Inferência Estatística (revisão)

CE219 - Controle Estatístico de Qualidade

Prof. Cesar Taconeli
taconeli@ufpr.br

Prof. Walmes Zeviani
walmes@ufpr.br

Laboratório de Estatística e Geoinformação
Departamento de Estatística
Universidade Federal do Paraná

Métodos estatísticos para a análise da qualidade

- O objetivo aqui é fazer uma (breve) revisão de **modelos probabilísticos** e **métodos estatísticos** com aplicação na descrição, modelagem e produção de inferências para processos.
- Serão abordados:
  1. Métodos de análise descritiva.
  2. Probabilidade e principais modelos probabilísticos.
  3. **Inferência estatística aplicada à qualidade do processo**.
- As próximas aulas serão intercaladas com ilustrações no R e os scripts disponibilizados na página da disciplina.

Inferências sobre a qualidade do processo

Introdução

- Modelos probabilísticos são aplicáveis na modelagem de variáveis que **caracterizam** a qualidade de processos.
- Na prática, os parâmetros que determinam tais modelos são **desconhecidos**.
- Utilizamos **amostras** selecionadas do processo como base para estimação dos parâmetros e teste de hipóteses.
- No Controle Estatístico Qualidade, a **inferência** sobre parâmetros do processo é fundamental para efeito de monitoramento, avaliação do desempenho e identificação de causas atribuíveis de variação em processos.

O processo de inferência

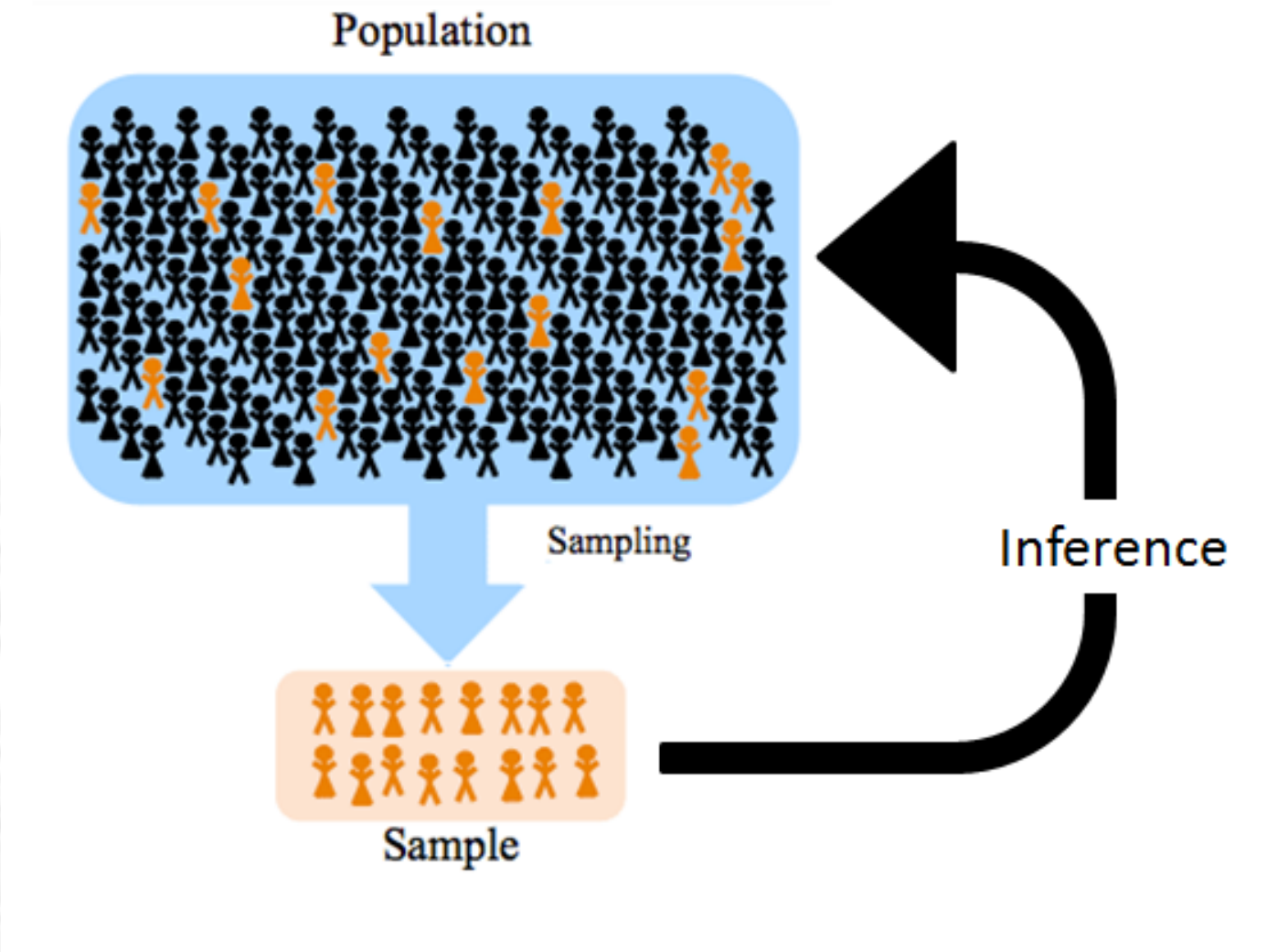


Figura 1. População e amostra.

Estatísticas e distribuições amostrais

- **Estatísticas** são funções dos dados amostrais que independem de parâmetros desconhecidos.
- Assim como as características da qualidade configuram variáveis aleatórias, às quais assumimos modelos probabilísticos apropriados, as estatísticas também configuram variáveis aleatórias, tendo suas respectivas distribuições de probabilidades.
- A distribuição de probabilidades de uma estatística é chamada **distribuição amostral**. Usaremos as distribuições amostrais de algumas estatísticas como base para a inferência de parâmetros do processo.

Amostra aleatória

- Considere $X_1, X_2, \cdots, X_n$ uma amostra aleatória de tamanho n da população sob estudo (ex: observações independentes de alguma característica da qualidade de um processo industrial).
- Vamos usar o termo **amostra aleatória** para nos referir a um conjunto de observações (variáveis aleatórias) independentes e identicamente distribuídas.
- Algumas das principais estatísticas usadas no monitoramento de processos, acompanhadas de suas distribuições amostrais, são discutidas na sequência.

Distribuição amostral da média amostral

- A média amostral, estimador não viciado da variância populacional, é definida por:

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

- Se assumirmos que $X_1, X_2, \cdots, X_n$ é uma amostra aleatória de uma distribuição Normal com parâmetros μ e σ^2, então a distribuição amostral da média amostral fica dada por:

$$\bar{X} \sim \text{Normal}\left(\mu_{\bar{X}} = \mu, \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}\right).$$

O TLC na distribuição amostral da média

- Ainda que a amostra não tenha sido produzida por uma distribuição (população) Normal, o Teorema Central do Limite (TCL) garante que, assintoticamente (quando $n \to \infty$), $\bar{X} \sim N(\mu, \sigma^2/n)$, uma vez que podemos escrever:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

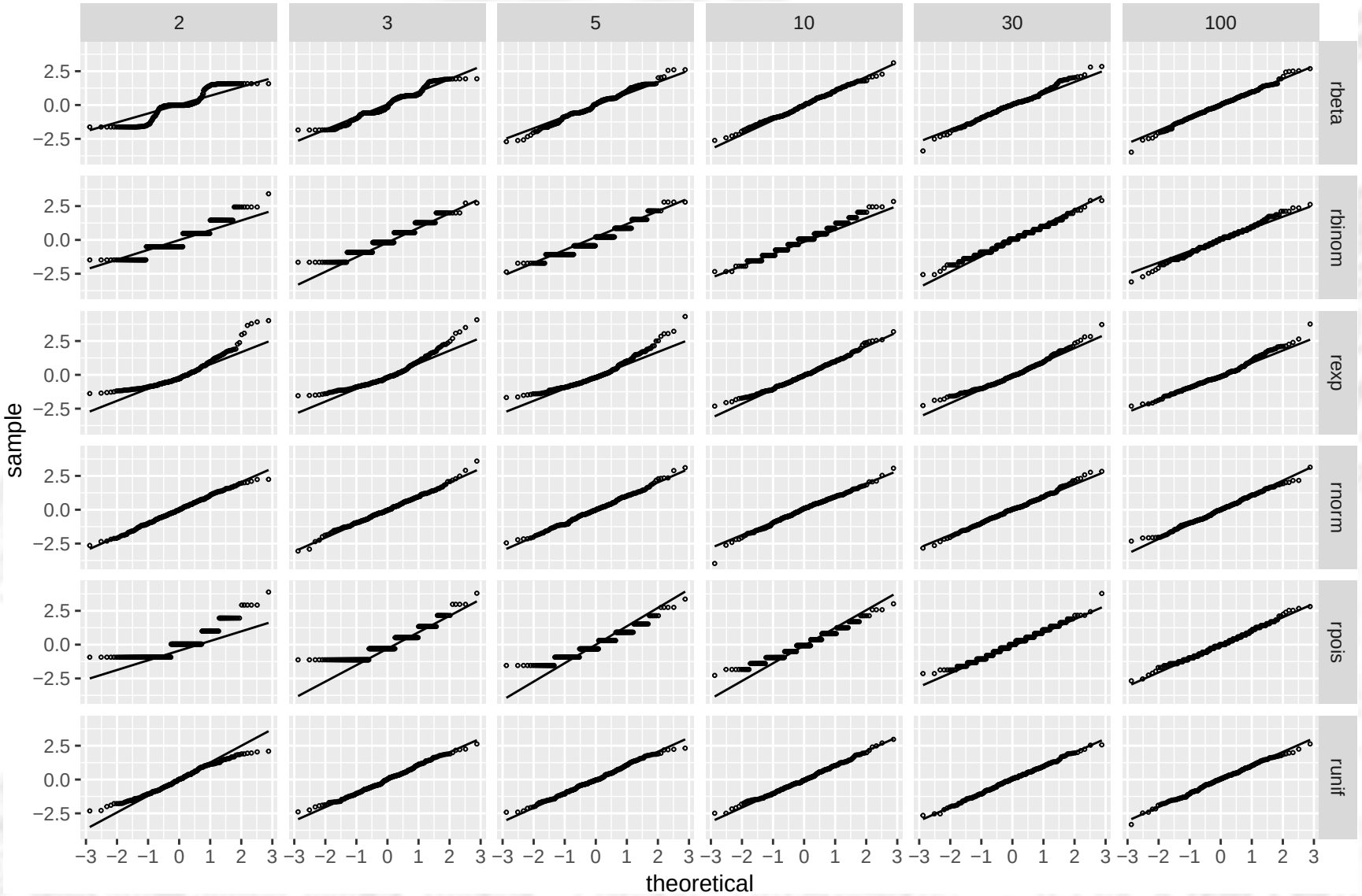
- Por mais que o TCL configure um resultado assintótico, a distribuição normal para a média amostral é verificada com boa aproximação para grandes amostras (n suficientemente grande).

Tamanho da amostra para o TLC

- O tamanho amostral necessário para se alcançar uma boa aproximação vai depender da distribuição da população sob estudo. É comum encontrar na literatura, como regra geral, que isso ocorre quando $n > 30$.
- No entanto, em boa parte dos casos temos uma aproximação satisfatória para tamanhos de amostra bem menores, até mesmo para $n = 10$ ou $n = 5$.

Estudo de simulação

```r
library(tidyverse)

simul <- function(dist, params, n = 10, replications = 250) {
    replicate(replications, {
        mean(invoke(dist, c(list(n = n), as.list(params))))
    })
}

tb_dist <- tibble(params = list(
                      "rnorm" = c(mean = 0, sd = 1),
                      "runif" = c(min = 0, max = 1),
                      "rbeta" = c(shape1 = 0.1, shape2 = 0.1),
                      "rexp" = c(rate = 1),
                      "rpois" = c(lambda = 0.5),
                      "rbinom" = c(size = 3, prob = 0.25)),
                  dist = names(params))
tb_n <- crossing(n = c(2, 3, 5, 10, 30, 100), dist = tb_dist[["dist"]])
tb <- inner_join(tb_dist, tb_n)

tb <- tb %>%
    mutate(x = pmap(list(dist, params, n), simul)) %>%
    mutate(y = map(x, scale)) %>%
    unnest(x, y)

ggplot(tb, aes(sample = y)) + facet_grid(facets = dist ~ n) +
    geom_qq(pch = 1, alpha = 1, size = 0.5) + geom_qq_line()
```


2
3
5
10
30
100
rbeta
rbinom
rexp
rnorm
rpois
runif
sample
theoretical
2.5
0.0
−2.5
−3 −2 −1 0 1 2 3

Distribuição amostral da variância amostral

- A variância amostral, como sabemos, é definida por:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- Se $X_1, X_2, \cdots, X_n$ é uma amostra aleatória de uma distribuição Normal com parâmetros μ e σ^2, então:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Distribuição amostral da variância amostral

- χ^2_{n-1} representa a distribuição qui-quadrado com $n-1$ graus de liberdade. Uma variável aleatória X com distribuição χ^2_k tem função densidade de probabilidade:

$$f(y) = \frac{1}{2^{k/2}\Gamma(\frac{k}{2})} y^{(k/2)-1} e^{-y/2}, \quad y > 0.$$

- A média da distribuição amostral de s^2 é $\mu_{s^2} = \sigma^2$. Baseado nesse fato, dizemos, como será discutido na sequência, que s^2 é um estimador não viciado de σ^2.

Distribuição amostral da proporção amostral

- Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória de uma distribuição Bernoulli, de parâmetro p.
- A distribuição Bernoulli permite modelar um experimento do tipo *sucesso vs fracasso*, por meio de uma variável aleatória que assume valor 0 para um dos desfechos e 1 para o outro:

$$\Pr(X = x) = \begin{cases} p, & x = 1 \\ (1-p), & x = 0 \end{cases}$$

Distribuição amostral da proporção amostral

- Nesse caso, a proporção amostral nada mais é que a média da amostra:

$$\hat{p} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

- A distribuição amostral exata para a proporção amostral pode ser obtida a partir da distribuição binomial. No entanto, usando o TCL temos que, assintoticamente,

$$\hat{p} \sim N\left(\hat{\mu}_{\hat{p}} = p, \quad \hat{\sigma}^2_{\hat{p}} = \frac{p(1-p)}{n}\right).$$

Estimação pontual de parâmetros do processo

- Estatísticas utilizadas para estimar parâmetros populacionais desconhecidos são denominadas **estimadores**.
- Um **estimador pontual** é uma estatística que produz um único número como estimativa para o parâmetro que desconhecemos.
- Ao valor do estimador avaliado numa particular amostra damos o nome de **estimativa**.
- Assim, para monitorar a média de um processo podemos considerar como estimador a média amostral ($\bar{X}$).
- Ao coletar uma amostra, calculamos, com base nos dados amostrais, $\bar{X} = 10$. Esse valor é a estimativa da média do processo naquele momento da produção.

Estimação pontual de parâmetros do processo

- Dentre as propriedades desejadas de um estimador pontual, destacamos:
  - Ausência de viés: dizemos que um estimador é não viesado se a média de sua distribuição amostral (seu valor esperado) for igual ao parâmetro que estamos estimando. Assim, $\hat{\theta}$ é um estimador não viesado de θ se:

$$\mu_{\hat{\theta}} = \mathrm{E}(\hat{\theta}) = \theta.$$

  - Apresentar variância mínima: além de não apresentar viés, deseja-se que um estimador produza estimativas que apresentem baixa variabilidade. Um estimador $\hat{\theta}$ é de mínima variância se:

$$\sigma^2_{\hat{\theta}} = \mathrm{Var}(\hat{\theta}) < \mathrm{Var}(\hat{\theta}^*),$$

para qualquer outro estimador $\hat{\theta}^*$.

Estimação pontual de parâmetros do processo

- Como vimos anteriormente, a média e a variância amostrais são estimadores não viciados dos correspondentes parâmetros populacionais:

$$\mathrm{E}(\bar{X}) = \mu; \quad \mathrm{E}(s^2) = \sigma^2.$$

- No entanto, o desvio padrão amostral:

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

não é um estimador não viciado do desvio padrão populacional σ.

Estimação pontual de parâmetros do processo

- Pode-se mostrar que, se a distribuição do processo for normal:

$$\mathrm{E}(s) = \left(\frac{2}{n-1}\right)^{1/2} \frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}\sigma = c_4\sigma,$$

  sendo c_4 uma constante que depende do tamanho da amostra.
- Assim, um estimador não viciado de σ é dado por:

$$\hat{\sigma}_s = \frac{s}{c_4}.$$

- **Nota:** c_4 se aproxima de 1 a medida que n aumenta, refletindo que s é assintoticamente não viciado.

Estimação pontual de parâmetros do processo

- É comum, em problemas de CEP, usar a amplitude amostral para estimar o desvio padrão da população. A amplitude amostral é definida por:

$$R = \max(X_i) - \min(X_i) = X_{(n)} - X_{(1)}.$$

- Denominamos $W = R/\sigma$ como *amplitude relativa*. A distribuição amostral de W é conhecida, sabendo-se que, sob distribuição normal:

$$\mathrm{E}(W) = d_2,$$

sendo d_2 uma constante que depende apenas do tamanho amostral (assim como c_4).

Estimação pontual de parâmetros do processo

- Valores de d_2 e c_4 podem ser calculados facilmente ou extraídos de tabelas nas referências de CEQ.
- Assim, um estimador não viesado para σ baseado na distribuição de W é dado por:

$$\hat{\sigma}_R = \frac{R}{d_2}.$$

- A eficiência da amplitude amostral na estimação do desvio padrão do processo cai rapidamente conforme se aumenta o tamanho da amostra.
- Na prática é recomendável usar o desvio padrão ao invés da amplitude amostral.

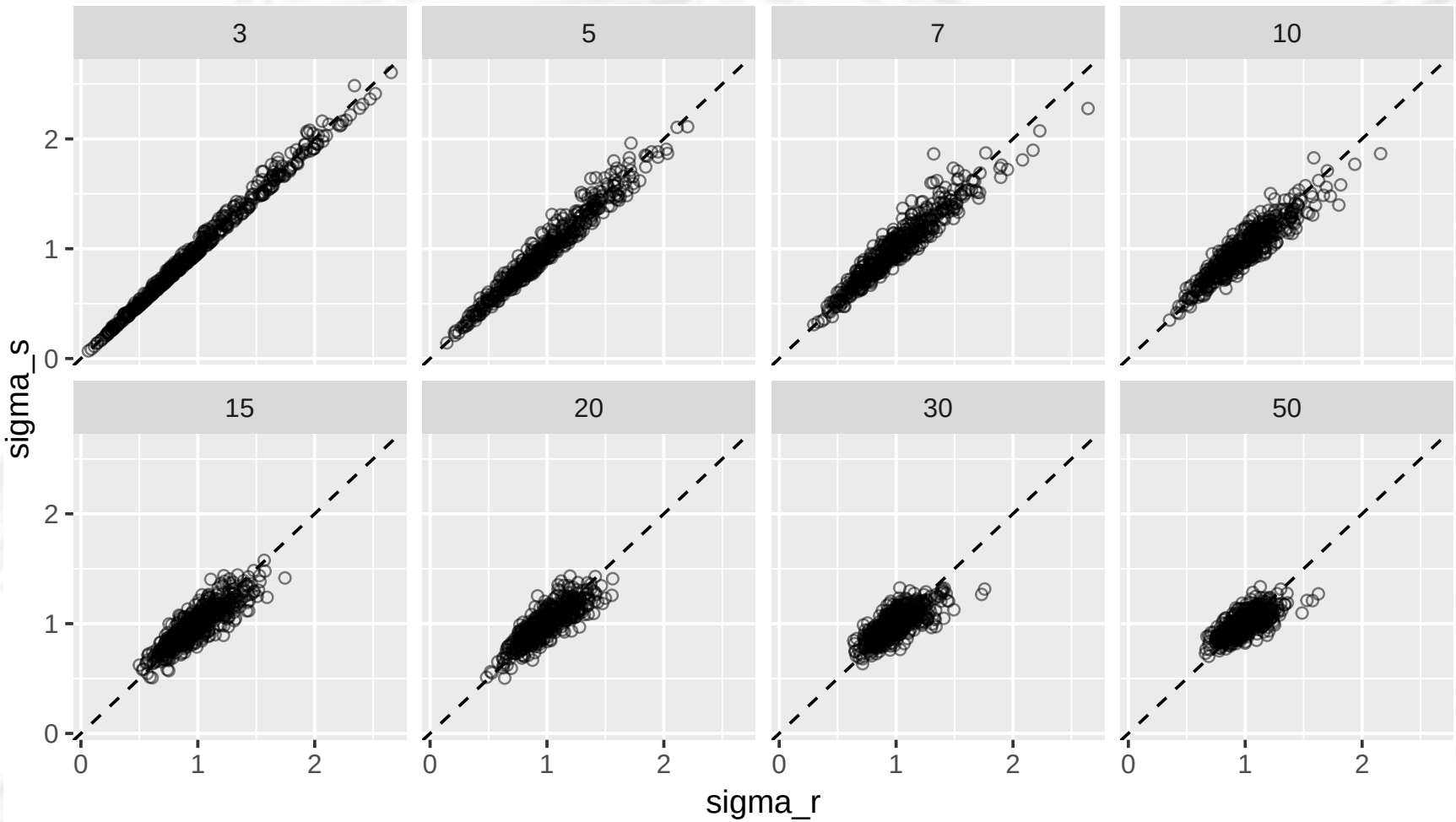
Estimação pontual de parâmetros do processo

Tabela 1. Eficiência relativa dos estimadores do desvio padrão.

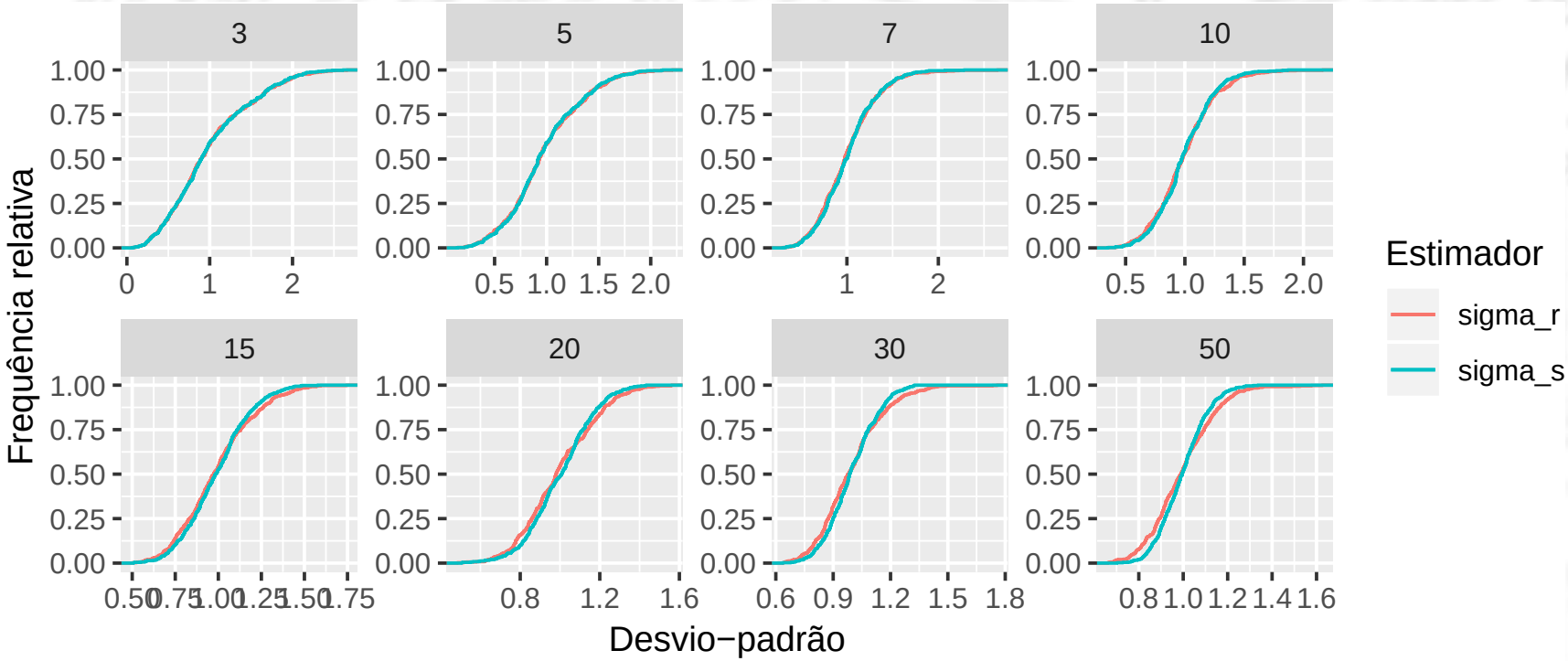
n	$\text{Var}(\hat{\sigma}_s)/\text{Var}(\hat{\sigma}_R)$
2	1.000
3	0.992
4	0.975
5	0.955
6	0.930
10	0.850

Estudo de simulação

```r
library(IQCC)
# ls("package:IQCC")

simul <- function(n, replications = 500) {
    res <- replicate(replications, {
        x <- rnorm(n)
        r <- diff(range(x))
        s <- sd(x)
        c(sigma_s = s/c4(n), sigma_r = r/d2(n))
    })
    res <- data.frame(n = n, as.data.frame(t(res)))
    return(res)
}

set.seed(12345)
n <- c(3, 5, 7, 10, 15, 20, 30, 50)
results <- map_df(n, simul)

ggplot(results, aes(x = sigma_r, y = sigma_s)) +
    facet_wrap(facets = ~n, nrow = 2) +
    geom_point(pch = 1, alpha = 0.5) +
    geom_abline(slope = 1, lty = 2)
```


3
5
7
10
15
20
30
50
sigma_s
sigma_r
0
1
2

```r
ggplot(results) +
    facet_wrap(facets = ~n, nrow = 2, scale = "free") +
    stat_ecdf(aes(x = sigma_r, color = "sigma_r")) +
    stat_ecdf(aes(x = sigma_s, color = "sigma_s")) +
    labs(color = "Estimador") +
    xlab("Desvio-padrão") +
    ylab("Frequência relativa")
```



```r
results %>%
    group_by(n) %>%
    summarise(var_sigma_s = var(sigma_s),
              var_sigma_r = var(sigma_r),
              eff = var_sigma_s/var_sigma_r)
```

```
# # A tibble: 8 x 4
#       n var_sigma_s var_sigma_r   eff
#   <dbl>       <dbl>       <dbl> <dbl>
# 1     3      0.261       0.268  0.975
# 2     5      0.131       0.138  0.951
# 3     7      0.0935      0.105  0.887
# 4    10      0.0558      0.0683 0.816
# 5    15      0.0354      0.0473 0.748
# 6    20      0.0273      0.0363 0.751
# 7    30      0.0174      0.0287 0.608
# 8    50      0.0115      0.0214 0.540
```

Inferência sobre a média de um processo - variância conhecida

- Considere a variável aleatória X normalmente distribuída, com média desconhecida μ e variância conhecida σ^2.
- Suponha que estejamos interessados em testar o seguinte par de hipóteses:

$$H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu \neq \mu_0,$$

sendo μ_0 algum valor especificado (ex: o alvo do intervalo de especificação).

- Dispondo-se de n observações independentes de X, o teste de hipóteses baseia-se na seguinte estatística:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Inferência sobre a média de um processo – variância conhecida

- Se a hipótese nula (H_0) for verdadeira, então a estatística Z tem distribuição normal padrão (N(0,1)).
- Essa distribuição serve como referência para testar a hipótese nula, de igualdade.
- Em qualquer teste de hipóteses estamos sujeitos a dois tipos de erros:
  - **Erro do tipo I:** Rejeitar a hipótese nula sendo que ela é verdadeira;
  - **Erro do tipo II**: Não rejeitar a hipótese nula sendo que ela é falsa.

Inferência sobre a média de um processo – variância conhecida

- Uma das forma de proceder o teste de hipóteses é fixar o **nível de significância** do teste, e tomar a decisão com base na regra correspondente.
- O nível de significância do teste é a probabilidade (α) que admitimos para o erro do tipo I.
- Assim, devemos rejeitar a hipótese nula, em favor da alternativa, se $|Z_0| > |z_{\alpha/2}|$, sendo $z_{\alpha/2}$ o quantil $\alpha/2$ da distribuição normal padrão.
- Usualmente utilizamos $\alpha = 5\%$ ou $\alpha = 1\%$.

Inferência sobre a média de um processo - variância conhecida

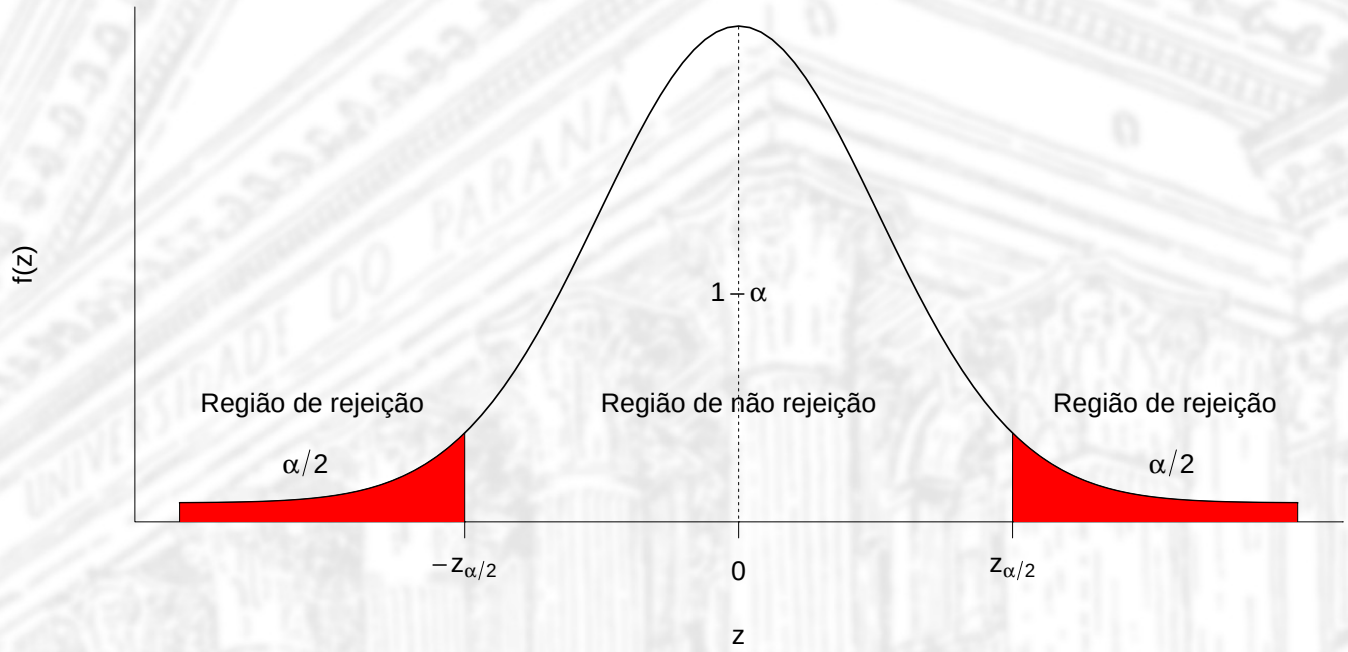


Figura 2. Teste de hipóteses - Tomada de decisão (nível. sig. α).

Inferência sobre a média de um processo - variância conhecida

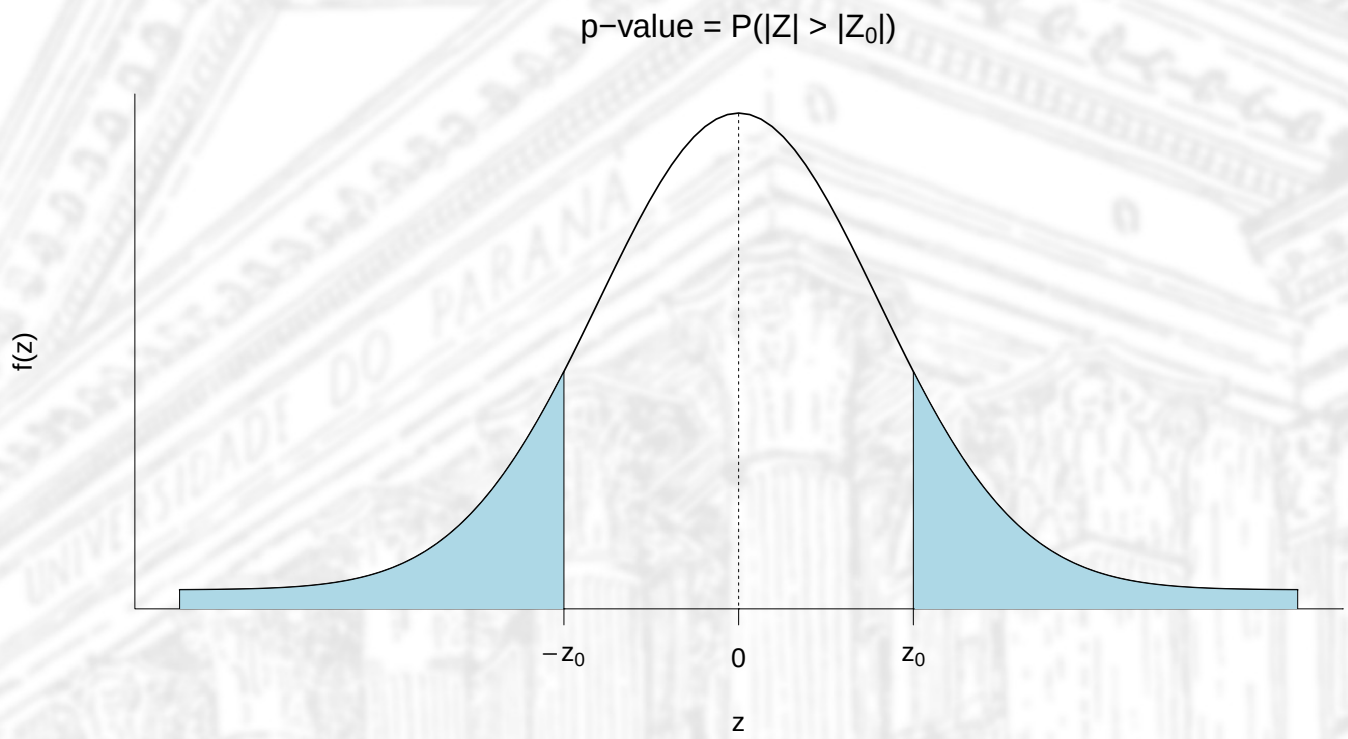


Figura 3. Teste de hipóteses - Ilustração do cálculo do *p*-valor.

Inferência sobre a média de um processo - variância conhecida

Exercício

O tempo de resposta de um sistema computacional é uma importante característica da qualidade. O gerente do sistema deseja saber se o tempo médio de resposta a um tipo específico de comando difere de 75 milisegundos. Da experiência passada ele sabe que o desvio padrão do tempo da resposta é 8 milisegundos. Sabendo que em $n = 25$ execuções do programa o tempo médio de resposta na amostra foi $\bar{x} = 78$ milisegundos:

1. Qual seria sua conclusão aos níveis de significância de 5% e 1%?
2. Calcule o valor p do teste. Interprete-o.

Solução

```r
# Quantidades conhecidas.
h0_mu <- 75
sd_x <- 8

# Quantidades determinadas no experimento.
x_bar <- 78
n <- 25

# Estatística de teste.
z_val <- (x_bar - 75)/(sd_x/sqrt(n))
z_val
```

```
# [1] 1.875
```

```r
# Valor p do teste de hipótese para H_0: \mu == 75 vs \mu != 75.
2 * pnorm(abs(z_val), lower.tail = FALSE)
```

```
# [1] 0.06079272
```

Inferência sobre a média de um processo - variância conhecida

- Dependendo do contexto, pode ser mais apropriado formular hipóteses unilaterais, como:

$$H_0 : \mu \leq \mu_0 \quad vs \quad H_1 : \mu > \mu_0,$$

rejeitando-se H_0, ao nível de significância α, se $Z_0 > Z_\alpha$ e

$$H_0 : \mu \geq \mu_0 \quad vs \quad H_1 : \mu < \mu_0,$$

rejeitando-se H_0, ao nível de significância α, se $Z_0 < -Z_\alpha$.

Inferência sobre a média de um processo - variância conhecida

- Um **intervalo de confiança** permite estimar parâmetros do processo usando conjecturas probabilísticas.
- Um intervalo de confiança $100(1-\alpha)\%$ para a média, considerando a variância populacional conhecida, é definido pelos seguintes limites:

$$\text{IC}(\mu; 100(1-\alpha)\%) = \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right).$$

- Interpretação: Para amostras aleatórias de tamanho n extraídas dessa população, em $100(1-\alpha)\%$ dos casos o intervalo calculado irá conter o valor desconhecido de μ.

Intervalos de confiança (95%) para a média

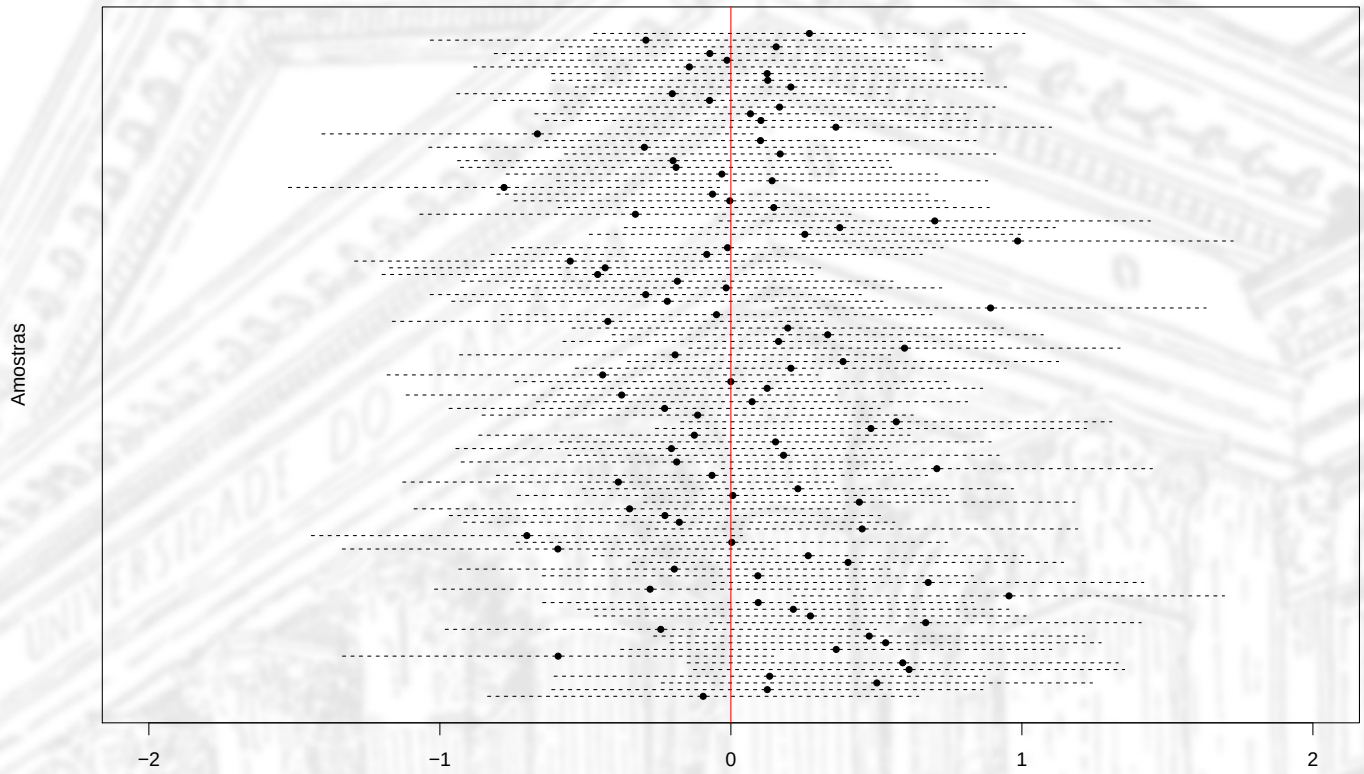
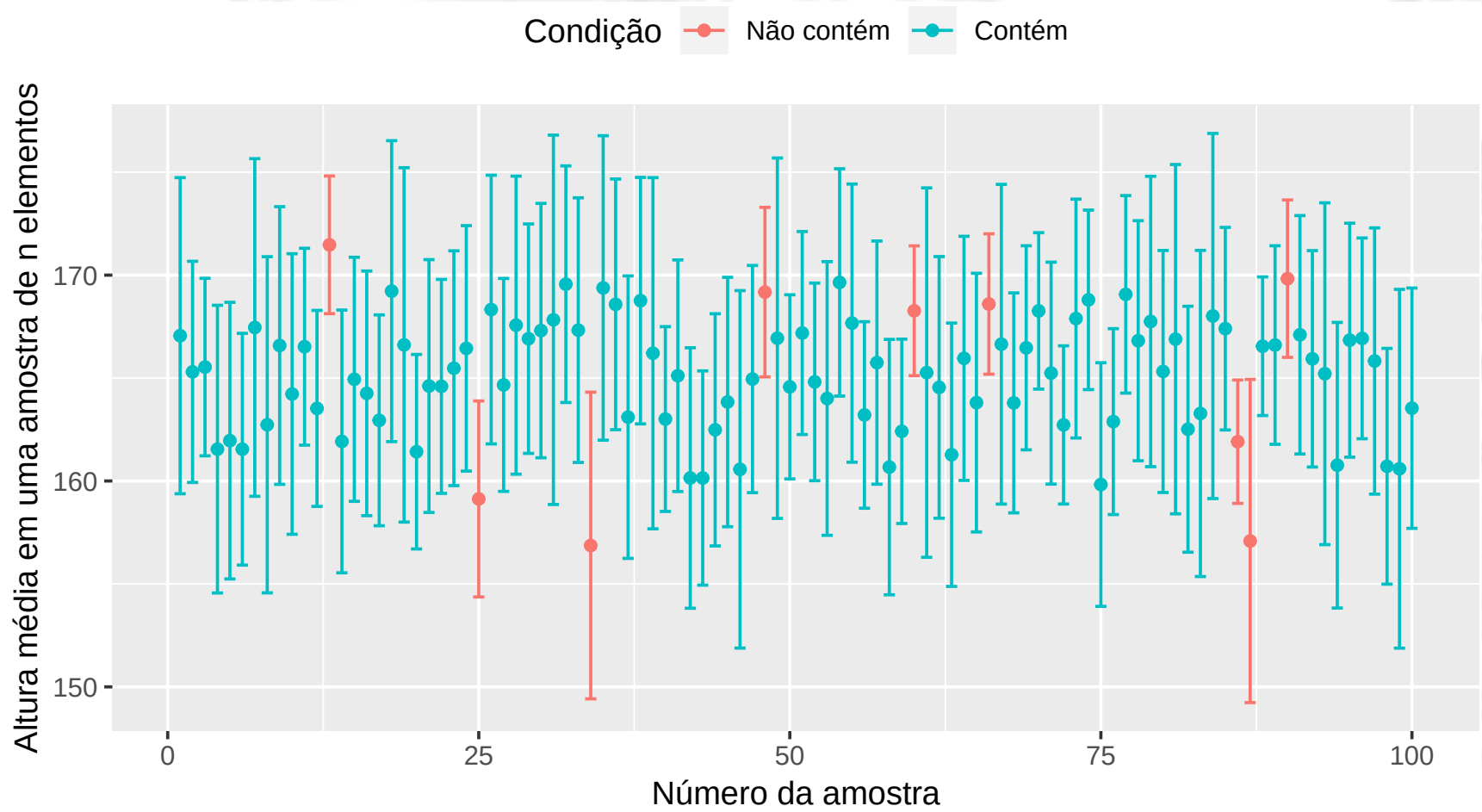


Figura 4. Intervalos de confiança.

Estudo de simulação

```r
simul <- function(n = 10, confidence = 0.95,
                  dist = "rnorm", params = c(mean = 0, sd = 1)) {
    x <- invoke(dist, c(list(n = n), as.list(params)))
    m <- mean(x)
    s <- sd(x)/sqrt(n)
    q <- qnorm((1 - confidence)/2)
    c(m + q * c(lwr = 1, est = 0, upr = -1) * s)
}

tb <- rerun(100, simul(n = 10, params = c(mean = 165, sd = 10))) %>%
    invoke(rbind, .x = .) %>%
    as_tibble() %>%
    mutate(i = 1:n(),
           status = (lwr < 165) & (upr > 165))

ggplot(tb, aes(x = i, y = est, color = status)) +
    geom_point() +
    geom_errorbar(aes(ymin = lwr, ymax = upr)) +
    xlab("Número da amostra") +
    ylab("Altura média em uma amostra de n elementos") +
    scale_color_discrete(name = "Condição",
                         labels = c("Não contém", "Contém")) +
    theme(legend.position = "top")
```


Condição
Não contém
Contém
Altura média em uma amostra de n elementos
170
160
150
0
25
50
75
100
Número da amostra

Inferência sobre a média de um processo – variância conhecida

Exercício

Considere novamente o exemplo anterior. Calcule intervalos de 95 e 99% de confiança para a média populacional.

Solução

```r
n <- 25
sd_x <- 8
x_bar <- 78
x_bar + c(lwr = -1, upr = 1) * qnorm(0.975) * sd_x/sqrt(n)
```

```
#      lwr      upr
# 74.86406 81.13594
```

Inferência sobre a média de um processo - variância desconhecida

- Caso a variância populacional seja desconhecida, deve ser estimada pela variância amostral $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$.
- A estatística teste fica definida por:

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}.$$

- A distribuição de referência, nesse caso, é a t-Student com $n-1$ graus de liberdade, devendo-se rejeitar H_0, ao nível de significância α, se $t_0 > |t_{n-1,\alpha/2}|$.

Inferência sobre a média de um processo - variância desconhecida

- De maneira semelhante, a distribuição t_{n-1} serve de referência para a construção de intervalos de confiança para μ :

$$\text{IC}(\mu; 100(1-\alpha)\%) = \left(\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}; \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \right).$$

Inferência para uma proporção populacional

- Suponha que se deseja testar o seguinte par de hipóteses:

$$H_0 : p = p_0 \quad vs \quad H_1 : p \neq p_0,$$

sendo p a proporção populacional desconhecida e p_0 algum valor especificado.

- Seja $\hat{p}$ a proporção amostral avaliada em uma amostra aleatória de tamanho n. O teste de hipóteses baseia-se na seguinte estatística teste:

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Inferência para uma proporção populacional

- De maneira similar ao teste de hipótese para a média com variância conhecida, deve-se rejeitar H_0, ao nível de significância α, se $|Z_0| > |z_{\alpha/2}|$.
- Testes de hipóteses unilaterais também podem ser aplicados à proporção.
- Um intervalo de confiança (assintótico) $100(1-\alpha)\%$ para a proporção tem limites:

$$\text{IC}(p; 100(1-\alpha)\%) = \left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right).$$

Exercício

Uma fundição produz cabos de aço usados na indústria automotiva. Deseja-se testar a hipótese de que a fração de itens não conformes é de 10%. Em uma amostra aleatória de 250 cabos, detectou-se que 32 estavam fora das especificações.

- Qual seria sua conclusão aos níveis de significância de 5% e 1%?
- Calcule o p-valor do teste. Interprete-o.
- Apresente um intervalo de confiança 95% para a fração de itens não conformes gerados pelo processo.

Solução

```r
# Dados fornecidos.
h0_p <- 0.1
n <- 250
p <- 32/n

# Estatística de teste.
z <- (p - h0_p)/sqrt(h0_p * (1 - h0_p)/n)
z
```

```
# [1] 1.47573
```

```r
# Valor p do teste para a hipótese H_0: p == 0.1 vs p != 0.1.
2 * pnorm(abs(z), lower.tail = FALSE)
```

```
# [1] 0.1400165
```

```r
# Intervalo de confiança.
p + c(-1, 1) * qnorm(0.975) * sqrt(p * (1 - p)/n)
```

```
# [1] 0.08658656 0.16941344
```

Solução pronta no R

```r
# Teste binomial exato para H_0: p == 0.1 vs p != 0.1.
binom.test(x = 32, n = 250, p = 0.1)
```

```
#
#   Exact binomial test
#
# data:  32 and 250
# number of successes = 32, number of trials = 250, p-value = 0.1399
# alternative hypothesis: true probability of success is not equal to 0.1
# 95 percent confidence interval:
#  0.08922514 0.17586971
# sample estimates:
# probability of success
#                  0.128
```

```r
# Teste para proporção de uma amostra com correção de contínuidade
# para H_0: p == 0.1 vs p != 0.1.
prop.test(x = 32, n = 250, p = 0.1)
```