

Projeto de MAE0315
Tecnologia de Amostragem

Fernando Henrique Ferraz Pereira da Rosa
Matheus Moreira Costa
Vagner Aparecido Pedro Junior

Universidade de São Paulo
Instituto de Matemática e Estatística

1 de julho de 2004

Resumo

Estudamos no presente trabalho, através do uso de simulações, o comportamento de diferentes planos amostrais comuns na literatura [1], comparando-os e verificando quais procedimentos se apresentam mais eficientes. Verificamos também a adequação das suposições teóricas para uso dos estimadores, assim como a adequação dos resultados simulados ao esperado na teoria. Na primeira parte comparamos o desempenho do estimador média amostral sob o plano aleatório simples com reposição, para duas populações finitas arbitrárias simuladas. Na segunda parte utilizamos um conjunto de dados disponibilizado pelo IBGE [2] para o estudo de diferentes planos amostrais e o comportamento dos estimadores apropriados dentro de cada plano.

Sumário

1	Parte 1 - AAS e o Teorema do Limite Central	2
1.1	População Normal	2
1.2	População Gama	7
2	Parte 2 - Comparação de Planos Amostrais	9
2.1	Descrição da População Utilizada	9
2.2	AASc - Amostra Aleatória Simples	10
2.2.1	Implementação	11
2.2.2	Simulação do Procedimento Amostral	13
2.3	AES - Amostra aleatória estratificada	14
2.4	RAAS - Estimador do Tipo Razão Simples	18
2.5	RAAS - Estimador do Tipo Razão Estratificado	20
2.6	RAAS - Estimador do Tipo Razão com Probabilidades Desiguais	24
2.7	RAAS - Estimador do Tipo Razão Estratificado com Probabilidades Desiguais	26
3	Considerações	29

1 Parte 1 - AAS e o Teorema do Limite Central

Nessa parte, geramos duas populações diferentes com distribuições distintas, e estudamos o efeito da variação do tamanho na amostra na distribuição da média e de sua precisão para a construção de intervalos de confiança. Os comandos utilizados foram executados no pacote estatístico R [3].

1.1 População Normal

Começamos definindo uma população com $N = 500$, através da geração de valores de uma normal com média 100 e variância 400 (desvio-padrão 20):

```
> pop1 <- rnorm(500,100,20)
> pop1[1:10]
[1] 62.90821 85.22546 135.45564 89.97311 65.04107 105.97195 73.98320
[8] 97.31953 104.27081 78.32411
```

Acima temos os primeiros 10 elementos da população. Na Figura 1 temos o histograma de densidades para essa população. Para essa população obtivemos ainda, μ e σ iguais respectivamente a:

```
> c(mean(pop1), sd(pop1))
[1] 99.72274 20.25131
```

A partir do sistema de referências para essa população $U = \{1, 2, \dots, 500\}$, vamos gerar 1000 amostras aleatórias simples com reposição de tamanho 10, 30 e 50. Para isso basta utilizar a função `sample()` para sortear 1000 conjuntos de números de 1 a 500, de tamanhos 10, 30 e 50 respectivamente.

```
amostras.tam10 <- matrix(sample(1:500,10*1000,replace=TRUE),ncol=10)
amostras.tam30 <- matrix(sample(1:500,30*1000,replace=TRUE),ncol=30)
amostras.tam50 <- matrix(sample(1:500,50*1000,replace=TRUE),ncol=50)
```

A 24ª amostra de tamanho 10, é, por exemplo:

```
> amostras.tam10[24, ]
[1] 63 468 250 482 455 425 237 57 175 270
```

Que nos retornaria os valores populacionais:

```
> pop1[amostras.tam10[24, ]]
[1] 80.00219 97.05419 91.10718 95.57220 95.10775 86.54849 95.10310
[8] 116.27813 128.56240 126.10750
```

Definimos agora a função `checa.ic`:

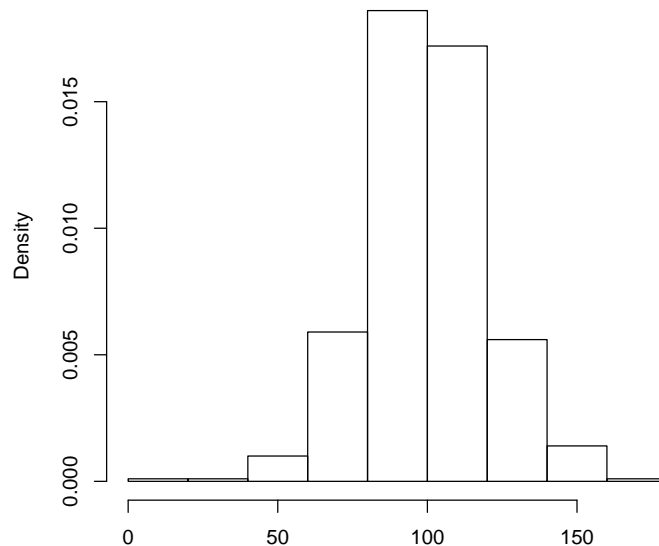


Figura 1: Histograma para a População 1

```

checa.ic <- function(amostra,pop) {
  amostra <- pop[amostra]
  med.real <- mean(pop)
  Z <- qnorm(0.975)
  conf.L <- mean(amostra) - Z*sqrt(var(amostra)/length(amostra))
  conf.U <- mean(amostra) + Z*sqrt(var(amostra)/length(amostra))
  if (med.real >= conf.L & med.real <= conf.U) {
    return(1)
  }
  return(0)
}

```

Essa função recebe como parâmetro um vetor com a amostra e outro com a população, e constrói um intervalo de confiança para a média da população, com $\gamma = 95\%$. Ela verifica então se o intervalo de confiança contém a média real da população, se sim, retorna 1, caso contrário retorna 0.

Para a amostra que tomamos como exemplo acima temos:

```

> c(mean(pop1[amostras.tam10[24, ]]), sd(pop1[amostras.tam10[24,
+   ]]))

```

```
[1] 101.14431 16.61803
```

```
> checa.ic(amostras.tam10[24, ], pop1)
```

```
[1] 1
```

Como a função retornou 1, o intervalo de confiança com $\gamma = 95\%$ contém a média real da população 1. Basta agora aplicar a função `checa.ic()` a todas as amostras de tamanho 10, 30 e 50 e verificar quantas continham a média real. Para isso basta utilizarmos os comandos abaixo:

```
> mean(apply(amostras.tam10, 1, checa.ic, pop = pop1))
```

```
[1] 0.916
```

```
> mean(apply(amostras.tam30, 1, checa.ic, pop = pop1))
```

```
[1] 0.928
```

```
> mean(apply(amostras.tam50, 1, checa.ic, pop = pop1))
```

```
[1] 0.949
```

Donde obtemos que para as amostras de tamanho 10, a cobertura foi de 91.6%, para as amostras de tamanho 30 foi de 92.8% e para as de tamanho 50 94.4%. Observa-se que quanto maior o tamanho da amostra mais próximo o nível de cobertura observado foi do nível de confiança nominal de 95%. Esse resultado era esperado pelo Teorema do Limite Central [4], que diz que a soma de n variáveis aleatórias independentes converge em distribuição para uma variável aleatória normal, conforme n cresce.

Nas Figuras 2, 3 e 4 temos os histogramas e QQPlots para a média amostral para as 1000 amostras de tamanho 10, 30 e 50 respectivamente. Observa-se que a aproximação normal é cada vez melhor, e que a variabilidade decresce com o aumento da amostra.

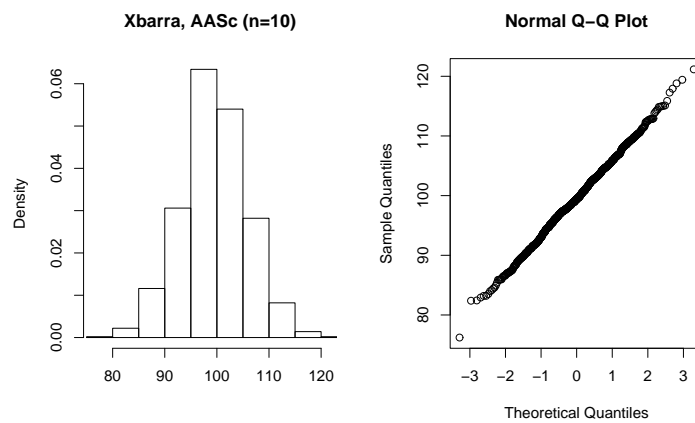


Figura 2: Distribuição de \bar{X} com amostras de tamanho 10, População Normal

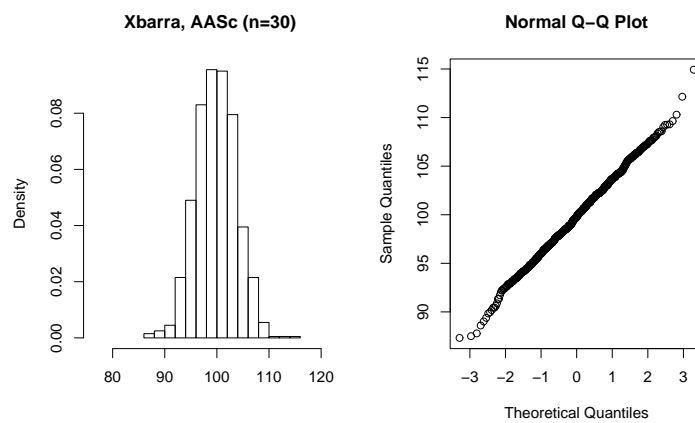


Figura 3: Distribuição de \bar{X} com amostras de tamanho 30, População Normal

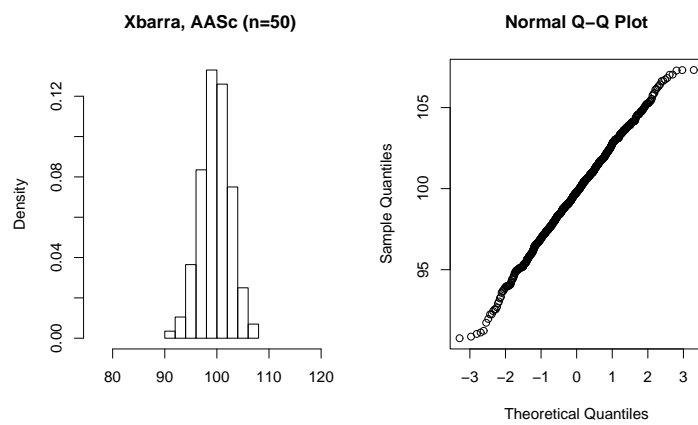


Figura 4: Distribuição de \bar{X} com amostras de tamanho 50, População Normal

1.2 População Gama

Queremos gerar uma população a partir de uma amostra de uma Gama com média 100 e variância 400 (desvio-padrão 20). Começemos notando que se $X \sim \text{Gama}(r, \lambda)$, $E(X) = r/\lambda$ e $\text{Var}(X) = r/\lambda^2$. Resolvendo o sistema:

$$\begin{cases} \frac{r}{\lambda} = 100 \\ \frac{r}{\lambda^2} = 400 \end{cases} \Rightarrow r = 25, \lambda = \frac{1}{4}$$

Geramos então $N = 500$ valores de uma variável aleatória Gama com $r = 25$ e $\lambda = 1/4$:

```
> pop2 <- rgamma(500, 25, 1/4)
```

Temos ainda para essa população $\mu = 100.9$ e $\sigma = 19.77$. Na Figura 5 temos o seu histograma de densidade.

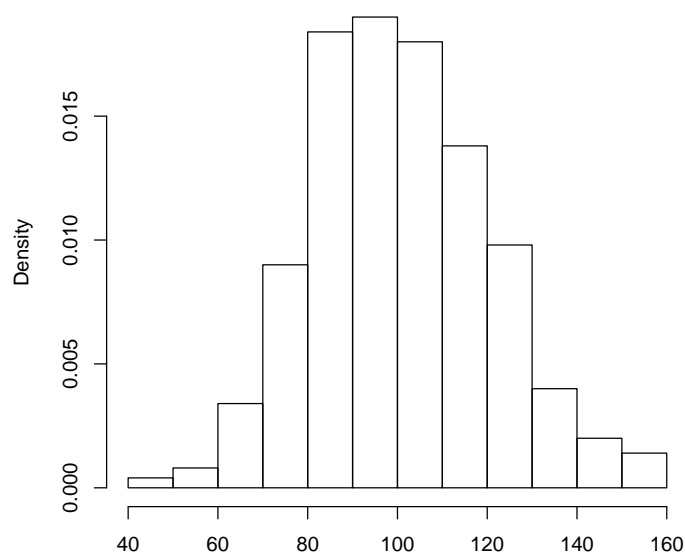


Figura 5: Histograma para a População 2

De forma análoga à realizada para a população 1, construímos 1000 intervalos de confiança com tamanho de amostra 10, 30 e 50, obtendo os seguintes níveis de cobertura:

```
> mean(apply(amostras.tam10, 1, checa.ic, pop = pop2))
```

```
[1] 0.912
```

```
> mean(apply(amostras.tam30, 1, checa.ic, pop = pop2))
```

```
[1] 0.946
```

```
> mean(apply(amostras.tam50, 1, checa.ic, pop = pop2))
```

```
[1] 0.948
```

Tivemos então que para as amostras de tamanho 10, a cobertura foi de 91.2%, para as amostras de tamanho 30 foi de 94.6% e para as de tamanho 50 94.8%. Observa-se novamente que quanto maior o tamanho da amostra mais próximo o nível de cobertura observado foi do nível de confiança nominal de 95%

Nas Figuras 6, 7 e 8 temos os histogramas e QQPlots para a média amostral para as 1000 amostras de tamanho 10, 30 e 50 respectivamente. Observa-se que a aproximação normal é cada vez melhor, e que a variabilidade decresce com o aumento da amostra. Nota-se que a aproximação não é tão boa comparada com quando usamos uma população normal, principalmente por causa dos pontos extremos, entretanto para amostras maiores essa influência dos pontos extremos também perde importância.

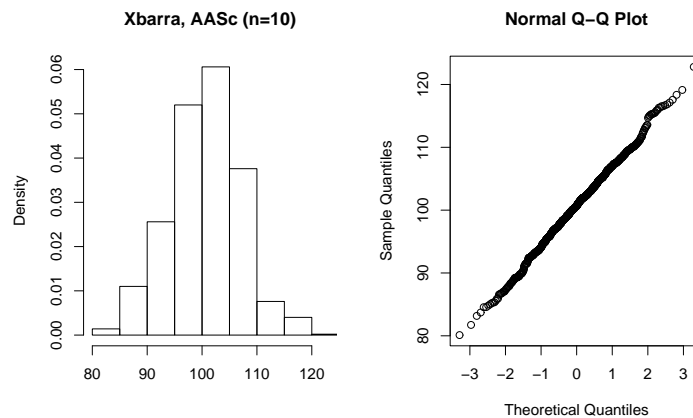


Figura 6: Distribuição de \bar{X} com amostras de tamanho 10, População Gama

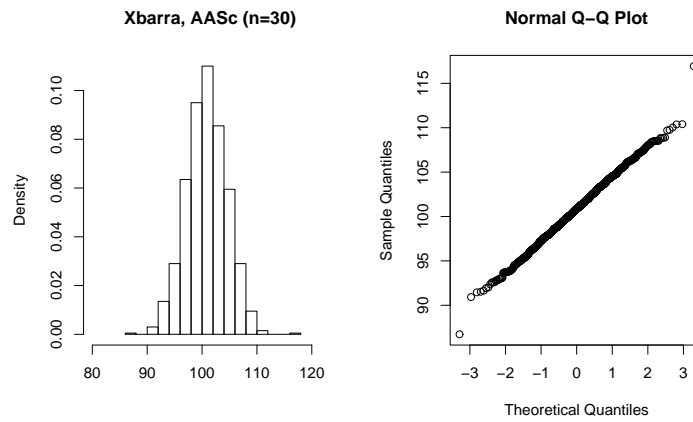


Figura 7: Distribuição de \bar{X} com amostras de tamanho 30, População Gama

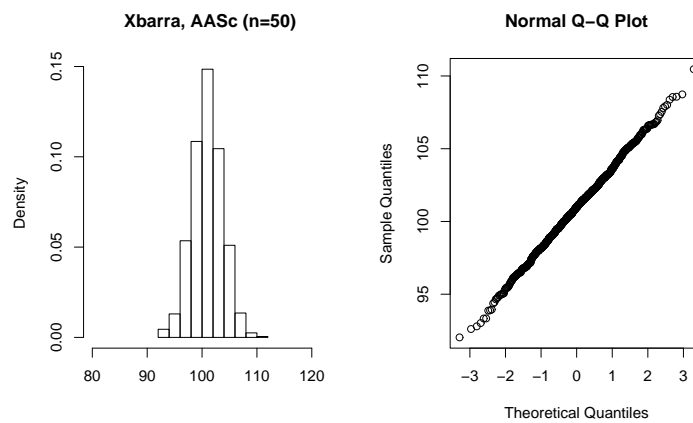


Figura 8: Distribuição de \bar{X} com amostras de tamanho 50, População Gama

2 Parte 2 - Comparação de Planos Amostrais

2.1 Descrição da População Utilizada

Obtivemos a partir do site do IBGE [2], a população descrita na Tabela 2.1. São dados por estado (números obtidos em 2002), da população e do número de telefones celulares. Os estados estão agrupados por região. Nas análises que se seguem, utilizaremos diferentes procedimentos amostrais com o intuito de estimar o total populacional do número de celulares ($\tau = 23802867$), através de uma amostra de tamanho $n = 10$. Compararemos então os diferentes planos amostrais através da distribuição do estimador expansão apropriado para cada

caso.

	populacao	celulares	regiao
Rondônia	1377792.00	70267.39	Norte
Acre	557226.00	49035.89	Norte
Amazonas	2813085.00	337570.20	Norte
Roraima	324152.00	34035.96	Norte
Pará	6189550.00	420889.40	Norte
Amapá	475843.00	56149.47	Norte
Tocantins	1155913.00	48548.35	Norte
Maranhão	5642960.00	220075.44	Nordeste
Piauí	2841202.00	116489.28	Nordeste
Ceará	7418476.00	489619.42	Nordeste
Rio Grande do Norte	2771538.00	257753.03	Nordeste
Paraíba	3439344.00	244193.42	Nordeste
Pernambuco	7911937.00	933608.57	Nordeste
Alagoas	2819172.00	256544.65	Nordeste
Sergipe	1781714.00	162135.97	Nordeste
Bahia	13066910.00	940817.52	Nordeste
Minas Gerais	17866402.00	1965304.22	Sudeste
Espírito Santo	3094390.00	371326.80	Sudeste
Rio de Janeiro	14367083.00	3965314.91	Sudeste
São Paulo	36969476.00	7098139.39	Sudeste
Paraná	9558454.00	1185248.30	Sul
Santa Catarina	5349580.00	797087.42	Sul
Rio Grande do Sul	10181749.00	2026168.05	Sul
Mato Grosso do Sul	2074877.00	255209.87	Centro-Oeste
Mato Grosso	2502260.00	265239.56	Centro-Oeste
Goiás	4996439.00	504640.34	Centro-Oeste
Distrito Federal	2043169.00	731454.50	Centro-Oeste

Tabela 1: População: Conjunto de Dados do IBGE sobre número de telefones celulares no Brasil em 2002.

2.2 AASc - Amostra Aleatória Simples

Como primeira opção consideramos a amostra aleatória simples com reposição. Obteremos amostras aleatórias com reposição de tamanho $n = 10$, e utilizaremos o estimador expansão populacional:

$$T(s) = \hat{\tau} = N\bar{y}, \quad (1)$$

que é um estimador não viesado do total populacional [1].

2.2.1 Implementação

Para simular os procedimentos amostrais estudados, vamos utilizar funções em R para cada um deles. Para tal, começamos criando uma classe “amostra”, cujos objetos vão armazenar o resultado da simulação de um dado procedimento amostral. Criamos então uma função para cada procedimento amostral. Cada função retorna um objeto da classe “amostra”, que nos permite facilmente gerar relatórios e gráficos sobre um dado procedimento. Para começar, tomemos a função mais simples, a da amostra aleatória simples com reposição:

```
> AASc

function (N, n, variavel, estatistica = c("media", "tau"))
{
  estatistica <- match.arg(estatistica)
  Est <- numeric(N)
  for (j in 1:N) {
    yb <- mean(sample(variavel, n, replace = TRUE))
    Est[j] <- switch(estatistica, media = yb, tau = length(variavel) *
      yb)
  }
  res <- list(resultados = Est, estatistica = estatistica,
    N = N, n = n, y = variavel, y.name = deparse(substitute(variavel)),
    tipo.amostra = "AASc")
  class(res) <- c("amostra", "list")
  return(res)
}
```

Essa função leva como parâmetros o número de vezes que o procedimento amostral será aplicado, o tamanho da amostra, a variável de interesse, e a estatística que desejamos estudar. Usamos “tau” para estimar o total populacional através do estimador expansão (equação 1), ou “media” para estimar a média populacional através de \bar{y} .

Temos agora duas funções, uma para fazer um gráfico a partir do objeto amostra retornado pela função AASc() e outra para imprimir um sumário na tela:

```
> plot.amostra

function (res, qqplot = FALSE, ...)
{
  if (qqplot == TRUE) {
    par(mfrow = c(1, 2))
  }
  titulo <- paste(res$tipo.amostra, "\n", res$N, "amostras de tamanho",
    sum(res$n))
  xlab <- paste(res$y.name, "\nEstatística: ", res$estatistica)
```

```

parametro <- switch(res$estatistica, tau = sum(res$y), media = mean(res$y),
  razao = sum(res$y)/sum(res$x))
hist(res$resultados, main = titulo, xlab = xlab, prob = T,
  ...)
abline(v = parametro, lty = 2)
if (qqplot == TRUE) {
  qqnorm(res$resultados)
}
}

> print.amostra

function (res)
{
  cat("  Análise do Procedimento Amostral\n\n")
  cat("  Procedimento amostral: ", res$tipo.amostra, "\n",
    sep = "")
  cat("  Variavel: ", res$y.name, "\n")
  cat("  N = ", length(res$y), ", n = ", sum(res$n), "\n",
    sep = "")
  if (!is.null(res$estratos)) {
    cat("  estratos: ", res$estratos.name, " (", length(res$n),
      " níveis)\n", sep = "")
    cat("  nh:", res$n, "\n")
  }
  if (!is.null(res$x)) {
    cat("  Variável auxiliar X: ", res$x.name, "\n")
  }
  cat("  Numero de amostras simuladas: ", res$N, "\n")
  cat("  Estatística utilizada: ", res$estatistica, "\n")
  cat("\tvalor populacional: ", switch(res$estatistica, tau = sum(res$y),
    media = mean(res$y), razao = sum(res$y)/sum(res$x)),
    "\n", sep = "")
  cat("\tmedia: ", format(mean(res$resultados), nsmall = 0),
    "\n")
  cat("\tvariância: ", var(res$resultados), "\n")
  return(invisible(res))
}

```

A primeira gera um histograma de frequência com o vetor das estatísticas, indicando no gráfico através de uma linha tracejada qual o valor do parâmetro populacional. Caso o parâmetro opcional qqplot seja especificado, é feito também um qqplot do vetor. A segunda gera uma saída com a descrição do plano amostral e algumas estatísticas do vetor de estatísticas.

2.2.2 Simulação do Procedimento Amostral

Vamos então simular 1000 amostras de tamanho $n = 10$, através da amostra aleatória simples com reposição, da população descrita na Tabela 2.1, tendo como interesse estimar o total populacional do número de celulares no Brasil. Com as funções definidas acima, basta utilizar:

```
> aas <- AASc(1000,10,celulares,estatistica="tau")
```

Temos armazenada na variável `aas` as informações sobre a simulação. Basta então pedir para ver seu conteúdo, que o método de sumário definido acima toma conta do resto:

```
> aas
```

```
      Análise do Procedimento Amostral
```

```
Procedimento amostral: AASc
Variavel:  celulares
N = 27, n = 10
Numero de amostras simuladas: 1000
Estatistica utilizada: tau
      valor populacional: 23802867
      media: 24137619
      variância: 1.545e+14
```

Observamos que o valor ficou próximo do total populacional, e que a variância do estimador ficou na ordem de 10^{14} . Podemos também ver o histograma e o qqplot na Figura 9 de densidades para o estimador expansão. Notamos que a distribuição é razoavelmente assimétrica à esquerda, revelando que a aproximação normal não seria tão boa.

```
> plot(aas, qqplot = TRUE)
```

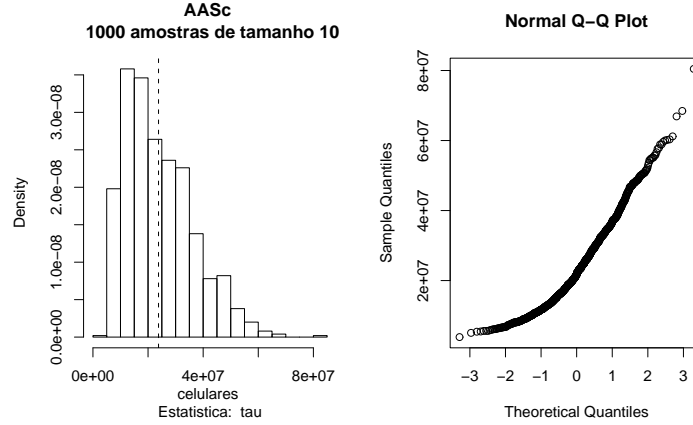


Figura 9: Distribuição do estimador expansão na AASc

2.3 AES - Amostra aleatória estratificada

A amostragem estratificada consiste na divisão de uma população em estratos segundo características conhecidas da população, de forma que atinja uma melhoria da precisão das estimativas. Dentro de cada estrato, obtem-se amostras seguindo algum plano amostral, e utilizando-se as informações sobre o tamanho de cada estrato e a amostra dentro de cada um, constrói-se um estimador estratificado. Quanto maiores as diferenças entre os estratos e quanto mais homogênea a característica de interesse dentro dos estratos, mais eficiente será o procedimento de amostragem estratificada.

Através da estrutura dos dados da Tabela 2.1, uma estratificação que se sugere é a por região do país. Como o número de celulares está relacionado com o desenvolvimento tecnológico de cada lugar, é de se esperar que determinadas regiões tenham maior número de celulares enquanto outras menos. Para verificar isso, fazemos o boxplot do número de celulares por cada região. O resultado pode ser visto na Figura 10. Nota-se claramente que o número de celulares varia bastante de região para região, e que um procedimento amostral estratificado pode melhorar bastante a precisão das estimativas.

Para estimar o total populacional através desse procedimento amostral, utilizaremos o estimador expansão estratificado:

$$T_{es} = \sum_{h=1}^H N_h \bar{y}_h, \quad (2)$$

onde aqui $H = 5$ (5 regiões), $\bar{y}_h = \frac{1}{n_h} \sum_{i \in s_h} Y_{hi}$, N_h é o tamanho do estrato h e n_h é o tamanho da amostra (aleatória simples com reposição) retirada do estrato h . Note-se que conhecemos N_h , mas n_h precisamos determinar. Começemos

```
> boxplot(celulares ~ regioao)
```

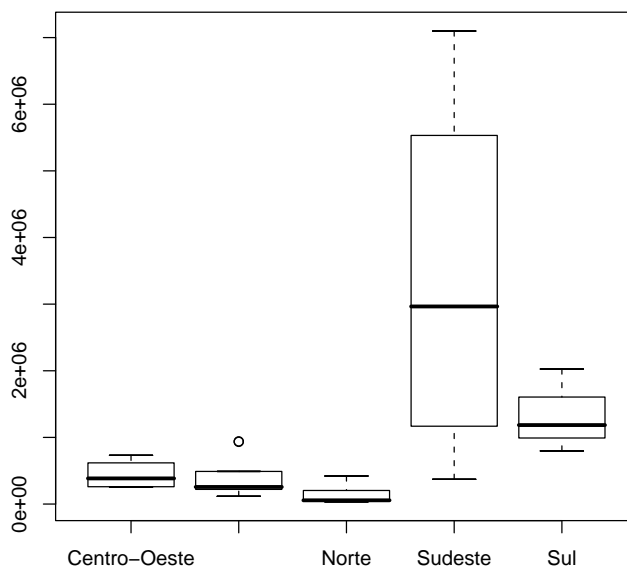


Figura 10: Número de celulares por região

utilizando a alocação proporcional. Nessa alocação tem-se que:

$$n_h = n \frac{N_h}{N}$$

Utilizando-se os dados da Tabela 2.1, temos:

```
> Nh <- tapply(as.numeric(regiao), regiao, length)
> n <- 10
> N <- 27
> round(n * Nh/N)
```

Centro-Oeste	Nordeste	Norte	Sudeste	Sul
1	3	3	1	1

Notemos que no arredondamento perdemos uma amostra. Nós sabemos que apesar de acordo com a Tabela 2.1, o Distrito-Federal ter sido considerado como um estado do Centro-Oeste, deixando a região com $N_h = 4$, na verdade o Distrito-Federal não é um estado de fato. Como na região Sudeste temos

$N_h = 4$ e os 4 são estados de fato, deixemos essa amostra adicional para a região Sudeste. definamos então n_h como:

```
> nh.proporcional <- c(1, 3, 3, 2, 1)
```

Tendo o número de amostras por estrato, basta agora utilizar a função:

```
> AES
```

```
function (N, nh, variavel, estratos, estatistica = c("media",
"tau"))
{
  Est <- numeric(N)
  Nh <- tapply(as.numeric(estratos), estratos, length)
  for (j in 1:N) {
    ybh <- numeric(max(as.numeric(estratos)))
    for (i in 1:max(as.numeric(estratos))) {
      ybh[i] <- mean(sample(variavel[as.numeric(estratos) ==
i], nh[i], replace = TRUE))
    }
    Est[j] <- switch(estatistica, media = sum(Nh * ybh)/sum(Nh),
tau = sum(Nh * ybh))
  }
  res <- list(resultados = Est, estatistica = estatistica,
N = N, n = nh, y = variavel, y.name = deparse(substitute(variavel)),
estratos = estratos, estratos.names = deparse(substitute(estratos)),
tipo.amostra = "AES")
  class(res) <- c("amostra", "list")
  return(res)
}
```

Com o comando abaixo geramos 1000 amostras de tamanho $n = 10$, com alocação proporcional das amostras entre os estratos, utilizando o estimador 2 para estimar o total populacional do número de celulares.

```
> aes.prop <- AES(1000,nh.proporcional,celulares,regiao,estatistica="tau")
```

```
> aes.prop
```

Análise do Procedimento Amostral

```
Procedimento amostral: AES
Variavel: celulares
N = 27, n = 10
estratos: regioao (5 níveis)
nh: 1 3 3 2 1
Numero de amostras simuladas: 1000
```

```

Estatística utilizada: tau
valor populacional: 23802867
media: 24027986
variância: 5.646e+13

```

Já foi possível notar uma sensível melhora na precisão do estimador. A média ficou mais próxima do parâmetro populacional, e a variância caiu para ordem de 5×10^{13} . Na Figura 11 temos o histograma e o qqplot para o vetor de estimativas. Além da redução da variabilidade, é possível notar que a distribuição do estimador não é mais assimétrica, aparentando ser razoavelmente aproximada por uma normal.

```
> plot(aes.prop, qqplot = TRUE)
```

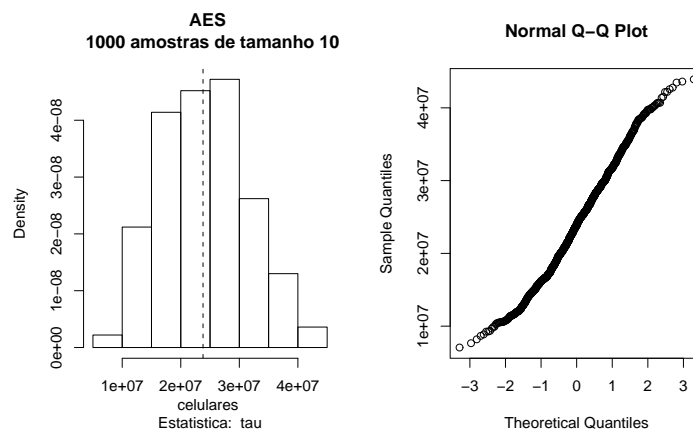


Figura 11: Distribuição do estimador expansão na AES com alocação proporcional.

Utilizemos agora a alocação ótima de Neyman, que é dada por:

$$n_h = n \frac{N_h \sigma_h}{\sum_{h=1}^H N_h \sigma_h}$$

Novamente utilizando os dados da Tabela 2.1, temos:

```

> sigmah <- tapply(celulares, regioao, sd)
> Nh <- tapply(as.numeric(regiao), regioao, length)
> n <- 10
> N <- 27
> round(n * Nh * sigmah/sum(Nh * sigmah))

```

Centro-Oeste	Nordeste	Norte	Sudeste	Sul
0	2	1	6	1

Dessa vez $\sum n_h = 10$, entretanto ficamos com uma amostra de tamanho 6 para a região Sudeste e uma amostra de tamanho 0 para a região Centro-Oeste. Tiramos uma da região Sudeste, passando para a região Centro-Oeste, já que a região Sudeste só tem 4 estados, e que temos que amostrar pelo menos um indivíduo de cada estrato para que o estimador estratificado seja não viciado. Temos então:

```
> nh.otima <- c(1, 2, 1, 5, 1)
```

Com o comando abaixo geramos 1000 amostras de tamanho $n = 10$, com alocação ótima de Neyman dos tamanhos das amostras dentro dos estratos, utilizando o estimador 2 para estimar o total populacional do número de celulares.

```
> aes.otim <- AES(1000,nh.otima,celulares,regiao,estatistica="tau")
```

```
> aes.otim
```

Análise do Procedimento Amostral

```
Procedimento amostral: AES
Variavel: celulares
N = 27, n = 10
estratos: regiao (5 níveis)
nh: 1 2 1 5 1
Numero de amostras simuladas: 1000
Estatistica utilizada: tau
      valor populacional: 23802867
      media: 23775386
      variância: 2.751e+13
```

Notamos que o estimador ficou mais preciso ainda, tendo a sua variância caído pela metade, além da sua média estar mais próxima do parâmetro populacional. Na Figura 12 temos o histograma de densidades e o gráfico QQPlot, mostrando a convergência para a normal e a redução na variância.

2.4 RAAS - Estimador do Tipo Razão Simples

Estimadores do tipo razão são apropriados quando a cada elemento i da população tem-se associado o par (X_i, Y_i) , onde a variável X é conhecida para cada elemento, sendo introduzida para se melhorar a precisão das estimativas.

No caso dos dados considerados na Tabela 2.1, a variável *populacao*, que é conhecida para todos os estados, pode ser utilizada para tornar a precisão da estimativa do número de celulares mais precisa. Para isso, consideramos o estimador razão do total populacional:

$$\hat{\tau}_Y = \frac{\bar{y}}{\bar{x}} \tau_X$$

```
> plot(aes.otim, qqplot = TRUE)
```

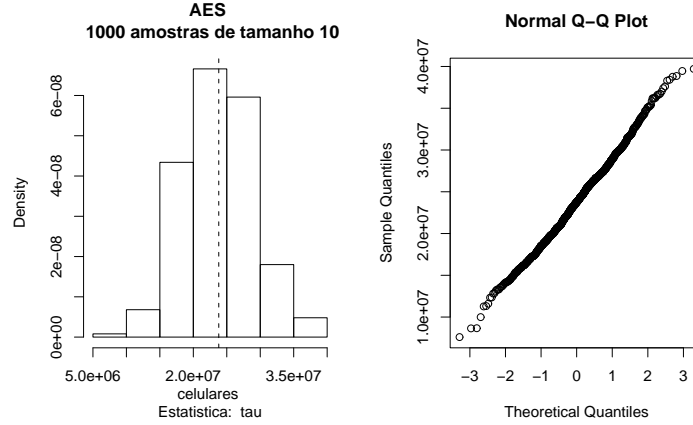


Figura 12: Distribuição do estimador expansão na AES com alocação ótima.

onde $\tau_X = 169590693$, \bar{y} e \bar{x} são obtidos através de uma amostra de tamanho $n = 10$. Notemos entretanto que esse estimador é viesado para τ_Y . De acordo com [1], para ter viés pequeno devemos ter:

$$\rho(X, Y) \frac{CV(X)}{CV(Y)} \simeq 1,$$

que no caso da população acima é dado por 0.6658. Logo o viés provavelmente não será desprezível. Por outro lado, tem-se [1] que o estimador razão do total populacional, será mais preciso que o estimador expansão na AASc quando $CV(X)/CV(Y)$ estiver entre 0.5 e 1.3 e $\rho(X, Y)$ for maior que 0.6. Novamente para os dados dessa população obtemos:

$$\frac{CV(X)}{CV(Y)} = 0.7112 \quad \rho(X, Y) = 0.9361$$

Indicando que o estimador total razão terá menor variabilidade que o estimador expansão na AASc. Para realizar a amostragem e aplicar o estimador consideramos a função:

```
> RAAS
```

```
function (N, n, x, y, estatistica = c("razao", "tau", "media"))
{
  if (length(x) != length(y)) {
    stop("x and y are not the same size.")
  }
  Est <- numeric(N)
```

```

for (i in 1:N) {
  amostra <- sample(1:length(x), n, replace = TRUE)
  r <- mean(y[amostra])/mean(x[amostra])
  Est[i] <- switch(estadistica, razao = r, tau = r * sum(x),
    media = r * mean(x))
}
res <- list(resultados = Est, estadistica = estadistica,
  N = N, n = n, y = y, y.name = deparse(substitute(y)),
  x = x, x.name = deparse(substitute(x)), tipo.amostra = "RAAS")
class(res) <- c("amostra", "list")
return(res)
}

```

Com o comando abaixo simulamos 1000 amostras de tamanho 10, estimando o número total de celulares pelo estimador razão do total populacional:

```

raas <- RAAS(1000,10,populacao,celulares,estadistica="tau")

> raas

```

Análise do Procedimento Amostral

```

Procedimento amostral: RAAS
Variavel: celulares
N = 27, n = 10
Variável auxiliar X: populacao
Numero de amostras simuladas: 1000
Estatística utilizada: tau
  valor populacional: 23802867
  media: 22279765
  variância: 2.799e+13

```

Observa-se que a variabilidade foi bem reduzida em relação à AASc, obtendo valores quase tão precisos quanto os da amostragem estratificada com alocação ótima de Neyman. Entretanto o viés foi relativamente alto, ficando a média das estimativas distante do valor real na ordem de 10^7 .

Na Figura 13 temos o histograma de densidade e o QQPlot para o procedimento amostral considerado. Ambos os gráficos indicam evidências de fuga de normalidade, sugerindo uma distribuição bimodal para o estimador. Logo, apesar da redução na variabilidade, tanto a fuga da normalidade quanto o viés foram pontos negativos desse procedimento.

2.5 RAAS - Estimador do Tipo Razão Estratificado

Uma alternativa ao estimador razão, quando a população está estratificada, é aproveitar essa estratificação para produzir um estimador razão que leve a estratificação em conta. Sejam \bar{y}_h , \bar{x}_h e τ_{X_h} as médias amostrais correspondentes

```
> plot(raas, qqplot = TRUE)
```

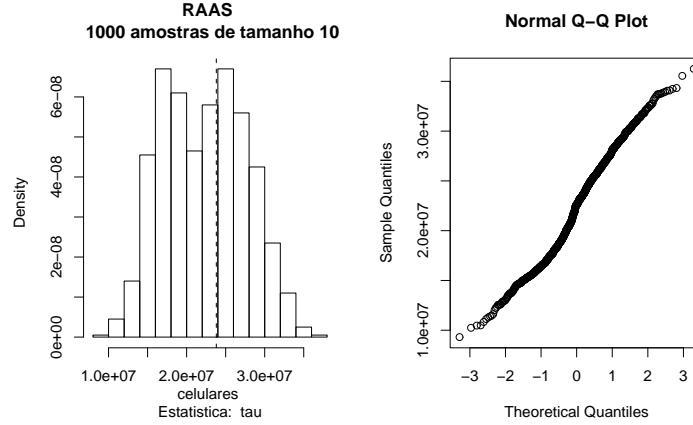


Figura 13: Distribuição do estimador razão simples do total populacional.

as variáveis Y e X e o total da variável X , respectivamente, no estrato h . Como um estimador do total populacional τ_Y , pode-se considerar:

$$T_{re} = \sum_{h=1}^H N_h \bar{y}_{Rh}$$

Onde temos que:

$$\begin{aligned} Var(T_{re}) &= Var\left(\sum_{h=1}^H N_h \bar{y}_{Rh}\right) = \sum_{h=1}^H N_h^2 Var(\bar{y}_{Rh}) \\ &\simeq \sum_{h=1}^H \frac{N_h^2}{n_h} (\sigma_{Y_h}^2 + R_h^2 \sigma_{X_h}^2 - 2R_h \rho(X, Y)_h \sigma_{Y_h} \sigma_{X_h}) \end{aligned}$$

Para AASc dentro dos estratos e tamanho de amostra suficientemente grande dentro de cada estrato [1]. Para n fixado, é possível minimizar essa variância, tomando o procedimento ótimo de Neyman:

$$n_h = n \frac{N_h \sigma_{R_h}}{\sum_{h=1}^H N_h \sigma_{R_h}}$$

Onde na AASc:

$$\sigma_{R_h}^2 = \sigma_{Y_h}^2 - 2R_h \rho(X, Y)_h \sigma_{Y_h} \sigma_{X_h} + R_h^2 \sigma_{X_h}^2$$

Calculando esses valores para a nossa população em estudo temos:

```

> Nh <- tapply(as.numeric(regiao), regiao, length)
> n <- 10
> N <- 27
> TYh <- tapply(celulares, regiao, sum)
> TXh <- tapply(populacao, regiao, sum)
> Rh <- TYh/TXh
> sigma2yh <- tapply(celulares, regiao, var)
> sigmayh <- tapply(celulares, regiao, sd)
> sigma2xh <- tapply(populacao, regiao, var)
> sigmaxh <- tapply(populacao, regiao, sd)
> rhoh <- numeric(5)
> for (i in 1:5) {
+   indices <- as.numeric(regiao) == i
+   rhoh[i] <- cor(celulares[indices], populacao[indices])
+ }
> sigma2Rh <- sigma2yh - 2 * Rh * rhoh * sigmayh * sigmaxh + Rh^2 *
+   sigma2xh
> sigmaRh <- sqrt(sigma2Rh)
> n * Nh * sigmaRh/sum(Nh * sigmaRh)

```

Centro-Oeste	Nordeste	Norte	Sudeste	Sul
1.3886	1.5866	0.4979	5.1977	1.3291

```

> round(n * Nh * sigmaRh/sum(Nh * sigmaRh))

```

Centro-Oeste	Nordeste	Norte	Sudeste	Sul
1	2	0	5	1

Como após o arredondamento $\sum n_h = 9 < 10$, e tendo a região Norte ficado com $n_h = 0$, deixamos a alocação como:

```

> nhr.otima <- c(1, 2, 1, 5, 1)

```

Definida a alocação, implementamos a função abaixo para o estimador razão estratificado:

```

> RAAS.estr

```

```

function (N, nh, x, y, estratos, estatistica = c("tau", "media"))
{
  if (length(x) != length(y)) {
    stop("x and y are not the same size.")
  }
  Est <- numeric(N)
  Nh <- tapply(as.numeric(estratos), estratos, length)
  for (i in 1:N) {
    yrh <- numeric(max(as.numeric(estratos)))

```

```

    for (j in 1:max(as.numeric(estratos))) {
      amostra <- sample(1:Nh[j], nh[j], replace = TRUE)
      fatores <- as.numeric(estratos)
      r <- mean(y[fatores == j][amostra])/mean(x[fatores ==
        j][amostra])
      yrh[j] <- r * mean(x[fatores == j])
    }
    Est[i] <- switch(estatistica, tau = sum(Nh * yrh), media = sum(Nh *
      yrh)/sum(Nh))
  }
  res <- list(resultados = Est, estatistica = estatistica,
    estratos = estratos, estratos.name = deparse(substitute(estratos)),
    N = N, n = nh, y = y, y.name = deparse(substitute(y)),
    x = x, x.name = deparse(substitute(x)), tipo.amostra = "RAAS.estr")
  class(res) <- c("amostra", "list")
  return(res)
}

```

Com o comando abaixo simulamos 1000 amostras de tamanho $n = 10$ de acordo com a alocação ótima de Neyman, utilizando o estimador total razão estratificado para estimar o número total de celulares.

```

raas.estr <- RAAS.estr(1000,nhr.otima,populacao,celulares,regiao,estatistica="tau")
> raas.estr

```

Análise do Procedimento Amostral

```

Procedimento amostral: RAAS.estr
Variavel: celulares
N = 27, n = 10
estratos: regioao (5 níveis)
nh: 1 2 1 5 1
Variável auxiliar X: populacao
Numero de amostras simuladas: 1000
Estatistica utilizada: tau
  valor populacional: 23802867
  media: 24010423
  variância: 6.82e+12

```

O que mostra que não só a estimativa ficou quase 100 vezes mais precisa (a variância caiu para a ordem de 10^{12}) em relação a AASc, como o viés parece ter desaparecido. A estimativa do total populacional foi precisa até a ordem de 10^5 . Na figura 14 o histograma de densidade e o QQ-Plot para as estimativas dentro desse delineamento amostral. A distribuição se mostra simétrica mas com caudas mais pesadas que a distribuição normal, sugerindo que a construção de intervalos de confiança baseada na suposição de normalidade não seria muito aconselhável para esse tamanho de amostra.

```
> plot(raas.estr, qqplot = TRUE)
```

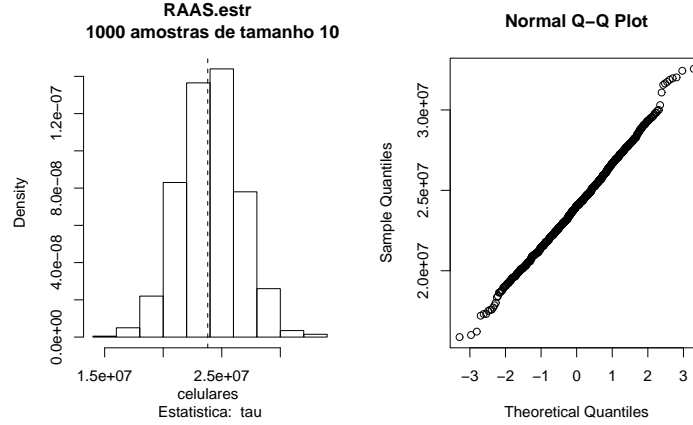


Figura 14: Distribuição do estimador razão estratificado do total populacional.

2.6 RAAS - Estimador do Tipo Razão com Probabilidades Desiguais

Consideremos agora o estimador razão com probabilidades desiguais. Consideremos uma amostra s de tamanho n selecionada de uma população na qual ao elemento i temos associado o par (X_i, Y_i) , onde a variável X é sempre conhecida. Tomemos o estimador razão dado por:

$$\bar{r} = \frac{1}{n} \sum_{i \in s} \frac{Y_i}{X_i} = \frac{1}{n} \sum_{i \in s} R_i$$

onde $R_i = X_i/Y_i, i = 1, \dots, N$. Num plano amostral com probabilidades desiguais, associamos a cada elemento i da população uma probabilidade arbitrária Z_i de seleção. No caso do estimador razão, aproveitamos o fato que os valores de X_i são conhecidos e, impomos que os valores de R_i serão selecionados com probabilidades:

$$Z_i = \frac{X_i}{\sum_{i=1}^N X_i} = \frac{X_i}{\tau_X} \quad (3)$$

onde é imediato que $\sum Z_i = 1$. De acordo com esse planejamento \bar{r} é um estimador não viesado para a razão populacional R [1]. A partir do problema 9.8 proposto em [1], obtemos o estimador para τ_Y , baseado em \bar{r} , dado por:

$$\hat{\tau}_Y = \bar{r}\tau_X. \quad (4)$$

Utilizando o Teorema 9.5 de [1], temos que sob o delineamento amostral com probabilidades finidas de acordo com a equação 3:

$$E[\hat{\tau}_Y] = E[\bar{r}\tau_X] = \tau_X E[\bar{r}] = \tau_X R = \tau_X \frac{\tau_Y}{\tau_X} = \tau_Y.$$

Através da função implementada abaixo, simulamos esse delineamento amostral N vezes:

```
> RAAS.desig

function (N, n, x, y, estatistica = c("razao", "tau", "media"))
{
  if (length(x) != length(y)) {
    stop("x and y are not the same size.")
  }
  Est <- numeric(N)
  for (i in 1:N) {
    amostra <- sample(1:length(x), n, replace = TRUE, prob = x/sum(x))
    r <- sum(y[amostra]/x[amostra])/n
    Est[i] <- switch(estatistica, razao = r, tau = r * sum(x),
      media = r * mean(x))
  }
  res <- list(resultados = Est, estatistica = estatistica,
    N = N, n = n, y = y, y.name = deparse(substitute(y)),
    x = x, x.name = deparse(substitute(x)), tipo.amostra = "RAAS.desig")
  class(res) <- c("amostra", "list")
  return(res)
}
```

Com o comando abaixo simulamos 1000 amostras de tamanho $n = 10$ da população descrita na Tabela 2.1, de acordo com o delineamento amostral acima e utilizando o estimador definido na equação (4) para estimar o total populacional do número de celulares:

```
raas.desig <- RAAS.desig(1000,10,populacao,celulares,estatistica="tau")
> raas.desig
```

Análise do Procedimento Amostral

```
Procedimento amostral: RAAS.desig
Variavel: celulares
N = 27, n = 10
Variável auxiliar X: populacao
Numero de amostras simuladas: 1000
Estatistica utilizada: tau
  valor populacional: 23802867
  media: 23763472
  variância: 1.411e+13
```

Notamos que em relação ao procedimento amostral utilizando o estimador razão simples, a variância da estimativa é maior. Entretanto o estimador não é mais viesado, obtendo um valor bem próximo do valor real do parâmetro. Na Figura 15 temos o histograma de densidades e o QQPlot para as 1000 estimativas. Não há indícios fortes de fuga da normalidade.

```
> plot(raas.desig, qqplot = TRUE)
```

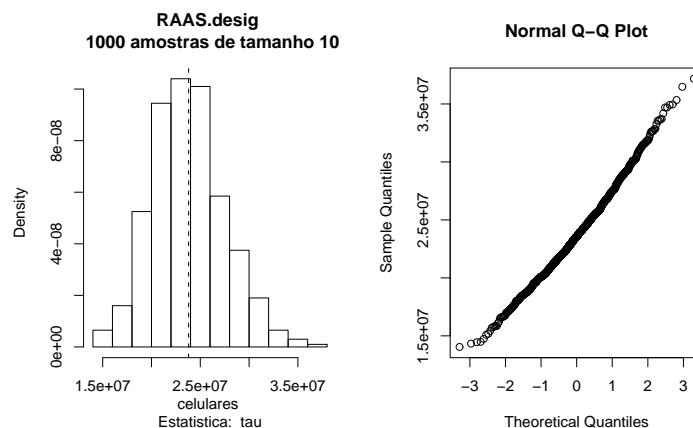


Figura 15: Distribuição do estimador razão com probabilidades desiguais do total populacional.

2.7 RAAS - Estimador do Tipo Razão Estratificado com Probabilidades Desiguais

Consideremos um delineamento amostral em que a população é dividida em estratos, e dentro desses estratos cada amostra é retirada de acordo com o estimador razão com probabilidades desiguais, de tal forma que dentro do estrato h , a probabilidade de se amostrar o elemento i é dada por:

$$Z_{hi} = \frac{X_{hi}}{\sum_{i=1}^{N_h} X_{hi}}, \quad i = 1, \dots, N_h$$

Seja então o estimador razão com probabilidades desiguais dentro do extrato h , dado por:

$$\bar{r}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{Y_{hi}}{X_{hi}} = \sum_{i=1}^{n_h} R_{hi},$$

onde os valores de X_{hi} são conhecidos para todo h e i . De forma análoga aos estimadores anteriores definimos o estimador do total populacional dentro do estrato h por:

$$\tau_{\hat{Y}_h} = \bar{r}_h \tau_{X_h}$$

Pelo teorema 9.5 de [1], dentro do estrado h , esse estimador é não viesado para o total populacional:

$$E[\tau_{\hat{Y}_h}] = E[\bar{r}_h \tau_{X_h}] = \tau_{X_h} R_h = \tau_{Y_h} \quad (5)$$

De forma que definimos o estimador razão estratificado com probabilidades desiguais do total populacional por:

$$\tau_{\hat{Y}_{ppz}} = \sum_{j=1}^H \tau_{\hat{Y}_j} \quad (6)$$

Que é não viesado para τ_Y de acordo com a equação 5. Na função abaixo temos esse delineamento amostral implementado:

```
> RAAS.estr.desig

function (N, nh, x, y, estratos, estatistica = c("tau", "media"))
{
  if (length(x) != length(y)) {
    stop("x and y are not the same size.")
  }
  Est <- numeric(N)
  Nh <- tapply(as.numeric(estratos), estratos, length)
  for (i in 1:N) {
    yrh <- numeric(max(as.numeric(estratos)))
    for (j in 1:max(as.numeric(estratos))) {
      fatores <- as.numeric(estratos)
      amostra <- sample(1:Nh[j], nh[j], replace = TRUE,
        prob = x[fatores == j]/sum(x[fatores == j]))
      r <- mean(y[fatores == j][amostra])/mean(x[fatores ==
        j][amostra])
      yrh[j] <- r * mean(x[fatores == j])
    }
    Est[i] <- switch(estatistica, tau = sum(Nh * yrh), media = sum(Nh *
      yrh)/sum(Nh))
  }
  res <- list(resultados = Est, estatistica = estatistica,
    estratos = estratos, estratos.name = deparse(substitute(estratos)),
    N = N, n = nh, y = y, y.name = deparse(substitute(y)),
    x = x, x.name = deparse(substitute(x)), tipo.amostra = "RAAS.estr.desig")
  class(res) <- c("amostra", "list")
  return(res)
}
```

Com o comando abaixo obtemos 1000 amostras de tamanho $n = 10$, com tamanho de amostras n_h iguais ao utilizados no caso do estimador razão estratificado, usando o estimador proposto na equação 6 para estimar o número total de celulares:

```
raas.estr.desig <-  
  RAAS.estr.desig(1000,nhr.otima,populacao,celulares,regiao,estatistica="tau")  
> raas.estr.desig
```

Análise do Procedimento Amostral

```
Procedimento amostral: RAAS.estr.desig  
Variavel: celulares  
N = 27, n = 10  
estratos: regioao (5 níveis)  
nh: 1 2 1 5 1  
Variável auxiliar X: populacao  
Numero de amostras simuladas: 1000  
Estatística utilizada: tau  
  valor populacional: 23802867  
  media: 23881678  
  variância: 4.648e+12
```

Na Figura 16 temos o histograma de densidades e o QQ-Plot para as 1000 estimativas de $\tau\gamma$. Observamos que tanto o estimador foi mais preciso, com variância na ordem de 10^{12} , quanto não viesado, obtendo média bem próxima do total populacional.

```
> plot(raas.estr.desig, qqplot = TRUE)
```

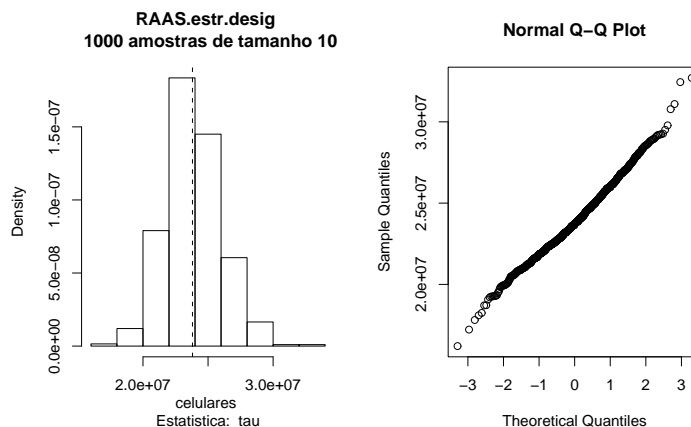


Figura 16: Distribuição do estimador razão estratificado com probabilidades desiguais do total populacional.

3 Considerações

O objetivo de trabalho mais frequente do estatístico, é o controle e a redução da variabilidade - seja com a intenção de prever o comportamento de um sistema, estimar uma quantidade populacional desconhecida, identificar padrões de comportamento entre muitas outras possibilidades.

Pudemos notar através das simulações executadas que o delineamento de um plano amostral adequado, influencia decisivamente a qualidade e precisão das estimativas que forem feitas sobre a amostra coletada. O acréscimo de informações adicionais conhecidas, quando feito de forma adequada, é a melhor ferramenta nesse sentido. Quando temos a população dividida em grupos distintos (como no caso dos estados divididos em regiões), e essa divisão é relevante na variável ou parâmetro que temos interesse em estimar, um delineamento amostral que leve isso em conta (estratificado, conglomerados) pode melhorar bastante a qualidade das estimativas. A presença de uma característica populacional conhecida, como no caso do conjunto de dados em estudo, o número de habitantes por estado, pode ser levada em conta através de estimadores razão, e também melhora a precisão das estimativas, quando a variável auxiliar tem correlação com a variável de interesse.

Na Tabela 2 e na Figura 17 temos um resumo comparativo dos principais delineamentos utilizados. Analisando os dois notamos que o plano aleatório simples teve a maior variância das estimativas, além de uma distribuição claramente assimétrica a direita. Em média o total populacional foi superestimado na ordem de 10^5 . Essa superestimativa é esperada pois como podemos notar na Figura 10 os valores na região Sudeste são muito maiores que no resto do país,

de forma que amostras que contenham vários estados dessa região vão inflar a estimativa do total populacional.

Em seguida temos o plano estratificado, com alocação ótima das unidades amostrais. Além da variância cair da ordem 10^{14} para 10^{13} , o erro médio com que estimamos o total populacional caiu para ordem de 10^4 . A distribuição das estimativas deixou de ser assimétrica, apesar de ter caudas mais leves que uma distribuição normal.

No caso do estimador razão simples, a variância continuou na ordem de 10^{13} , entretanto o valor populacional menos a média dos estimadores teve o maior desvio de todos os planos estudados, ficando na ordem de 10^6 . Isso reflete o fato do estimador razão ser viesado para a razão populacional, e n ser pequeno. Temos também uma distribuição das estimativas longe da normal, aparentando uma distribuição bimodal.

No caso do estimador razão estratificado, a variância melhorou, caindo para ordem de 10^{12} . A distribuição das estimativas não mostrou grande evidência de fuga da normalidade. Entretanto a média das estimativas ficou razoavelmente distante do total populacional, ficando o erro na ordem de 10^5 , superestimando o total populacional assim como na amostra aleatória simples.

O estimador razão com probabilidades desiguais mostrou um desvio menor em relação ao parâmetro populacional, ficando na ordem de 10^4 . A distribuição se mostrou razoavelmente normal, entretanto a variância cresceu para ordem de 10^{13} .

Por fim o estimador razão estratificado com probabilidades desiguais obteve tanto um erro médio razoável, na ordem de 10^4 quanto a variância baixa, na ordem de 10^{12} . Podemos notar que esse delineamento uniu a variância baixa do plano amostral razão estratificado com o erro menor do estimador razão com probabilidades desiguais. A estratificação reduziu a variância e o uso do delineamento com probabilidades desiguais garantiu que o estimador fosse não viesado, obtendo erro médio menor.

Concluimos que o delineamento amostral que apresentou o melhor resultado foi portanto o razão estratificado com probabilidades desiguais. Esse resultado era esperado no sentido que esse delineamento foi o que levou em conta todas as informações que possuímos em relação a população: a estratificação, a correlação da população com o número de celulares por estado e o peso maior dado a estados com maior razão do número de celulares por população.

O único contratempo foi que a distribuição das estimativas para o estimador do total populacional nesse delineamento apresenta caudas um pouco mais leves que a da distribuição normal, de forma que a utilização da aproximação normal para análises inferenciais deve ser feita com cautela.

	media	variancia	$\tau_Y - media$
AAS	24137619.21	1.54E+14	-3.35E+05
AES	23775386.20	2.75E+13	2.75E+04
RAAS	22279764.99	2.80E+13	1.52E+06
RAAS.estr	24010422.67	6.82E+12	-2.08E+05
RAAS.desig	23763471.88	1.41E+13	3.94E+04
RAAS.estr.desig	23881677.76	4.65E+12	-7.88E+04

Tabela 2: Comparação dos diversos delineamentos amostrais simulados

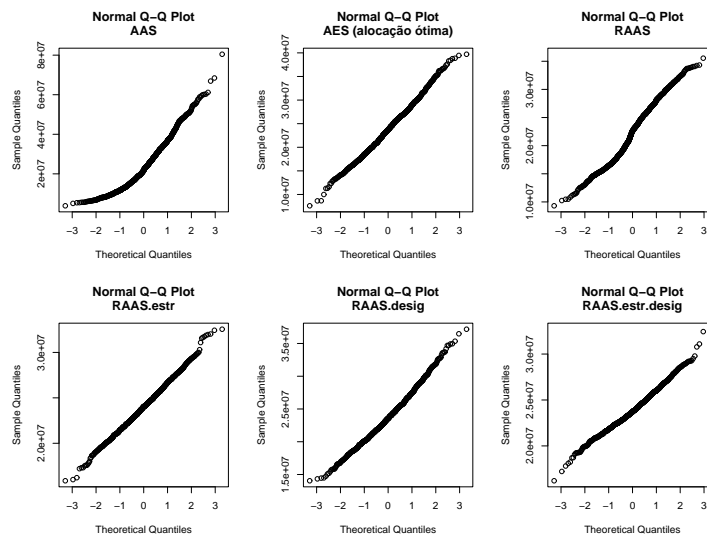


Figura 17: Q-Q Plots dos diversos delineamentos amostrais simulados

Referências

- [1] BOLFARINE, H.; BUSSAB, W. O. *Elementos de Amostragem*. São Paulo: Versão Preliminar, 2000.
- [2] ESTATÍSTICAS do Século XX. Disponível em: <<http://www.ibge.gov.br/>>.
- [3] R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, 2004. ISBN 3-900051-00-3. Disponível em: <<http://www.R-project.org/>>.
- [4] JAMES, B. R. *Probabilidade: Um curso em nível intermediário*. Rio de Janeiro: IMPA, 2002.

Sobre

A versão eletrônica desse arquivo pode ser obtida em <http://www.feferraz.net>

Copyright (c) 1999-2005 Fernando Henrique Ferraz Pereira da Rosa.
É dada permissão para copiar, distribuir e/ou modificar este documento sob os termos da Licença de Documentação Livre GNU (GFDL), versão 1.2, publicada pela Free Software Foundation;
Uma cópia da licença em está inclusa na seção intitulada "Sobre / Licença de Uso".