Estimação

Fernando B. Sabino da Silva

Contents

1 Estimação por ponto e por intervalo

1.1 Estimação por ponto e por intervalo

- Nós queremos investigar hipóteses sobre parâmetros (constantes populacionais). Exemplo: a média μ e o desvio-padrão σ.
  - Se μ for o tempo médio de espera em uma fila, então pode ser relevante estudar se o tempo médio excede, por exemplo, a 2 minutos.
- Com base em uma amostra, nós calculamos uma **estimativa pontual** que de maneira coloquial é o nosso melhor "chute" para o valor do parâmetro.
  - Por exemplo, nós usamos $\bar{y}$ como uma estimativa para μ e s como uma estimativa para σ.
- Em geral, nós estamos também interessados em calcular uma **estimativa por intervalo** (também chamada de **intervalo de confiança**). Este intervalo é construído em torno da estimativa por ponto.
- A estimativa do parâmetro (e o intervalo de confiança) pode ser utilizado para investigar a hipótese.

1.2 Estimadores por ponto: Viés

- Se queremos estimar a média da população μ, nós temos várias possibilidades (a princípio). Exemplos:
  - a média amostral $\bar{y}$
  - a média amostral y_T dos quartis superior (Q_3) e inferior (Q_1).

- Vantagem de y_T: Pouca influência de outliers (observações com valor muito alto/baixo), i.e. não teremos praticamente nenhum efeito se houver alguns erros no banco de dados.
- Desvantagem de y_T: Se a distribuição da população for assimétrica, então y_T será um estimador **viesado**, o que significa que no longo prazo este estimador sistematicamente será acima ou abaixo do verdadeiro valor de μ.
- Geralmente, nós preferimos que um estimador seja **não viesado**, isto é, que a sua distriuição seja centrada em volta do verdadeiro valor do parâmetro.
- Relembre que para uma população com média μ, a média amostral $\bar{y}$ também tem média μ, i.e., $\bar{y}$ é um estimador não viesado da média populacional μ.

1.3 Estimadores por ponto: Eficiência

- Nós já sabemos (mostramos em aula - verifique se você sabe **provar** isto) que o erro padrão de $\bar{y}$ é $\frac{\sigma}{\sqrt{n}}$, i.e. o erro padrão converge para zero quando o tamanho da amostra aumenta (você saberia explicar a intuição disto?).
- É difícil expressar o erro padrão de y_T, mas é possível provar que ele será maior do que o erro padrão de $\bar{y}$. Este é um bom motivo para que, usualmente, $\bar{y}$ seja um estimador preferível.
- Em geral, nós preferimos que um estimador seja **eficiente**. De maneira coloquial, isto significa que o erro padrão converge para zero quando o tamanho da amostra aumenta. O estimador **mais eficiente** será aquele que converge para zero a uma velocidade maior. No nosso exemplo, para a maioria das populações, y_T é ineficiente.

1.4 Notação

- O símbolo ˆ acima de um parâmetro é frequentemente utilizado para denotar uma estimativa (pontual) de um parâmetro. Exemplos:
  - estimativa da média populacional: $\hat{\mu} = \bar{y}$
  - estimativa do desvio padrão: $\hat{\sigma} = s$
- Quando observamos uma variável binária (0/1), o que, por exemplo, é utilizada para denotar sim/não ou masculino/feminino, então nós usamos a notação

$$p = P(Y = 1)$$

  para a proporção da população com a característica $Y = 1$.
- A estimativa $\hat{p} = (y_1 + y_2 + \ldots + y_n)/n$ é a frequência relativa amostral de $Y = 1$.

1.5 Intervalo de Confiança

- A definição geral de um intervalo de confiança para um parâmetro populacional é a seguinte:
  - Um **intervalo de confiança** para um parâmetro é um intervalo construído com base na amostra, isto é, ele é aleatório (depende da amostra em particular). Esperamos que este intervalo contenha o verdadeiro valor do parâmetro (que não é aleatório, é uma constante populacional).
  - A "probabilidade" de que esta construção produza um intervalo que inclua o verdadeiro valor do parâmetro é chamada de **nível de confiança** (ou de cobertura). Tipicamente, o nível de confiança escolhido é de 95%.
  - (1-nível de confiança) é chamado de **nível de significância** (neste exemplo, $1-0.95 = 0.05$, i.e. 5%).
- Frequentemente, o intervalo é construído como um intervalo simétrico em torno de uma estimativa por ponto:
  - **estimativa por ponto ±margem de erro**
  - Regra de bolso: Com uma margem de erro de aproximadamente 1.96 vezes o erro padrão você obtém um intervalo de confiança de aproximadamente 95%.

– i.e: **estimativa por ponto ± 1.96 x erro padrão** tem nível de confiança de aproximadamente 95%.

1.6 Intervalo de Confiança para a Proporção

- Considere uma população com uma distribuição onde a probabilidade de ter uma determinada característica seja p e a probabilidade de não ter seja $1 - p$.
- Quando as categorias *não*/*sim* são denotadas por 0/1, i.e. y é 0 ou 1, a distribuição de y tem um desvio padrão de :

  $$\sigma = \sqrt{p(1-p)}.$$

  Isto é, o desvio padrão não é um parâmetro "livre" para uma variável 0/1, pois o seu valor está diretamente ligado a probabilidade p.
- Com uma amostra de tamanho n o erro padrão de $\hat{p}$ será (dado que $\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n}$):

  $$\sigma_{\hat{p}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}.$$

- Nós não sabemos o valor de p, mas se inserirmos a estimativa iremos obter o **erro padrão estimado** de $\hat{p}$:

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

- A regra de bolso nos diz que o intervalo

  $$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  tem nível de confiança de aproximadamente 95%. i.e., antes que os dados sejam conhecidos, o intervalo aleatório dado pela fórmula acima tem aproximadamente 95% de "probabilidade" de conter o verdadeiro valor de p.

1.6.1 Exemplo: Estimativa por ponto e por intervalo para a proporção

- Vamos dar uma olhada em dados de uma pesquisa nacional conduzida no Chile entre Abril e Maio de 1988. Informações sobre os dados podem ser encontradas aqui.

```
Chile <- read.delim("C:/Users/fsabino/Desktop/Codes/papers/Introductory_Stat_II/notebook/Chile.txt")
```

- Concentremo-nos na variável `sex`, i.e. a distribuição de gênero na amostra.

```
library(mosaic)
```

```
## Warning: package 'mosaic' was built under R version 3.4.3
## Warning: package 'dplyr' was built under R version 3.4.3
## Warning: package 'ggformula' was built under R version 3.4.3
## Warning: package 'ggplot2' was built under R version 3.4.3
## Warning: package 'mosaicData' was built under R version 3.4.3
```

```
tally( ~ sex, data = Chile)
```

```
## sex
##    F    M
## 1379 1321
```

```
tally( ~ sex, data = Chile, format = "prop")
```

```
## sex
##         F         M
## 0.5107407 0.4892593
```

- Proporção da população (desconhecida) de mulheres (F), p.
- Estimativa de p: $\hat{p} = \frac{1379}{1379+1321} = 0.5107$
- Regra de bolso : $\hat{p} \pm 1.96 \times ep = 0.5107 \pm 2\sqrt{\frac{0.5107(1-0.5107)}{1379+1321}} = (0.4919, 0.5296)$ é um intervalo de confiança aproximado de 95% para p.

1.6.2 Exemple: Intervalos de confiança para a proporção no R

- **R** automaticamente calcula o intervalo de confiança para a proporção de pessoas do sexo feminino quando nós fazemos um teste de hipóteses (voltaremos a isso mais adiante):

```
prop.test( ~ sex, data = Chile, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  Chile$sex  [with success = F]
## X-squared = 1.2459, df = 1, p-value = 0.2643
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4918835 0.5295675
## sample estimates:
##         p
## 0.5107407
```

- O argumento `correct = FALSE` é necessário para pedir ao **R** para fazer uma aproximação para a distribuição normal como feito nestas notas. Quando `correct = TRUE` (o default) uma correção matemática que você não aprendeu se aplica e os resultados serão ligeiramente diferentes.

1.7 Intervalos de confiança aproximados para a proporção

- Com base no teorema central do limite (CLT), nós temos :

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

ep $n\hat{p}$ e $n(1-\hat{p})$ são grandes o suficiente para que a aproximação seja válida (maiores do que 15, por exemplo).

- Para construir um intervalo de confiança com nível de confiança (aproximado) $1-\alpha$:
  1) Encontre o valor crítico z_{crit} para o qual a probabilidade na cauda superior da distribuição normal seja $\alpha/2$.
  2) Calcule $ep = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

3) Então $\hat{p} \pm z_{crit} \times ep$ é um intervalo de confiança com nível de confiança $1 - \alpha$.

1.7.1 Exemplo: Dados do Chile

Para os dados do `Chile` calcule os intervalos de confiança 99% e 95% para a probabilidade de que uma pessoa seja do sexo feminino:

- Para um nível de confiança de 99%, temos $\alpha = 1\%$ e
  1) z_{crit}=`qdist("norm", 1 - 0.01/2)`=2.576.
  2) Sabemos que $\hat{p} = 0.5107$ e $n = 2700$, então $ep = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0096$.
  3) Assim, um intervalo de confiança de 99% é: $\hat{p} \pm z_{crit} \times ep = (0.4859, 0.5355)$.
- Para um nível de confiança de 95%, temos $\alpha = 5\%$ e
  1) z_{crit}=`qdist("norm", 1 - 0.05/2)`=1.96.
  2) Novamente, $\hat{p} = 0.5107$ and $n = 2700$ e assim $ep = 0.0096$.
  3) Deste modo, nós encontramos um intervalo de confiança de 95%: $\hat{p} \pm z_{crit} \times ep = (0.4918, 0.5295)$ (como resultado de `prop.test`).

1.8 Intervalo de confiança para a média - amostra retirada de uma população com distribuição normal

- Quando é razoável supor que a distribuição da população é normal, nós temos o resultado **exato**

$$\bar{y} \sim \mathtt{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

  i.e. $\bar{y} \pm z_{crit} \times \frac{\sigma}{\sqrt{n}}$ não é mais apenas um intervalo de confiança aproximado (como no caso da proporção - por que é aproximado neste caso?), mas sim um intervalo de confiança exato para a média populacional, μ.
- Na prática, porém, **nós não conhecemos** σ e ao invés disso, nós somos obrigados a utilizar o desvio padrão da amostra s para encontrar o **erro padrão estimado** $ep = \frac{s}{\sqrt{n}}$.
- Esta incerteza extra, no entanto, implica que um intervalo de confiança exato para a média populacional μ não pode ser construído usando o escore-z.
- Um intervalo exato ainda pode ser construído usando o chamado **escore-t**, que além do nível de confiança depende dos **graus de liberdade** (degrees of freedom = df), que neste caso são $df = n - 1$. Isto é, o intervalo de confiança toma agora a forma

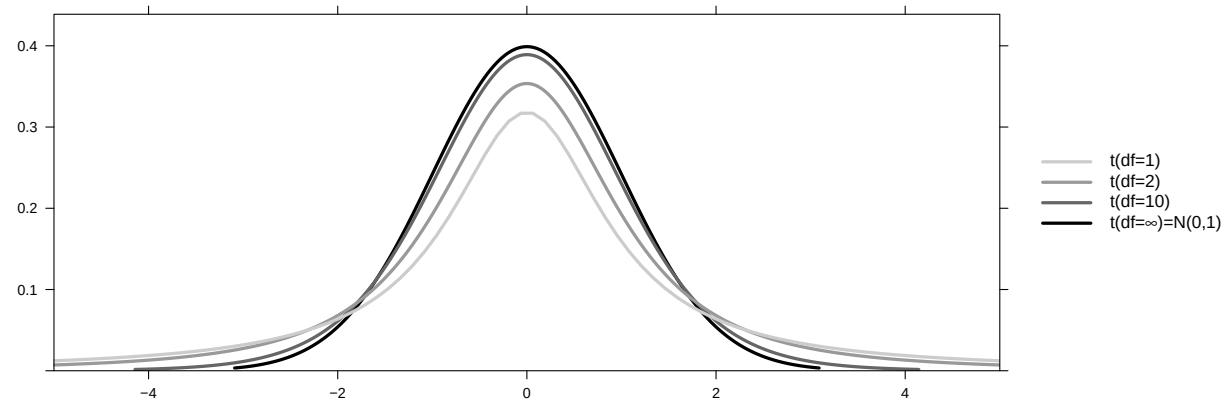
$$\bar{y} \pm t_{crit} \times ep.$$

- Nota: É importante aprender as relações entre as distribuições normal, t de Student, qui-quadrado e F. Veja mais detalhes em Costa Neto (Estatística) e Casella and Berger (Statistical Inference, traduzido para português). Mais detalhes serão vistos em sala de aula.

1.9 A distribuição t e o escore t

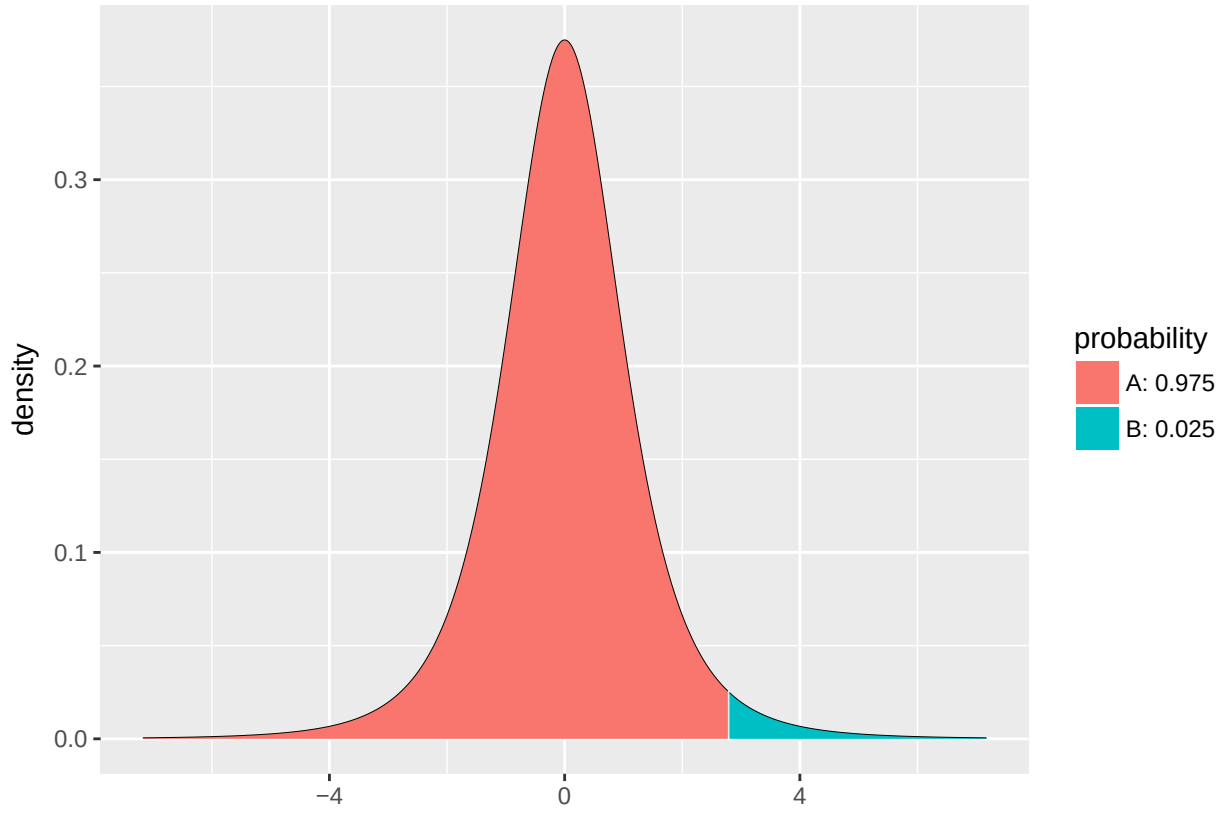
- O cálculo do escore t é baseado na **distribuição** t, que é semelhante a distribuição normal padrão z:
  - ela é simétrica em torno de zero e é uma função em forma de "sino", mas
  - como vimos tem caudas mais "pesadas" e portanto
  - um desvio padrão maior do que o desvio padrão da distribuição normal padrão.
  - Note que o desvio padrão da distribuição t decaí em função de seus **graus de liberdade** (que denotamos por df).
  - e quando df cresce a distribuição t se aproxima da distribuição normal padrão.

A expressão da função densidade não será indicada aqui (pode ser encontrada nos livros sugeridos ou no google). Ao invés disso, a distribução t é representada abaixo para $df = 1, 2, 10$ e ∞.



1.9.1 Cálculo do escore t no R

```
qdist("t", p = 1 - 0.025, df = 4)
```




```
## [1] 2.776445
```

- Um escore t é o quantil (i.e. o valor no eixo x) para o qual temos uma dada probabilidade na **cauda direita**.
- Para obter, por exemplo, um escore t correspondente a uma probabilidade na cauda direita de 2.5 % nós temos que procurar o quantil 97.5 % usando `qdist` with `p = 1 - 0.025`, pois `qdist` olha a área para o **lado esquerdo**.
- Os graus de liberdade são determinados pelo tamanho da amostra. No exemplo anterior usamos df = 4 para ilustração.
- Como um escore t para uma probabilidade na cauda direita de 2.5 % é 2.776 e a distribuição t é simétrica em torno do 0, nós temos que uma observação tem probabilidade de 1 - 2 · 0.025 = 95 % de estar entre -2.776 and 2.776 para uma distribuição t com 4 graus de liberdade.

1.10 Exemplo: Intervalo de Confiança para a média

- Em estatística I usamos o conjunto de dados de `Ericksen`. Queremos agora construir um intervalo de confiança de 95% para a média da população mean μ da variável `crime`.

```
Ericksen <- read.delim("C:/Users/fsabino/Desktop/Codes/papers/Introductory_Stat_I/notebook/datasets_Eri
stats <- favstats( ~ crime, data = Ericksen)
stats
```

```
##  min Q1 median Q3 max     mean       sd  n missing
##   25 48     55 73 143 63.06061 24.89107 66       0
```

```
qdist("t", 1 - 0.025, df = 66 - 1, plot = FALSE)
```

```
## [1] 1.997138
```

- i.e., nós temos
  - $\bar{y} = 63.061$
  - $s = 24.891$
  - $n = 66$
  - $df = n - 1 = 65$
  - $t_{crit} = 1.997$.
- O intervalo de confiança é $\bar{y} \pm t_{crit} \frac{s}{\sqrt{n}} = (56.94, 69.18)$
- Todos estes cálculos podem ser feitos automaticamente no **R**:

```
t.test( ~ crime, data = Ericksen, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  crime
## t = 20.582, df = 65, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  56.94162 69.17960
## sample estimates:
## mean of x
##  63.06061
```

1.11 Exemplo: Fazendo vários intervalos de confiança no R

- Vamos olhar o conjunto de dados `chickwts` que já vem integrado no **R**.
- `?chickwts` produz uma página com a seguinte informação

Um experimento foi conduzido para medir e comparar a eficácia de vários suplementos alimentares sobre a taxa de crescimento de galinhas Os filhotes recém-nascidos foram alocados aleatoriamente em seis grupos, e cada grupo recebeu um suplemento alimeantar diferente. Seus pesos em gramas após seis semanas são dados juntamente com os tipos de alimentação.

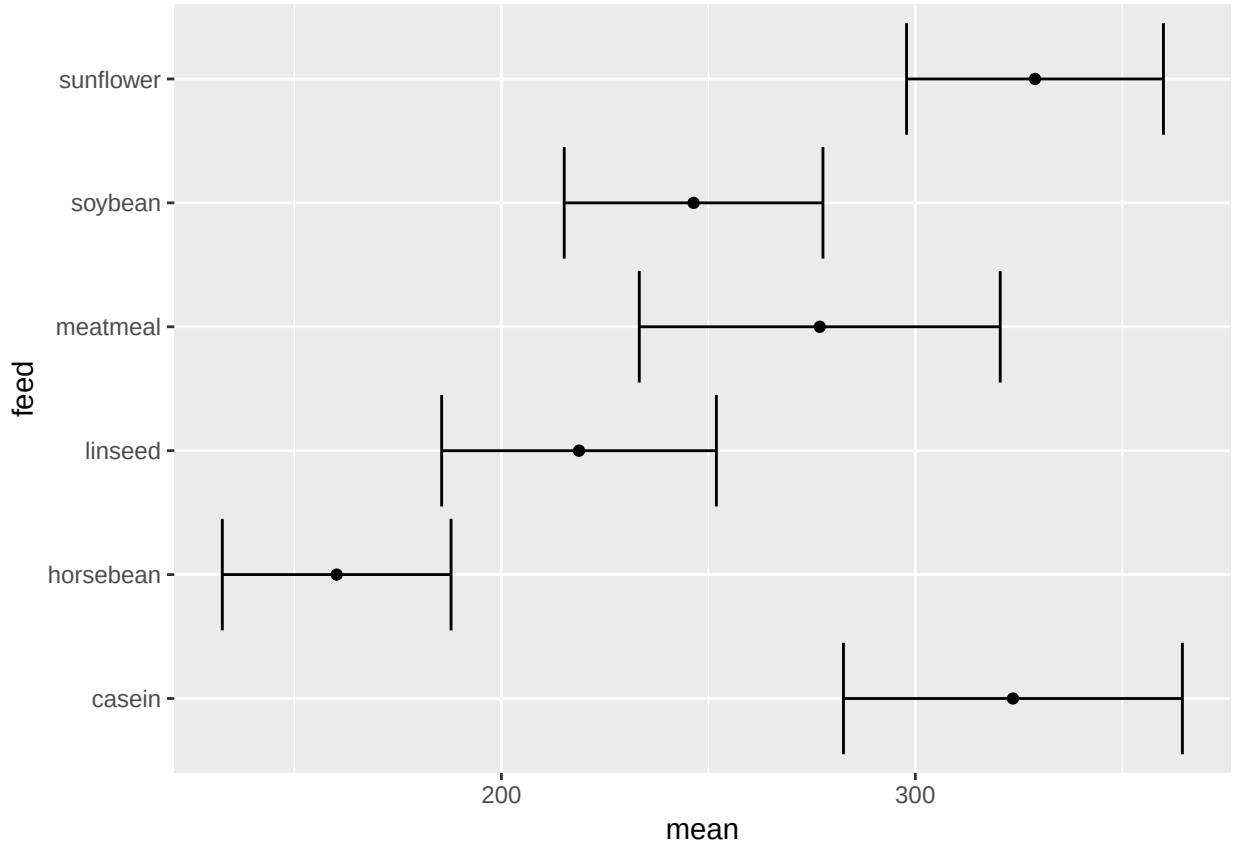
- `chickwts` é um data frame com 71 observações e 2 variáveis:
  - `weight`: é uma variável numérica que representa o peso do filhote.
  - `feed`: um fator (variável qualitativa/categórica) que representa o tipo de alimentação.
- Calcule um intervalo de confiança para o peso médio de cada alimentação separadamente; o intervalo de confiança é de inferior(`lower`) para superior (`upper`) dado por média ± escore t * erro padrão (`mean±tscore * ep`):

```
cwei <- favstats( weight ~ feed, data = chickwts)
ep <- cwei$sd / sqrt(cwei$n) # Erros padrão
tscore <- qdist("t", p = .975, df = cwei$n - 1, plot = FALSE) # Escores t para uma probabilidade de cad
cwei$lower <- cwei$mean - tscore * ep
cwei$upper <- cwei$mean + tscore * ep
cwei[, c("feed", "mean", "lower", "upper")]
```

```
##        feed     mean    lower    upper
## 1    casein 323.5833 282.6440 364.5226
## 2 horsebean 160.2000 132.5687 187.8313
## 3   linseed 218.7500 185.5610 251.9390
## 4  meatmeal 276.9091 233.3083 320.5099
## 5   soybean 246.4286 215.1754 277.6818
## 6 sunflower 328.9167 297.8875 359.9458
```

- Nós podemos traçar os intervalos de confinaça como segmentos de linhas horizontais usando a função `gf_errorbarh`:

```
gf_errorbarh(feed ~ mean + lower + upper, data = cwei) %>%
  gf_point(feed ~ mean)
```



2 Determinando o tamanho da amostra

2.1 Tamanho da amostra para a proporção

- O intervalo de confiança é da forma estimativa por ponto ± margem de erro estimada.
- Quando nós estimamos uma proporção, a margem de erro é

$$e = z_{crit}\sqrt{\frac{p(1-p)}{n}},$$

onde o escore z crítico, z_{crit}, é determinado pelo nível de confiança especificado.
- Imagine que nós queremos planejar um experimento, onde **desejamos obter uma certa margem de erro** e (e portanto uma largura específica do intervalo de confiança associado).
- Se nós resolvermos a equação acima, é possível notar:
  - Se $n = p(1-p)(\frac{z_{crit}}{e})^2$, nós obtemos uma estimativa de p com margem de erro e.
- Se não tivermos um bom palpite para o valor de p, nós podemos usar o valor do "pior" caso, isto é, $p = 50\%$ (1/4 é o maior valor possível que pode assumir a função $p(1-p)$). O tamanho amostral correspondente $n = (\frac{z_{crit}}{2e})^2$ garante que nós iremos obter uma estimativa com uma margem de erro que é no *máximo* e.

2.1.1 Exemplo

- Vamos escolher $z_{crit} = 1.96$, i.e o nível de confiança é de 95%.

- Qual o número de eleitores que devemos entrevistar para obter uma margem de erro igual a 1%?
- O pior caso é $p = 0.5$, o que nos leva a:

$$n = p(1-p)\left(\frac{z_{crit}}{e}\right)^2 = \frac{1}{4}\left(\frac{1.96}{0.01}\right)^2 = 9604.$$

- Se nós estivermos interessados na proporção de votos para o candidato de um partido cujo um bom palpite é de no máximo $p = 0.23$, teríamos

$$n = p(1-p)\left(\frac{z_{crit}}{e}\right)^2 = 0.23(1-0.23)\left(\frac{1.96}{0.01}\right)^2 = 6804.$$

- Se nós estivermos interessados na proporção de votos para o candidato de um partido cujo um bom palpite é de no máximo $p = 0.05$, teríamos

$$n = p(1-p)\left(\frac{z_{crit}}{e}\right)^2 = 0.05(1-0.05)\left(\frac{1.96}{0.01}\right)^2 = 1825.$$

2.2 Tamanho da amostra para a média

- O intervalo de confiança é da forma estimativa por ponto ± margem de erro estimada.
- Quando estimamos uma média com desvio-padrão conhecido a margem de erro é

$$e = z_{crit}\frac{\sigma}{\sqrt{n}},$$

  onde um escore z crítico, z_{crit}, é determinado pelo nível especificado de confiança.
- Imagine que nós queremos planejar um experimento, onde **desejamos obter uma certa margem de erro** e.
- Se nós resolvermos a equação acima, é possível notar:
  - Se $n = (\frac{z_{crit}\sigma}{e})^2$, nós obtemos uma estimativa com margem de erro e.
- Problema: Nós usualmente não sabemos σ. Possíveis soluções:
  - Com base em estudos similares realizados anteriormente, nós fazemos um palpite "educado" sobre o valor de σ.
  - Baseado em um estudo piloto estimamos o valor de σ.

3 Exercícios e Leituras Recomendadadas

1. Qual deve ser o valor do desvio padrão para que uma distribuição normal com média 9 cubra o intervalo 0-18 com probabilidade de 99.7%?

2. Leia a seção 3.4 de Costa Neto (doravante CN) e verifique o que foi dado em aula sobre o assunto.

3. Leia as seções 4.4 e 4.5 de CN e refaça os exemplos (tem solução).

4. Faça os exercícios 1, 3, 5, 7, 14, 16, 17, 18 e 21 da seção 4.6 (Exercícios propostos) de CN.

4 Apêndice

- Instale o pacote TeachingDemos: install.packages("TeachingDemos"). Após a instalação carregue o pacote para poder usá-lo: library(TeachingDemos).
- Use a função ci.examp() para visualizar 50 intervalos de confiança. Você pode visualizar os argumentos usados (by default) na função pedindo ajuda: ?ci.examp

- Tente alterar alguns dos argumentos da função e utilize method = “t”. Experimente, por exemplo, gerar 40 amostras aleatórias (reps = 40) e “depois peça os gráficos correspondentes com”plote" os correspondentes 40 intervalos de confiança usando um nível de confiança de 90% (conf.level = 0.90).
- Convença-se de que esperamos que 4 destes intervalos de confiança não contenham a verdadeira média populacional. Como isso se encaixa com o que você está vendo?
- Experimente também a função clt.examp(). Visualize os argumentos pedindo ajuda: ?clt.examp
- Use a função windows() para abrir um dispositivo gráfico.
  - clt.examp()
  - clt.examp(5)
  - clt.examp(30)
  - clt.examp(50)