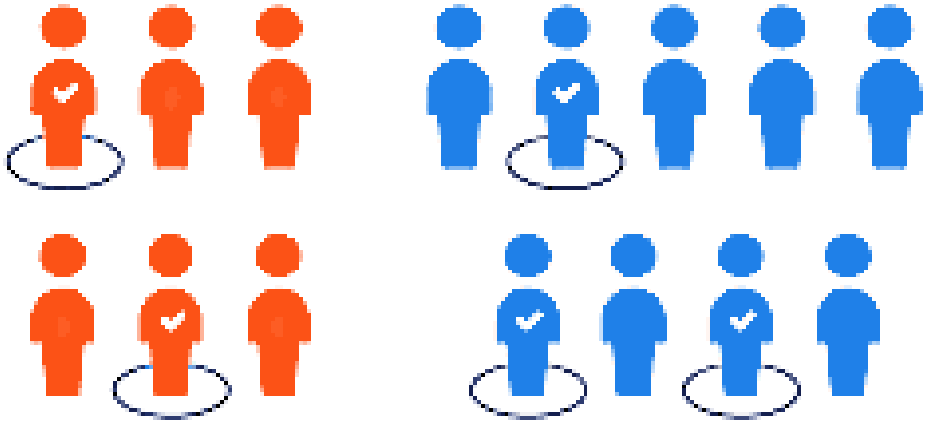


Colocando-se todos os números em um recipiente, por exemplo, e retirando um número de cada vez, até que os n elementos da amostra sejam selecionados.

O procedimento de sorteio pode ser **com** ou **sem** a **reposição** de cada elemento diante de cada sorteio; sendo que; a **AAS com reposição** apresenta uma **vantagem teórica**, por conduzir a obtenção de *observações independentes*.

É importante ressaltar, que o procedimento de sorteio não é prático para uma população muito grande; busca-se, então, simular tal sorteio, o que é feito pelo uso de uma tabela de dígitos pseudo-aleatórios ou pelo uso de funções aleatórias existentes em programas computacionais

2 - **Amostragem Estratificada**: Muitas vezes a população se divide em Sub-populações ou Estratos, sendo razoável supor que, de estrato para estrato, a(s) variável(is) de interesse apresente(m) comportamento(s) substancialmente diverso(s) (comportamento heterogêneo), tendo, entretanto, comportamento(s) razoavelmente homogêneo(s) dentro de cada estrato.

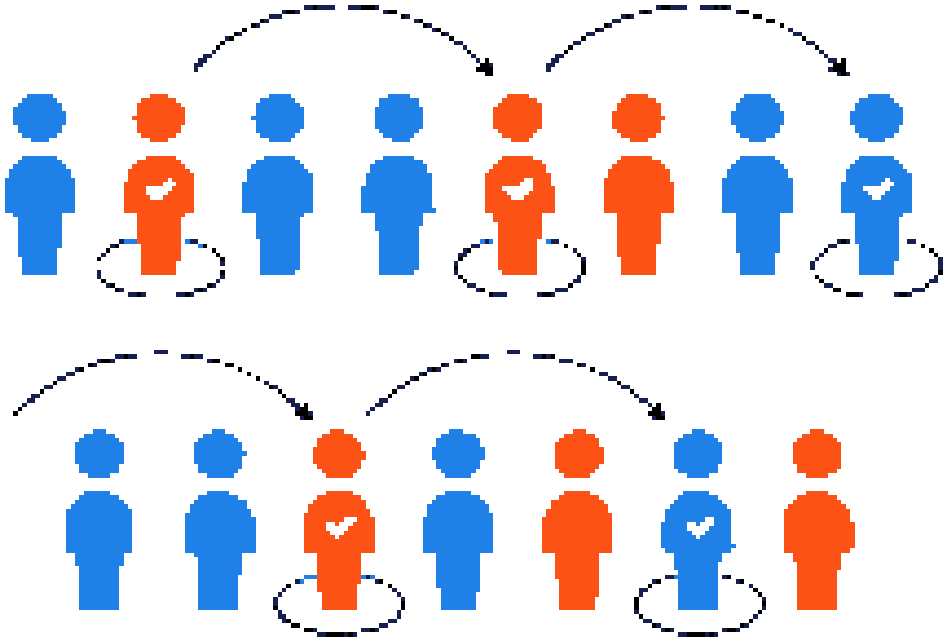


Em tais casos, se o sorteio dos elementos da amostra for realizado sem levar em consideração a existência dos estratos, pode acontecer que os diversos estratos não sejam convenientemente representados na amostra, a qual seria mais influenciada pelas características da variável nos estratos mais favorecidos pelo sorteio.

Evidentemente, a tendência à ocorrência de tal fato será tanto maior quanto menor o tamanho da amostra. Para evitar isso, pode-se adotar uma amostragem estratificada.

A amostragem estratificada consiste em identificar os estratos (grupos distintos da população), e em especificar quantos elementos da amostra serão retirados em cada estrato através de uma amostragem aleatória simples. É costume considerar três tipos de amostragem estratificada: uniforme, proporcional e ótima.

3 - **Amostragem Sistemática**: Uma amostragem é sistemática quando a retirada dos elementos da população é feita periodicamente, sendo o *intervalo de seleção* calculado, para uma populaçõa finita, por meio da divisão do tamanho da população, N, pelo tamanho da amostra a ser selecionada, n.



- **Amostragem Não-Probabilística:** Na amostragem não-probabilística não é possível calcular a probabilidade de cada elemento pertencer a amostra.

Este tipo de amostragem é muitas vezes empregado em trabalhos estatísticos, por simplicidade ou por impossibilidade de se obterem amostras probabilísticas, como seria desejável.

Como em muitos casos os efeitos da utilização de uma amostragem não-probabilística podem ser considerados equivalentes aos de amostragem probabilística, resulta que os processos não-probabilísticos de amostragem têm também sua importância. Apresentamos a seguir alguns casos de amostragem não-probabilística.

- **Inacessibilidade a toda a população** Essa situação ocorre com muita freqüência na prática. Somos então forçados a colher a amostra na parte da população que nos é acessível.

Surge aqui, portanto, uma distinção entre população-objeto e população amostrada.

A população-objeto é aquela que temos em mente ao realizar o trabalho estatístico.

Apenas uma parte dessa população, porém, está acessível para que dela retiremos a amostra. Essa parte é a população amostrada.

Se as características da(s) variável(is) de interesse forem as mesmas na população-objeto e na população amostrada, então esse tipo de amostragem equivalerá a uma amostragem probabilística.

Uma situação muito comum em que ficamos diante da inacessibilidade a toda a população é o caso em que parte da população é ainda hipotética.

Assim, por exemplo, seja a população que nos interessa constituída por todas as peças produzidas por certa máquina. Ora, mesmo estando a máquina em funcionamento normal, existe uma parte da população que é formada pelas peças que ainda vão ser produzidas.

Ou, então, se nos interessar a população de todos os portadores do vírus HIV, estaremos diante de um caso semelhante.

Deve-se notar que, em geral, estudos realizados com base nos elementos da população amostrada terão, na verdade, seu interesse de aplicação voltado para os elementos restantes da população-objeto.

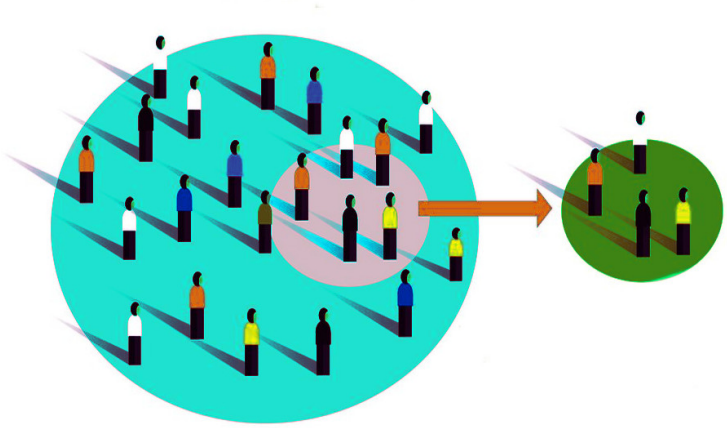
Esse fato realça a importância de se estar convencido de que as duas populações podem ser consideradas como tendo as mesmas características.

O presente caso de amostragem não-probabilística pode ocorrer também quando, embora se tenha a possibilidade de atingir toda a população, retiramos a amostra de uma parte que seja prontamente acessível. Assim, se fôssemos recolher uma amostra de um monte de minério, poderíamos por simplificação retirar a amostra de uma camada próxima à superfície exterior do monte, pois o acesso às porções interiores seria problemático.

Amostragem a esmo ou sem norma É a amostragem em que o amostrador, para simplificar o processo, procura ser aleatório sem, no entanto, realizar propriamente o sorteio usando algum dispositivo aleatório confiável.Poderia ser observar uma amostra de hotéis de uma determinada cidade, e que, por um motivo qualquer, não possuímos a listagem dos hotéis.

Então, poderíamos proceder a uma amostragem simplesmente a esmo ou ao acaso, buscando hotéis localizados em diferentes bairros, de diferentes tamanhos e estrelas, e caso tivéssemos interessados em pesquisar hotéis numa única rua, procuraríamos observar hotéis tanto do lado direito, quanto do lado esquerdo da rua, e evitaríamos observar hotéis que fossem vizinhos.

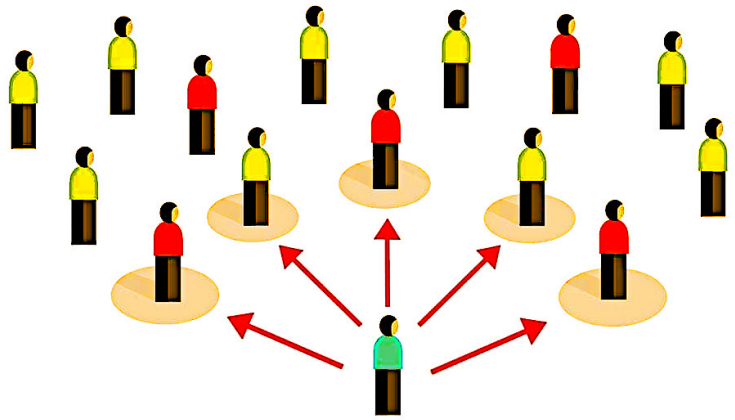
Os resultados da amostragem a esmo são, em geral, equivalentes aos de uma amostragem probabilística se a população é homogênea (elementos com características bastante semelhantes) e se não existe a possibilidade de o amostrador ser inconscientemente influenciado por alguma característica dos elementos da população.



Amostragens intencionais Enquadram-se aqui os diversos casos em que o amostrador *deliberadamente* escolhe certos elementos para pertencer à amostra, por julgar tais elementos bem representativos da população. O perigo desse tipo de amostragem é obviamente grande, pois o amostrador pode facilmente se equivocar em seu pré-julgamento.

Apesar disso, o uso de amostragens intencionais, ou parcialmente intencionais, é bastante freqüente, ocorrendo em vários tipos de situações reais. Exemplos freqüentes ocorrem na área empresarial, em que os administradores de uma empresa desejam que determinados elementos de uma população não fiquem fora da amostra.

Devemos, chamar a atenção que esta intencionalidade pode ser usada tanto para garantir a representatividade da amostra, como também para induzir resultados. Um exemplo deste último objetivo seria a intencionalidade de um político para que pertençam a amostra, uma ou mais comunidades em que ele suspeita que a maioria dos eleitores são favoráveis a sua candidatura.



Planejamento de Experimentos No Planejamento de Experimentos, os procedimentos para a obtenção da amostra tem a interferência por parte do pesquisador, cujo principal objetivo é analisar o efeito de uma ou mais variáveis sobre outra.

Na prática de pesquisas científicas, é de fundamental importância o planejamento do experimento, pois na falta de um planejamento adequado, análises posteriores podem até mesmo serem impossíveis de realizar.

Levantamentos Observacionais Este tipo de procedimento ocorre com bastante freqüência em pesquisas sociais e econômicas, pelo fato do pesquisador não ter controle algum sobre as informações, exceto eventualmente sobre possíveis erros grosseiros. Por exemplo, em pesquisa de séries de dados temporais, os dados efetivamente já ocorreram dentro uma infinidade de possibilidades.

Como podemos ver, há diferentes maneiras pelas quais as amostras podem ser selecionadas, cada qual com vantagens e desvantagens.

É importante ressaltar que a **definição do tamanho da amostra** a ser retirada da população é um outro problema associado à amostragem. O tamanho amostral deve minimizar os custos operacionais da amostragem e será tanto maior quanto for a variabilidade das características populacionais a serem estudadas.

Para concluir, formalizaremos a definição de amostra aleatóra simples com reposição (AAS), por se tratar da amostragem que leva à observações independentes, fato este que facilitará o desenvolvimento teórico para a solução de problemas da inferência estatística.

Distribuição amostral

As distribuições amostrais são o fundamento da estatística inferencial, pois é a partir do conhecimneto das distribuições amostrais que se poderá entender a relação entre as estatísticas e os parâmetros.

- **Amostra Aleatória Simples - AAS:** Variáveis aleatórias $X_1, X_2, \ldots X_n$ constituem uma **amostra aleatória simples** de tamanho n, ou simplesmente amostra aleatória (a.a.) de uma variável aleatória (v.a) X, quando satisfazem as seguintes condições:

1. As variáveis aleatórias $X_1, X_2, \ldots X_n$ são independentes, e

2. Cada uma das variáveis aleatórias $X_i, i = 1, 2, \ldots, n$ tem a mesma distribuição de probabilidade da variável X.

- **Distribuição Amostral** é a distribuiçõa de probabilidade que re-presenta o comportamento de uma estatística ao realizar repetidas amostragens.

Observação: A distribuição amostral de uma estatística (ou estimador) depende da distribuição de probabilidade da população, do tamanho da amostra e do método de seleção da amostra.

Distribuição Amostral da Média ($\overline{X}$)

A **média amostral** é um dos estimadores (estatística) mais utilizados na estatística pelo fato do grande interesse em se **estimar** o **valor médio** de uma certa **característica populacional**.

Para entender a relação entre a média amostral ($\overline{X}$) e a média populacional (μ), precisamos pensar na distribuição da média amostral, que descreve o conjunto de todas as possíveis médias amostrais extraídas de uma população.

Teorema: Seja X uma variável aleatória com distribuição de probabilidade **qualquer** tal que $E(X) = \mu$ e $Var(X) = \sigma^2$. Se $X_1, X_2, \ldots, X_n$ é uma amostra aleatória de X e

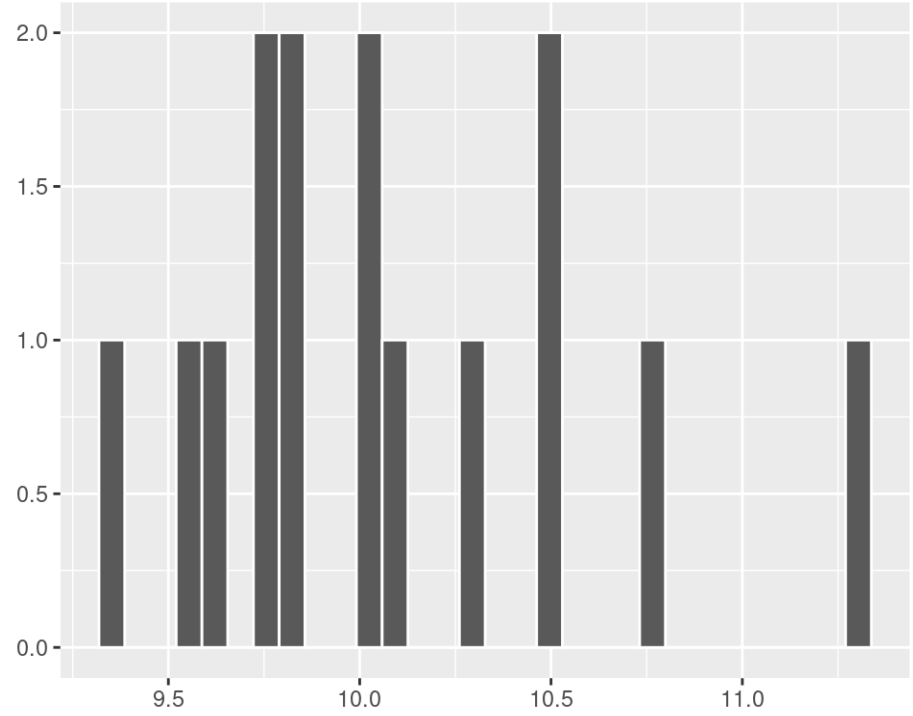
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

é a média amostral, então, i) $E(\overline{X}) = \mu$, ii) $Var(\overline{X}) = \frac{\sigma^2}{n}$ e se n é sufientemente grande (geralmente $n \geq 30$), pelo *Teorema Central do Limite*, tem-se que:

$$\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

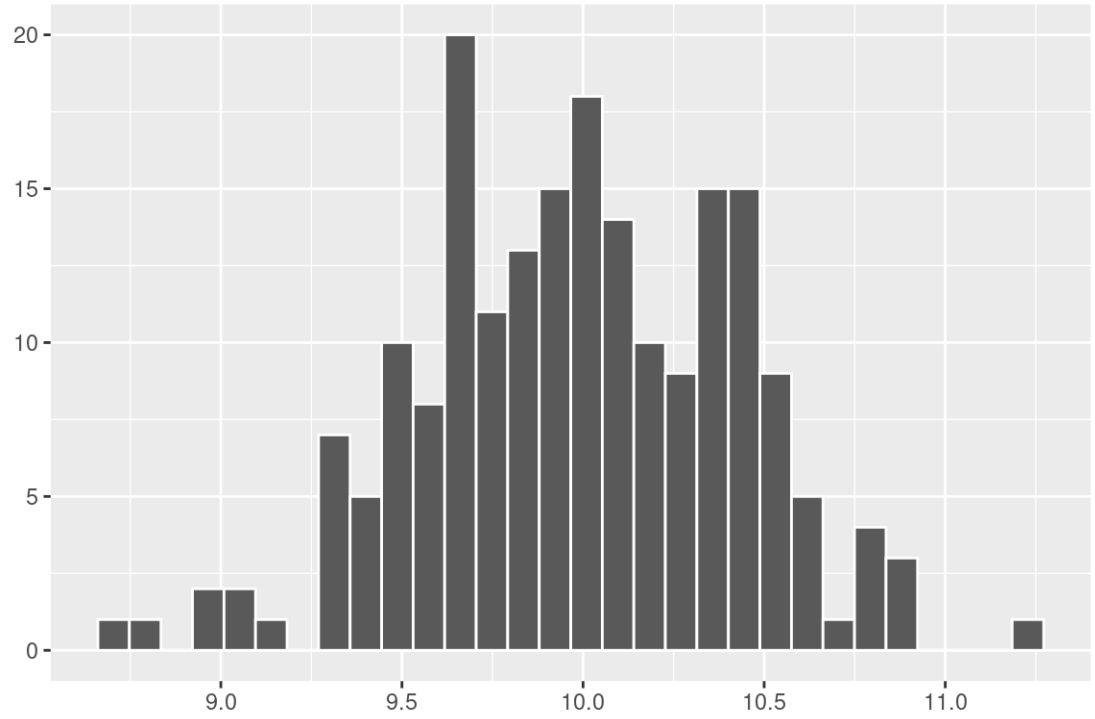
em que $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$ é a média populacional (parâmetro) e $\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$ é a variância populacional (parâmetro), com N representando o total de elementos da população.

Considere uma população normal com média $\mu = 10$ e variância $\sigma^2 = 4$. Vamos realizar um estudo de simulação para a distribuição da média amostral considerando amostras de tamanho 20 dessa população. Primeiramente, considere que são retiradas 15 amostras de tamanho 20 dessa população.



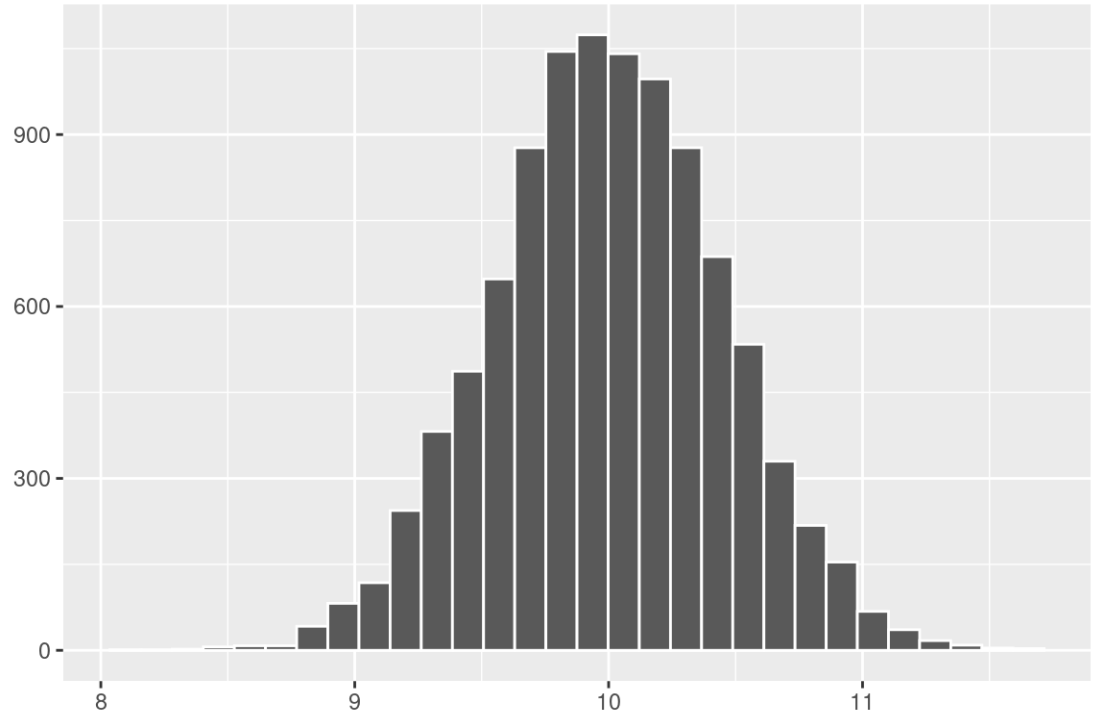
Nessa simulação obtivemos $\bar{x} = 10.09$ e $s = 0.52$.

Suponha agora que façamos o mesmo processo, porém ao invés de considerarmos 15 amostras de tamanho 20, consideramos 200 amostras.



Agora, obtivemos $\bar{x} = 9.98$ e $s = 0.44$.

Realizando o mesmo experimento, porém agora considerando 10000 amostras de tamanho 20, a distribuição da média amostral pode ser vista segundo o histograma abaixo.



Para este caso, a média das médias amostrais foi $\bar{x} = 10$ e o desvio padrão foi $s = 0.45$. Então, empiricamente, podemos perceber que a distribuição da média amostral se aproxima de uma distribuição normal com média $\mu = 10$ e desvio padrão $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{20}} = 0,4472$.

Distribuição da Variância Amostral

Vimos que a estatística

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

é um estimador não viciado da variância (ou seja que se aproxima do verdadeiro valor da variância populacional) σ^2, portanto

$$E(S^2) = \sigma^2$$

. Vejamos o comportamento a distribuição amostral de S^2. e para isso precisamos estudar a distribuição qui-quadrado.

Distribuição Qui-quadrado

Se X é uma variável aleatória com densidade

$$f_X(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right)}\left(\frac{1}{2}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, k > 0, x > 0,$$

em que $\Gamma(w) = \int_0^\infty x^{w-1} e^{-x} dx, w > 0$. Então, X tem uma distribuição qui-quadrado com k graus de liberdade, onde o parâmetro k é um número inteiro.

Para entender a ideia de graus de liberdade, consideremos um conjunto de dados qualquer. Graus de liberdade é o número de valores deste conjunto de dados que podem variar após terem sido impostas certas restrições a todos os valores.

Por exemplo, consideremos que 10 estudantes obtiveram em um teste média 8. Assim, a soma das 10 notas deve ser 80 (restrição). Portanto, neste caso, temos um grau de liberdade de $10 - 1 = 9$, pois 9 notas podem variar livremente desde que a soma seja 80, no entanto 1 nota sempre será [80-(soma das 9 outras notas)].

Se as variáveis aleatórias X_i, $i = 1, 2, \ldots, n$ são independentes e normalmente distribuídas com médias μ_i e variâncias σ_i^2, isto é $X_i \sim N(\mu_i, \sigma_i^2)$, então

$$U = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma^2} \right)^2$$

tem uma distribuição qui-quadrado com n graus de liberdade.

Além disso, se $X_1, \ldots, X_n$ é uma a.a. de uma distribuição normal padrão, então, valem as seguintes propriedades:

1. $\bar{X}$ e $\sum_{i=1}^{n} (X_i - \bar{X})^2$ são independentes;
2. $\sum_{i=1}^{n} (X_i - \bar{X})^2$ tem uma distribuição qui-quadrado com $n - 1$ graus de liberdade.

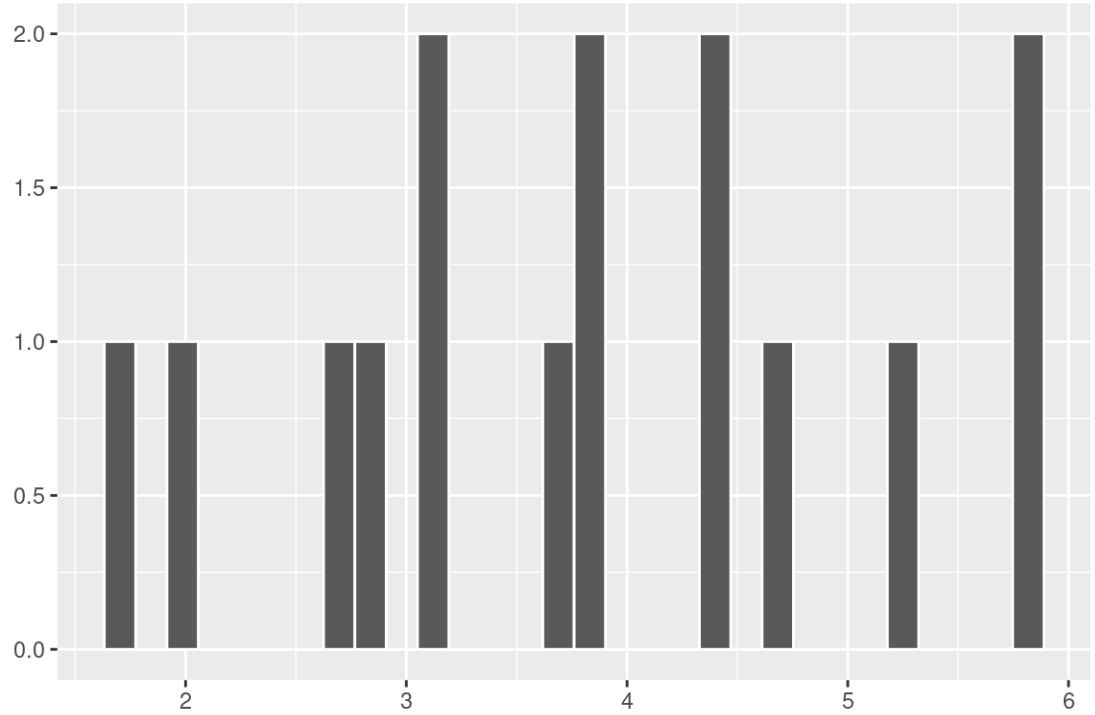
Assim, chegamos que se S^2 é a variância amostral de uma amostra aleatória $X_1, \ldots, X_n$ de uma distribuição normal com média μ e variância σ^2, então

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

ou seja, U tem uma distribuição qui-quadrado com $n - 1$ graus de liberdade.

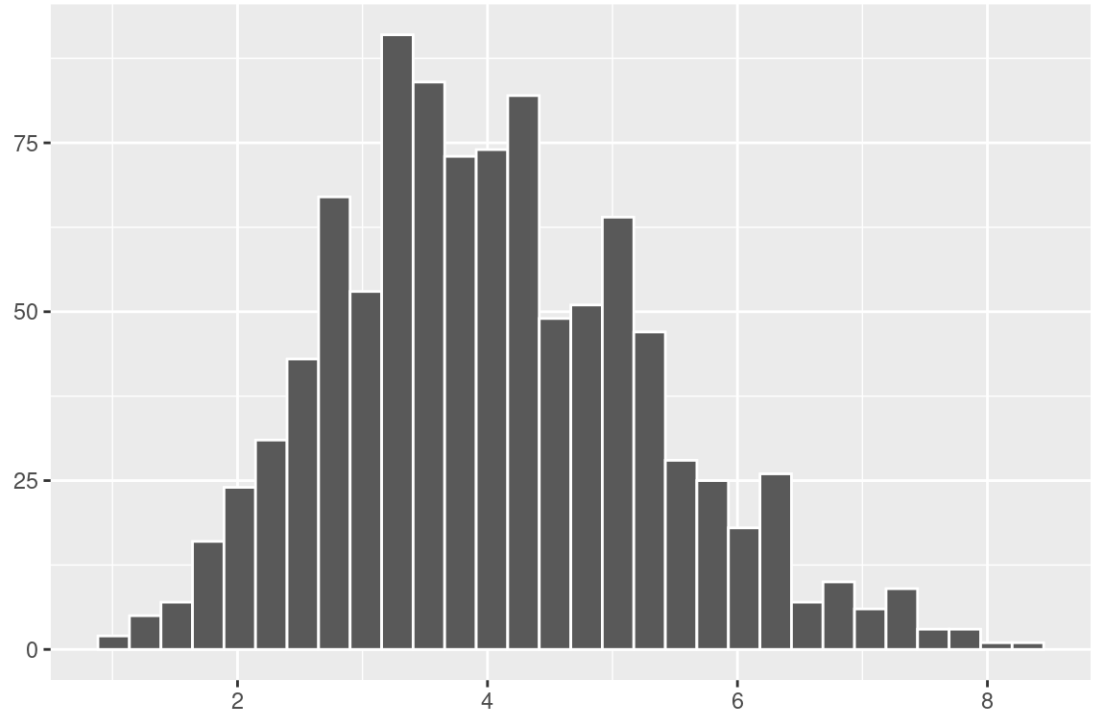
Analogamente ao estudo de simulação realizado no caso da média amostral, considere uma população normal com média $\mu = 10$ e variância $\sigma^2 = 4$.

Primeiramente, considere que são retiradas 15 amostras de tamanho 20 dessa população.



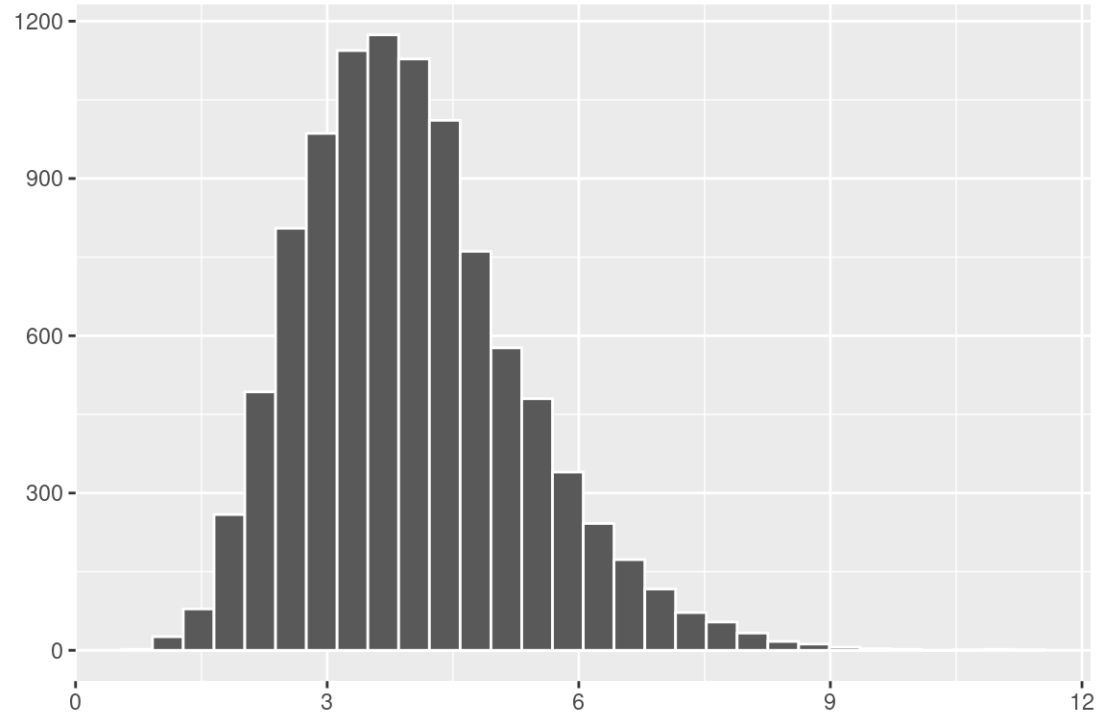
Nessa simulação obtivemos a média das variâncias igual a 3.82 e a variância das variâncias igual a 1.62.

Suponha agora que façamos o mesmo processo, porém ao invés de considerarmos 15 amostras de tamanho 20, consideramos 1000 amostras.



Neste caso a média das variâncias foi igual a 4.04 e a variância das variâncias igual a 1.63.

Realizando o mesmo experimento, porém agora considerando 10000 amostras de tamanho 20, a distribuição da variância amostral pode ser vista segundo o histograma abaixo.



Neste caso, a média das variâncias é 4.01 e a variância é 1.72. Então, realmente, podemos perceber que a distribuição da variância amostral se aproxima de uma distribuição qui-quadrado com média $\mu = 4$ e variância $\frac{2\sigma^4}{n-1} = \frac{2\times 16}{19} = 1,684$. Destas propriedades temos que $V(S^2) = \frac{2\sigma^4}{n-1}$.

Distribuição Amostral da Proporção ($\hat{p}$)

Considere que a proporção de elementos numa população com determinada característica é p. Assim, cada elemento da população pode ser representado por uma variável aleatória X, tal que

$$X = \begin{cases} 1, & \text{se o elemento apresenta a característica (com probabilidade } p); \\ 0, & \text{se o elemento não apresenta a característica (com probabilidade } 1-p). \end{cases}$$

Note que, $X \sim Bernoulli(p) = Binomial(1, p)$, e portanto, $E(X) = \mu = p$ e $Var(X) = \sigma^2 = p(1-p)$.

Seja $X_1, X_2, \ldots, X_n$ uma A.A.S. retirada com reposição dessa população, e seja

$$S_n = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$$

o total de elementos que apresentam a característica de interesse na amostra de tamanho n, então

$$S_n \sim Binomial(n, p).$$

Agora, se definimos $\hat{p}$ como a proporção de elementos que apresentam a característica na amostra; isto é;

$$\hat{p} = \frac{S_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X},$$

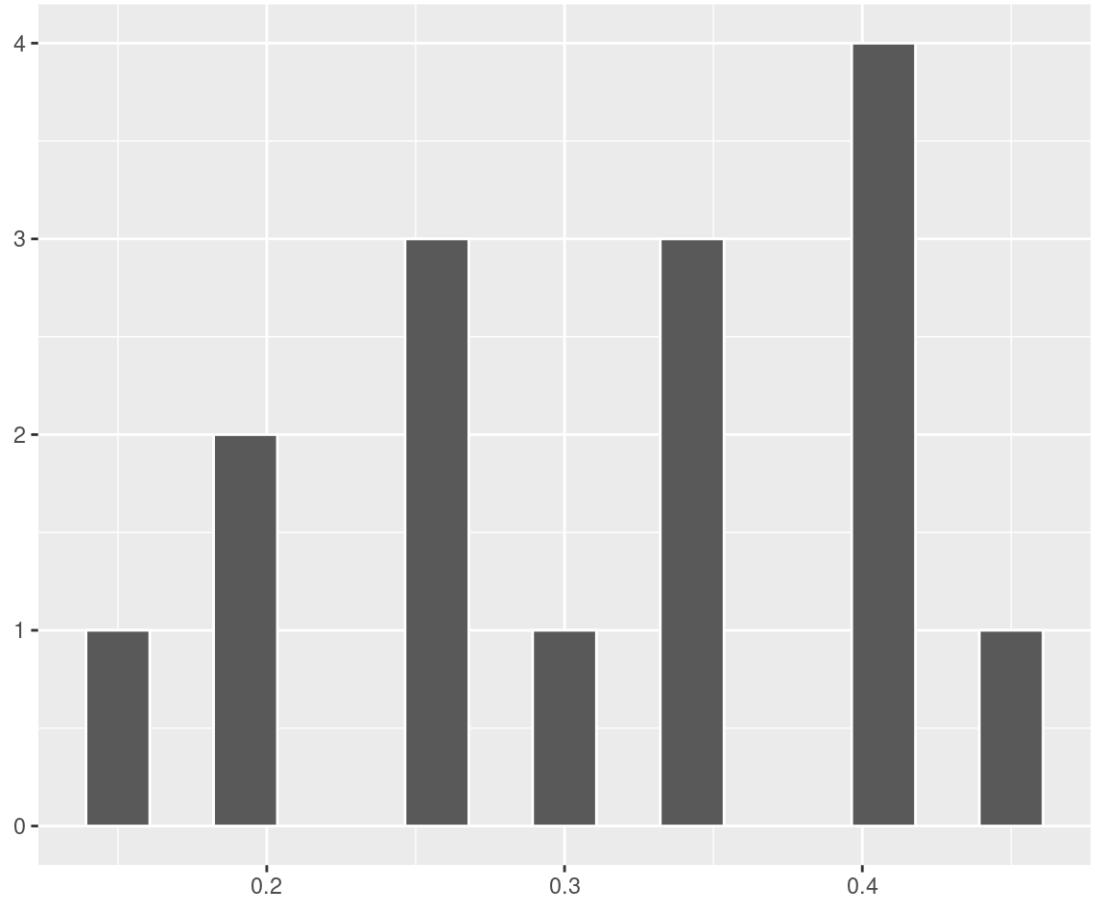
temos que, pelo **Teorema Central do Limite**,

$$\hat{p} = \overline{X} \sim \mathcal{N}(p, \frac{p(1-p)}{n}),$$

quando n é suficientemente grande. Pois, $\mu = p$ e $\sigma^2 = p(1-p)$.

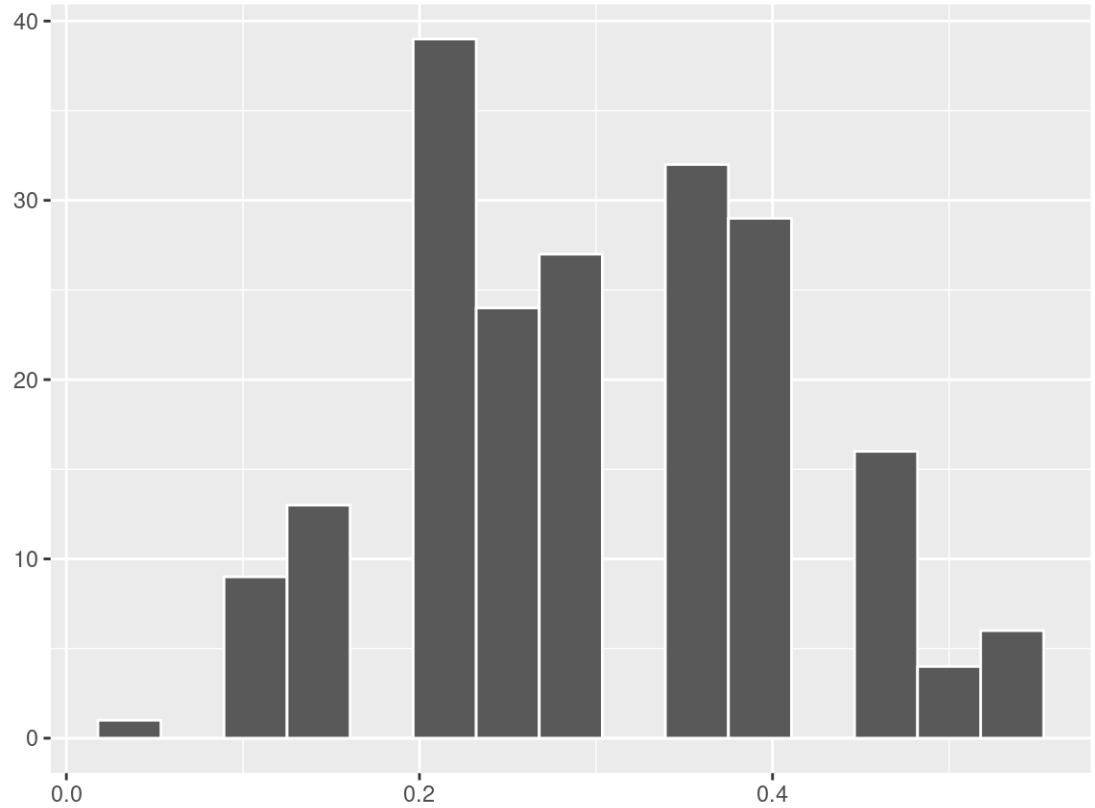
Considere uma população bernoulli com $p = 0,3$. Vamos realizar um estudo de simulação para a distribuição da proporção amostral considerando amostras de tamanho 20 dessa população.

Primeiramente, considere que são retiradas 15 amostras de tamanho 20 dessa população.



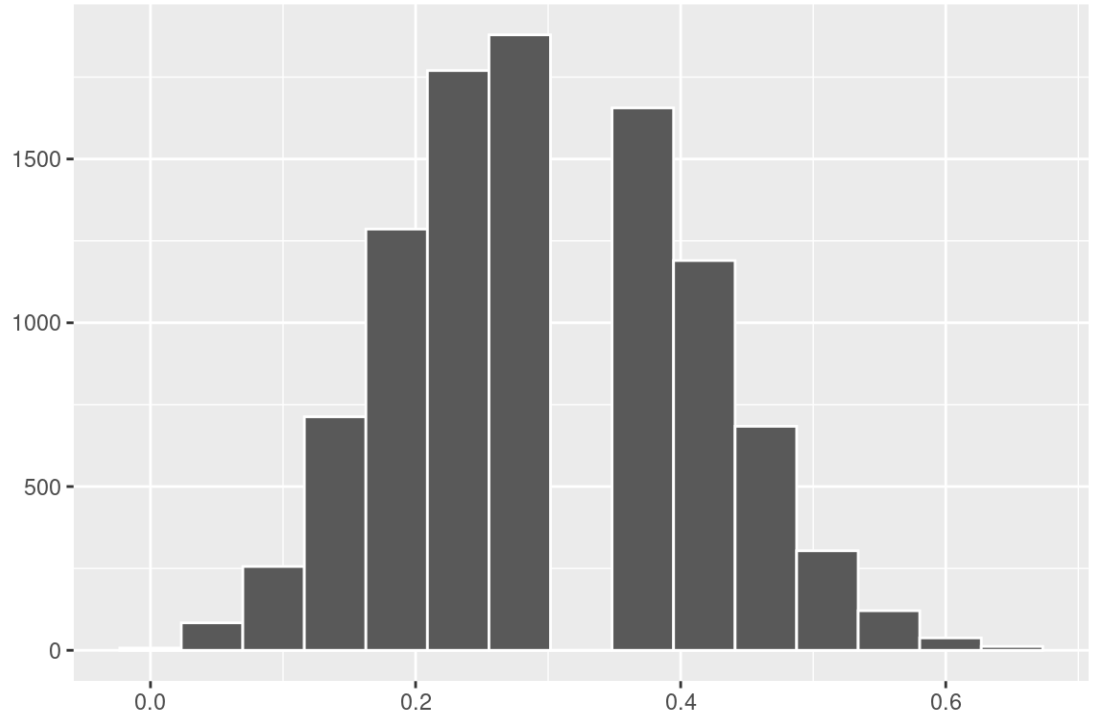
Nessa simulação obtivemos $\hat{p} = 0.31$.

Suponha agora que façamos o mesmo processo, porém ao invés de considerarmos 15 amostras de tamanho 20, consideramos 200 amostras.



Agora, obtivemos $\hat{p} = 0.3$.

Realizando o mesmo experimento, porém agora considerando 10000 amostras de tamanho 20, a distribuição da média amostral pode ser vista segundo o histograma abaixo.



Para este caso, a média das proporções amostrais foi $\hat{p} = 0.3$. Então, empiricamente, podemos perceber que a distribuição da média amostral se aproxima de uma distribuição normal com média $\mu = 0,3$ e desvio padrão $\frac{p(1-p)}{\sqrt{n}} = \frac{0.21}{\sqrt{20}} = 0,0105$.