### 4.5.2. Usando simulações para avaliar a evidência de um resultado

Nesta seção, iremos simular uma amostra de peixes (ou melhor, de seus pesos), usando o R. Para isto, precisamos considerar estes pesos como uma variável aleatória contínua, e atribuir um modelo para esta variável. Como uma maior parte das variáveis biológicas têm distribuições razoavelmente próximas do modelo normal (ou, pelo menos, unimodais e razoavelmente simétricas), usaremos um modelo normal para os pesos. No texto abaixo, mostramos os comandos em R (usando fonte `Courier New`), e comentamos os resultados. A rotina completa está disponível num arquivo *.R.

Os passos são:

1. Apagamos tudo o que estiver na memória do R (resultados de cálculos feitos anteriormente, etc.):
   ```
   rm(list=ls(all=TRUE))
   ```

2. Fixamos a "semente" (*seed*) do gerador de números aleatórios (para que os resultados que você obterá quando rodar a rotina sejam iguais aos que estão neste texto. Experimente depois com outros valores de semente):
   ```
   set.seed(1313)
   ```

3. Fixamos os valores de média ($\mu$) e desvio-padrão ($\sigma$) dados pela hipótese nula, que serão os parâmetros da distribuição normal do peso:
   ```
   # parametros
   mu=28.0
   sigma=4.0
   ```

4. Simulamos o peso de um peixe – uma observação de uma variável aleatória normal, com os parâmetros acima. O comando `rnorm` gera um número aleatório (*random*) de uma dis-tribuição *normal* (**r**andom **norm**al), com média (*mean*) igual a *mu*, e desvio-padrão (*stan-dard deviation, sd*) igual a *sigma* :
   ```
   x1=rnorm(1,mean=mu,sd=sigma)
   x1
   ```

5. Simulamos os pesos de mais dois peixes:
   ```
   x2=rnorm(1,mean=mu,sd=sigma)
   x2
   x3=rnorm(1,mean=mu,sd=sigma)
   x3
   ```

Os valores encontrados foram:
```
x1 = 32.75742
x2 = 30.20554
x3 = 29.41454
```

O gráfico de pontos da Fig. **1** mostra a posição destes três valores no eixo dos pesos (este gráfico não pode ser feito no R):

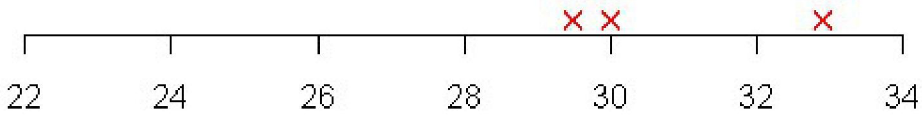


**Fig. 1. Pesos (simulados) de três peixes**

Note que os três valores simulados acima estão bem acima da média hipotética. Não é possível, contudo, tirar nenhuma conclusão válida a partir de apenas três valores.

Simulemos agora três amostras de n=60 peixes cada e calculemos suas médias :

```
# simulando uma amostra de n=60
n=60
x1=rnorm(n,mean=mu,sd=sigma)
med1=mean(x1)
med1
# simulando uma segunda amostra
x2=rnorm(n,mean=mu,sd=sigma)
med2=mean(x2)
med2
# simulando uma terceira amostra
x3=rnorm(n,mean=mu,sd=sigma)
med3=mean(x3)
med3
```

| | | |
|---|---|---|
| 14\| | 14\| | 14\| |
| 16\| | 16\| | 16\|0 |
| 18\|24 | 18\|1 | 18\| |
| 20\|57 | 20\| | 20\|17 |
| 22\|39189 | 22\|8579 | 22\|5633367 |
| 24\|457889123456778 | 24\|2558336 | 24\|150111779 |
| 26\|0123401225 | 26\|01334557900122346668 | 26\|3392359 |
| 28\|12568237 | 28\|022347224889 | 28\|046677113459 9 |
| 30\|12344911388 | 30\|688028 | 30\|04447789112356 |
| 32\|1244 | 32\|0234566 | 32\|06689 |
| 34\|116 | 34\|5 | 34\|9 |
| 36\| | 36\| | 36\| |
| 38\| | 38\|0 | 38\|4 |
| Amostra 1 | Amostra 2 | Amostra 3 |

**Fig. 2. Diagramas de ramo-e-folhas das três amostras retiradas**

Note, nos diagramas de ramo-e-folhas mostrados na Fig. 2, que em todas as amostras foram encontrados peixes com pesos iguais ou menores que 26 kg (o mínimo peso encontrado foi de 16 kg, na amostra 3); no entanto em nenhuma das amostras foi encontrada uma *média* $\overline{X}$ igual ou menor que 26 kg.

As médias destas amostras foram:

```
med1 = 27.43253
med2 = 28.12835
med3 = 28.06148
```

O gráfico de pontos da Fig. 3 mostra a posição destas médias:

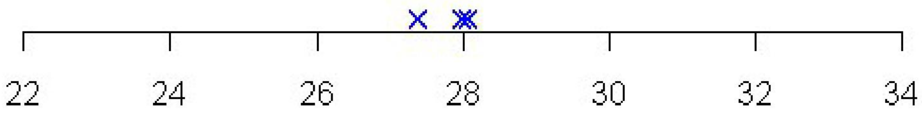


**Fig. 3. Médias de três amostras**

As três médias amostrais parecem bem próximas da média populacional hipotética (28 kg), de acordo com o que esperaríamos encontrar se hipótese fosse verdadeira; parecem portanto serem evidências a favor desta hipótese. Para termos mais evidências, simularemos agora mil amostras desta mesma população teórica:

```
# simulando n_am=1000 amostras de n=60
n_am=1000
media=c()
for (i in 1:n_am)
      {
      x= rnorm(n,mean=mu,sd=sigma);
      media[i]=mean(x)
      }
hist(media,breaks=20,xlim=c(21,35))
```

O gráfico da Fig. 4 mostra a histograma das médias destas 1000 amostras simuladas:

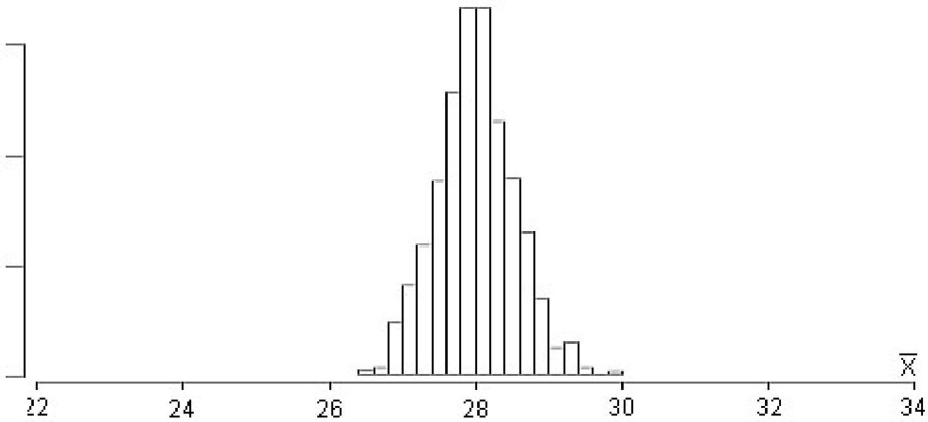


**Fig. 4. Histograma das médias das 1000 amostras**

Este gráfico mostrando como as médias das amostras se distribuem é uma simulação da *distribuição amostral das médias* ; veremos como é feito o cálculo teórico desta distribuição na próxima seção. Neste histograma, podemos notar, primeiro, que as médias das amostras realmente se concentram, como esperaríamos, em torno da média populacional, igual a 28,0 kg. (Veremos na próxima seção um teorema que diz que a média da distribuição amostral é igual à média populacional). Calculando com o R:

```
mean(media)
```

obtemos uma média 28.00033, que de fato foi praticamente idêntica ao que era esperado.

Em segundo lugar, podemos também observar que parece haver pouca variação entre estas 1000 médias. O teorema mencionado acima também diz que o desvio-padrão da distribuição amostral teórica é dado por:

```
errop=sigma/sqrt(n)
```

o que resulta em:

```
0.5163978
```

Calculando o desvio-padrão da distribuição simulada,

```
sd(media)
```

encontramos:

```
0.5209279
```

o que está bem próximo do valor teórico esperado.

Voltando ao problema do peso dos peixes: que podemos concluir, depois de todas estas simulações? Na amostra mencionada no enunciado do problema foi encontrada uma média de 26,0 kg; no entanto, nas 1000 amostras que simulamos, não encontramos nenhuma com uma média semelhante. Calculando com o R,

```
min(media)
```

Vemos que a mínima média observada foi de

```
26.46270
```

o que está acima de 26,0 kg. A amostra mencionada no problema, portanto, é muito diferente do que esperaríamos encontrar, se a hipótese nula fosse verdadeira.

Se o valor que encontramos na amostra não é o que esperávamos, o que de errado pode ter acontecido? Há vários fatores que podem ter influenciado no resultado:

1. Ocorreu algum erro na coleta dos dados dos peixes (erro na pesagem dos peixes, no registro dos pesos, nos cálculos, etc.). Erros deste tipo são geralmente chamados de “erros grosseiros”, e não há nada que os estatísticos possam fazer para corrigi-los; se quem obteve os dados (pescou os peixes e os pesou) não consegue identificar onde ocorreu o erro, a única solução seria começar tudo de novo, com outra amostra.

2. A amostra original não é uma *amostra aleatória simples*, e não podemos portanto compará-la com amostras simuladas (e também não podemos usar os teoremas que serão estudados na seção 4.5.3). Este é um problema mais complicado; na prática, as amostras usadas quase nunca serão “aleatórias simples” (uma amostra aleatória simples seria aquele em que todos os peixes do oceano fossem numerados, e depois fosse feito um sorteio para verificar quais iriam fazer parte da amostra; é obviamente impossível fazer isto). O que os estatísticos podem fazer é, com o auxílio de especialistas da área de interesse (no caso, de biólogos, ou de profissionais envolvidos na pesca daqueles peixes), decidir até que ponto uma amostra como a do problema poderia ser considerada *representativa* da população dos peixes. (Isto é discutido na seção 5.1).

3. A suposição de que os pesos dos peixes na população seguem uma distribuição normal está errada (fizemos todas as simulações a partir do modelo normal). Isto é relativamente fácil de verificar, pois existem vários métodos para testar a "normalidade" de uma popu-lação. Estes métodos vão desde a simples comparação visual do histograma da amostra com o gráfico de uma distribuição normal teórica (o histograma é unimodal e simétrico, como esperaríamos que acontecesse com uma amostra vinda de uma distribuição nor-mal?), até diversos testes estatísticos que avaliam a partir de amostras se existe evidência a favor ou contra a normalidade de uma população (veremos estes testes na seção 5.6). De qualquer forma, veremos na próxima seção que para uma amostra grande como esta (n=60) a suposição de normalidade não é realmente necessária, e conclusões podem ser tiradas mesmo que a população não seja normal.

4. O desvio-padrão da população não é $\sigma = 4{,}0$ kg, como era antes, segundo o enunciado. De fato, se o valor do desvio-padrão se alterou, todas as simulações acima são inúteis (experimente rodar a rotina fazendo sigma=20, e você verá que uma amostra de média 26,0 kg passa a ser possível, e não iria contradizer a hipótese). Esta suposição também pode ser testada; existem técnicas estatísticas que nos permitem estimarmos o desvio-padrão de uma população, a partir do desvio-padrão de uma amostra. Na prática, na maioria das vezes não sabemos qual é o desvio-padrão da população, e temos que estimá-lo a partir da amostra.

Se temos certeza de que resultado encontrado na amostra não foi afetado por nenhum dos quatro fatores mencionados acima, ele só pode ter acontecido devido a um último fator:

5. A hipótese de que o peso médio da população destes peixes continua sendo de 28,0 kg, como era no passado, é falsa!

Um teste feito desta maneira é didático, mas não tem aplicação prática, por que:

- Usa o desvio-padrão da população (*sigma*), que geralmente não conhecemos;
- Pressupõe que a distribuição dos pesos dos peixes seja normal;
- Não permite estimarmos as probabilidades de erro $\alpha$ e $\beta$;
- Não convence, porque não tem nenhuma base matemática formal.

Na próxima seção, veremos como fazer o teste de maneira analítica, usando teoremas matemáticos sobre a *distribuição amostral* teórica da média.