# ATIVIDADE 1

DUPLA: Fernando Andrade Lima Tavares e Nina Magalhães de Oliveira

Dataset utilizado: Breast Cancer (Diagnostic) Dataset

DOI: **10.24432/C51P4M**

Criadores: Matjaz Zwitter, Milan Soklic

## DESCRIÇÃO DO DATASET

- **Qual foi o método utilizado para a coleta dos dados?**

  O dataset foi coletado pela Universidade Medical Centre, Instituto de Oncologia na Ljubljana, Eslovênia, a partir de um estudo médico.

- **A base já foi empregada em trabalhos acadêmicos ou pesquisas científicas?**

  Sim, a base é utilizada em diversos trabalhos acadêmicos envolvendo técnicas estáticas e de machine learning.
  Alguns exemplos:

  *QUOTIENT: Two-Party Secure Neural Network Training and Prediction*

  *Target-Focused Feature Selection Using a Bayesian Approach*

  *AnomiGAN: Generative adversarial networks for anonymizing private medical data*

- **A base possui alguma certificação oficial ou validação reconhecida?**

  No site UC Irvine Machine Learning Repository a licença reconhecida do dataset é a Creative Commons Attribution 4.0 International (CC BY 4.0)

- **Você considera a base de dados confiável?**

  Ela é uma base bastante consolidada na esfera acadêmica, já que sua origem é bem registrada é sua usabilidade é alta. Portanto, pode-se dizer que a base é confiável.

- **Variáveis do dataset**

Os atributos são computados a partir de uma imagem digitalizada de uma punção aspirativa por agulha fina (procedimento médico minimamente invasivo onde uma agulha fina é usada para coletar uma amostra de células ou fluidos de uma massa suspeita) de uma massa mamária. Eles descrevem as características da célula tumoral presente na imagem.
O dataset possui, também, o ID do sujeito de estudo e o diagnóstico, sendo que o último possui valor binário: M para maligno e B para benigno.
O diagnóstico é a *target,* ou seja, é a variável a qual se deseja entender melhor e até, possivelmente, fazer previsões a partir dos atributos.

Cada coluna segue os padrões descritos abaixo, totalizando 32 colunas, sendo elas:

| Nome variável | Tipo | Descrição |
|---|---|---|
| **diagnosis** | Categórica binária | Diagnóstico (benigno ou maligno) da célula tumoral observada |
| **id** | Numérica descritiva | ID do sujeito de estudo. |
| **radius-mean**<br>**radius-se**<br>**radius-worst** | Numérica contínua | Distância média do centro do núcleo até pontos no perímetro. Mede o tamanho geral do núcleo. Um raio maior pode indicar um núcleo maior ou mais irregular. |
| **texture-mean**<br>**texture-se**<br>**texture-worst** | Numérica contínua | Desvio padrão dos valores em escala de cinza da imagem do núcleo, indicando variação na intensidade dos pixels. Representa quão rugosa ou uniforme é a textura da superfície do núcleo. |
| **perimeter-mean**<br>**perimeter-se**<br>**perimeter-worst** | Numérica contínua | Comprimento total da borda do núcleo. Núcleos maiores ou mais irregulares tendem a ter perímetros maiores. |
| **area-mean**<br>**area-se** | Numérica contínua | Área total do núcleo, medida em número de pixels ou unidades equivalentes. Indica o tamanho do núcleo em termos de superfície. |

| | | |
|---|---|---|
| **area-worst** | | |
| **compactness-mean**<br>**compactness-se**<br>**compactness-worst** | Numérica contínua | Calculado como (perímetro$^2$ / área) - 1,0. Mede o quão "compacto" ou denso é o núcleo. Núcleos irregulares tendem a ter valores maiores de compacidade. |
| **smoothness-mean**<br>**smoothness-se**<br>**smoothness-worst** | Numérica contínua | Medida da variação local nos comprimentos dos raios (distâncias do centro ao perímetro). Indica a quão lisa (suave) ou irregular é a borda do núcleo. |
| **concavity-mean**<br>**concavity-se**<br>**concavity-worst** | Numérica contínua | Grau da severidade das porções côncavas (entradas ou reentrâncias) do contorno do núcleo. Valores altos indicam bordas mais "dentadas" e irregulares. |
| **concave-points-mean**<br>**concave-points-se**<br>**concave-points-worst** | Numérica contínua | Número de porções côncavas no contorno do núcleo, indicando irregularidades na forma. |
| **symmetry-mean**<br>**symmetry-se**<br>**symmetry-worst** | Numérica contínua | Grau de simetria do núcleo. Núcleos malignos tendem a ser menos simétricos, com formas mais irregulares. |
| **fractal-dimension-mean**<br>**fractal-dimension-se**<br>**fractal-dimension-worst** | Numérica contínua | Aproximação da "rugosidade" ou complexidade do contorno do núcleo, baseada na dimensão fractal (similar à medição da complexidade da linha costeira). Valores próximos a 1 indicam contornos simples; valores maiores indicam contornos mais complexos e irregulares. |

- **Observações:**

1) O dataset consta com 357 diagnósticos benignos e 212 Malignos

2) Um tumor benigno é um crescimento não canceroso com células semelhantes às de origem, crescimento lento, limites definidos, que geralmente pode ser removido com cirurgia e não se espalha para outras

partes do corpo. Já um tumor maligno é canceroso, com células que crescem descontroladamente, podem invadir tecidos vizinhos (comportamento agressivo) e se espalhar para outros órgãos (metástase), exigindo tratamentos mais complexos como quimioterapia e radioterapia.

3) Por não possuir um papel importante dentro das análises que serão feitas do dataset, a variável *id* foi retirada.

- **Atributos derivados para cada núcleo celular (colunas 3 a 32)**

Para cada uma das 10 características abaixo, são calculadas três medidas diferentes:

1) **-mean** (média): média dos valores observados para a característica no núcleo
2) **-se** (standard error, erro padrão): medida da variabilidade ou incerteza da característica
3) **-worst** (pior): média dos três maiores valores da característica, representando o pior caso observado

Assim, temos um total de 30 atributos derivados (10 características × 3 medidas).

## ANÁLISES ESTATÍSTICAS

- **Média, mediana, moda, desvio padrão e variância das variáveis**

As estatísticas das variáveis do dataset foram calculadas a partir das seguintes funções da linguagem R: *mean* para a média, *median* para a mediana, *var* para a variância e *sd* para o desvio padrão. Para obter a moda, criou-se uma função (*getmode*) a partir dos valores únicos e de maior frequência de cada variável.

| | MÉDIA | MEDIANA | MODA | DESVIO PADRÃO | VARIÂNCIA |
|---|---|---|---|---|---|
| **Radius-mean** | 14.1272917398946 | 13.37 | 12.34 | 12.4189201295267 | 3.52404882621208 |

| | | | | | |
|---|---|---|---|---|---|
| **Texture-mean** | 19.2896485061511 | 18.84 | 15.7 | 18.4989086790515 | 4.30103576816695 |
| **Perimeter-mean** | 91.9690333919156 | 86.24 | 82.61 | 590.44047952177 | 24.2989810387549 |
| **Area-mean** | 654.889103690685 | 551.1 | 512.2 | 123843.554317681 | 351.914129181653 |
| **Smoothness-mean** | 0.0963602811950791 | 0.09587 | 0.1007 | 0.000197799700272903 | 0.0140641281376736 |
| **Compactness-mean** | 0.1043409841827777 | 0.09263 | 0.1206 | 0.00278918740043813 | 0.0528127579325122 |
| **Compactness-mean** | 0.1043409841827777 | 0.09263 | 0.1206 | 0.00278918740043813 | 0.0528127579325122 |
| **Concavity-mean** | 0.0887993158172232 | 0.06154 | 0 | 0.00635524790042313 | 0.0797198087078935 |
| **Concave-points-mean** | 0.0489191458699473 | 0.0335 | 0 | 0.00150566076916354 | 0.0388028448591536 |
| **Symmetry-mean** | 0.1811618629173399 | 0.1792 | 0.1769 | 0.000751542821171316 | 0.0274142813360357 |
| **Fractal-dimension-mean** | 0.0627976098418278 | 0.06154 | 0.05667 | 4.98487227982128e-05 | 0.0070603627950844 6 |
| **Radius-se** | 0.405172056239016 | 0.3242 | 0.286 | 0.0769023518762222 | 0.277312732986104 |
| **Texture-se** | 1.21685342706503 | 1.108 | 1.15 | 0.304315949077143 | 0.551648392617202 |
| **Perimeter-se** | 2.86605922671353 | 2.287 | 1.778 | 4.08789583770081 | 2.02185455404211 |
| **Area-se** | 40.337079086116 | 24.53 | 16.97 | 2069.43158286873 | 45.4910055161318 |
| **Smoothness-mean** | 0.0070409789103690 7 | 0.00638 | 0.006399 | 9.01511400307557e-06 | 0.0030025179438390 7 |
| **Compactness-se** | 0.0254781388400703 | 0.02045 | 0.0231 | 0.000320702886760619 | 0.0179081793256774 |
| **Concavity-se** | 0.0318937163444464 | 0.02589 | 0 | 0.000911198237823095 | 0.0301860603229884 |
| **Concave-points-se** | 0.0117961370826011 | 0.01093 | 0 | 3.80724191290626e-05 | 0.0061702851740468 7 |
| **Symmetry-se** | 0.0205422987697715 | 0.01873 | 0.01344 | 6.83328982521288e-05 | 0.0082663715287984 |
| **Fractal-dimension-se** | 0.0037949038 6643234 | 0.003187 | 0.003002 | 7.00169156287235e-06 | 0.0026460709670892 |
| **Radius-worst** | 16.26918980667 84 | 14.97 | 12.36 | 23.3602241751776 | 4.83324158046932 |
| **Texture-worst** | 25.677223198594 | 25.41 | 27.26 | 37.7764827687567 | 6.1462576230 3832 |
| **Perimeter-worst** | 107.2612126537 79 | 97.66 | 117.7 | 1129.13084694237 | 33.6025422690364 |
| **Area-worst** | 880.5831282952 55 | 686.5 | 1269 | 324167.385102168 | 569.356992669949 |
| **Smoothness-worst** | 0.1323685940246 05 | 0.1313 | 0.1312 | 0.000521319832526795 | 0.0228324294048355 |
| **Compactness-worst** | 0.2542650439 36731 | 0.2119 | 0.3416 | 0.0247547707437041 | 0.1573364889 13742 |

| | | | | | |
|---|---|---|---|---|---|
| **Concavity-worst** | 0.272188483304042 | 0.2267 | 0 | 0.0435240904592607 | 0.208624280608132 |
| **Concave-points-worst** | 0.114606223198594 | 0.09993 | 0 | 0.00432074067909974 | 0.00432074067909974 |
| **Symmetry-worst** | 0.00432074067909974 | 0.2822 | 0.3196 | 0.00382758353950593 | 0.0618674675375187 |
| **Fractal-dimension-worst** | 0.0839458172231986 | 0.08004 | 0.07427 | 0.000326209378248224 | 0.018061267348894 |

Como é possível observar na tabela, apenas as variáveis *area, perimeter, texture e radius* possuem valores significativos, já que são as únicas medidas maiores que zero do dataset.

- **Tabela de frequência**

Foram criadas tabelas de frequência para a única variável qualitativa do dataset (*diagnosis*) e para quatro variáveis quantitativas *(radius-mean, area-mean, perimeter-mean, texture-mean*). Elaborou-se uma função que define as frequências absolutas, relativas e percentuais e, desse modo, gera as tabelas de frequência.

A função primeiro verifica se o dado atribuído é numérico. Se for numérico, ela irá dividi-lo em uma sequência de 5 intervalos, sendo que cada um começa do valor mínimo (arredondado para baixo) e vai até o valor máximo (arredondado para cima). Os intervalos incluem apenas o limite superior. Finalmente, a função irá contar quantas vezes cada intervalo aparece no dado atribuído e armazenar esse valor na variável n_1, que é a frequência absoluta.

Se o dado não for numérico, como é o caso para *diagnosis,* a função irá contar a frequência absoluta de cada categoria. Para a variável diagnosis, serão contadas quantas vezes as categorias M (maligno) e B (benigno) aparecem no dataset.

Para a frequência relativa, calcula-se a proporção de cada intervalo em relação ao total, ou seja, a frequência relativa, armazenada na variável f_1, é obtida com base em n_1.

Já a frequência percentual, transforma-se as frequências relativas em porcentagem, portanto, multiplica-se o valor de f_1 por cem.

Definidos os valores das frequências absolutas (n-1), relativas (f_1) e percentuais (p_1), estes são colocados em vetores em conjunto com a soma deles, ou seja, o total de cada frequência. Os totais devem ser os mesmos para cada variável, então o total de n_1 deve ser 569 (número de linhas do dataset), de f_1 é 1 e de p_1 é 100.

Ao fim, os vetores são unificados em uma tabela que, após algumas modificações de normalização, são transformadas nas tabelas de frequência. Elas possuem os seguintes *layouts*:

*Layout* da tabela de frequência para as variáveis quantitativas:

| Nome da variável | Frequência absoluta | Frequência relativa | Frequência percentual |
|---|---|---|---|
| Intervalo 1 | | | |
| Intervalo 2 | | | |
| Intervalo 3 | | | |
| Intervalo 4 | | | |
| Intervalo 5 | | | |
| **Total** | **569** | **1** | **100** |

*Layout* da tabela de frequência para a variável qualitativa:

| Nome da variável | Frequência absoluta | Frequência relativa | Frequência percentual |
|---|---|---|---|
| categoria 1 | | | |
| categoria 2 | | | |
| ... | | | |
| categoria n | | | |
| **Total** | **569** | **1** | **100** |

Tabelas de frequência geradas no R:

```
Tabela de frequência para: diagnosis
  diagnosis n_i   f_i    p_i
1         B 357 0.627 62.742
2         M 212 0.373 37.258
3     Total 569     1    100
```

```
Tabela de frequência para: radius-mean
  radius-mean n_i   f_i    p_i
1    [6,11.8) 148  0.26 26.011
2 [11.8,17.5) 318 0.559 55.888
3 [17.5,23.2)  93 0.163 16.344
4   [23.2,29)  10 0.018  1.757
5       Total 569     1    100
```

```
Tabela de frequência para: texture-mean
  texture-mean n_i   f_i    p_i
1     [9,16.8) 167 0.293  29.35
2  [16.8,24.5) 334 0.587 58.699
3  [24.5,32.2)  64 0.112 11.248
4    [32.2,40)   4 0.007  0.703
5        Total 569     1    100
```

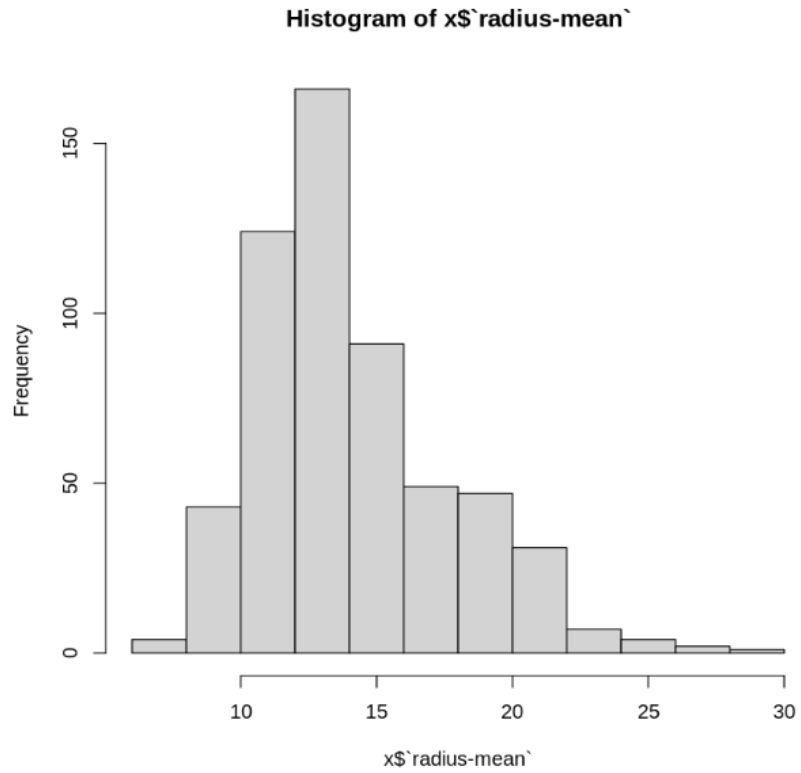
```
 Tabela de frequência para: perimeter-mean
   perimeter-mean n_i   f_i    p_i
 1      [43,79.5) 205  0.36 36.028
 2     [79.5,116) 264 0.464 46.397
 3      [116,152)  89 0.156 15.641
 4      [152,189)  11 0.019  1.933
 5          Total 569     1    100
```

```
Tabela de frequência para: area-mean
            area-mean n_i   f_i    p_i
1           [143,732) 413 0.727 72.711
2      [732,1.32e+03) 130 0.229 22.887
3 [1.32e+03,1.91e+03)  22 0.039  3.873
4  [1.91e+03,2.5e+03)   3 0.005  0.528
5               Total 568     1    100
```

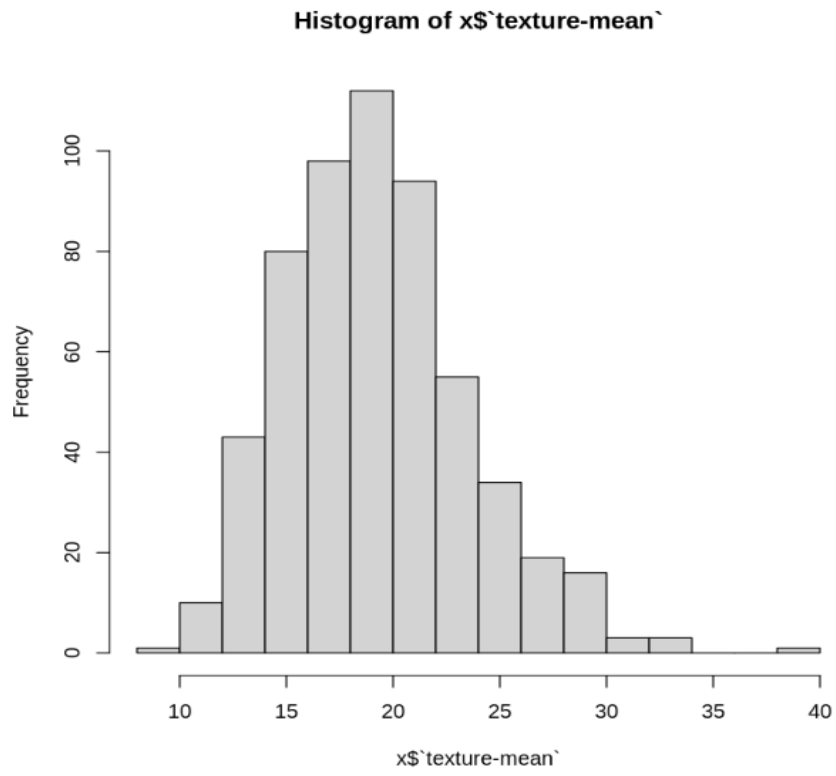
- **Histograma**

A fim de se obter uma representação gráfica da distribuição de frequências das variáveis contínuas *radius-mean* e *texture-mean* , criaram-se histogramas delas. Na linguagem R, utiliza-se a função *hist( )* para plotar os histogramas nos quais o eixo x representa os intervalos das variáveis contínuas e o eixo y representa a frequência de cada intervalo.

Histograma da variável *radius-mean*



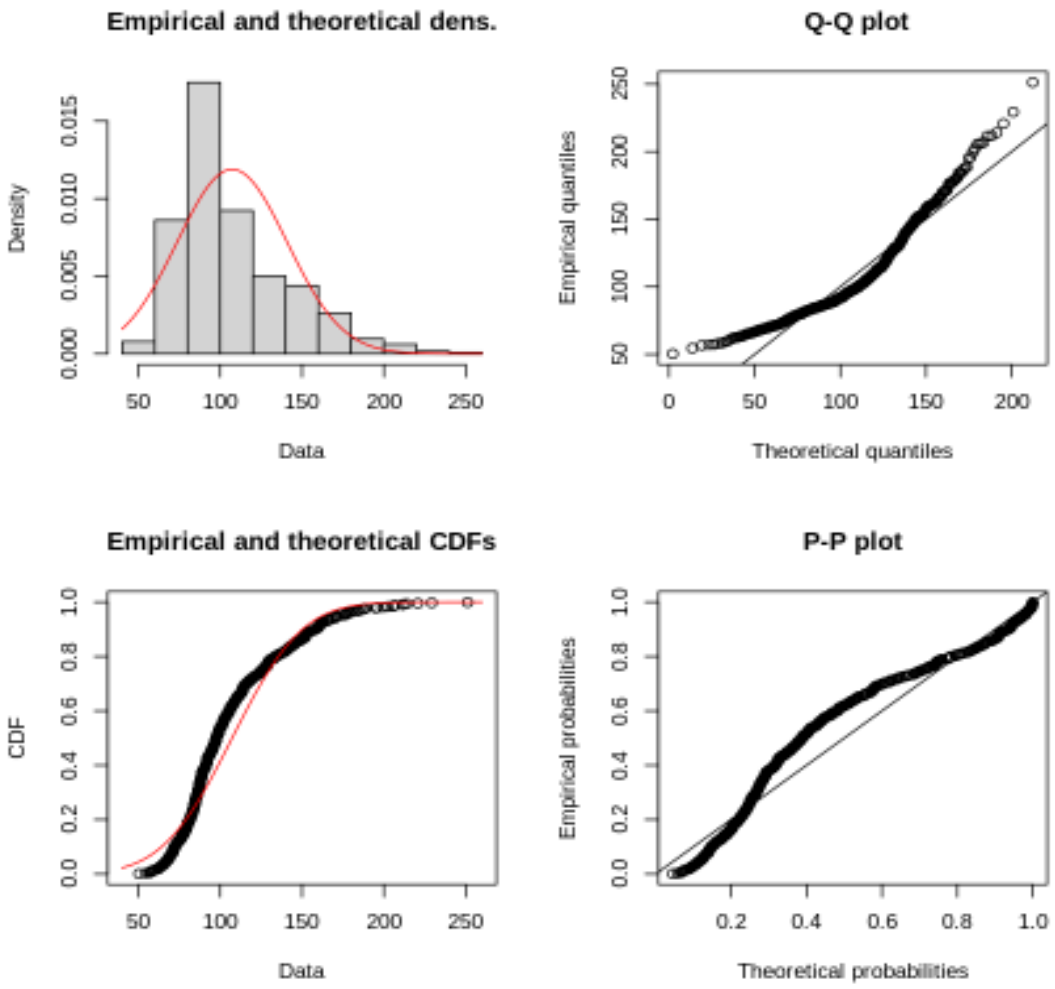

Histograma da variável *texture-mean*



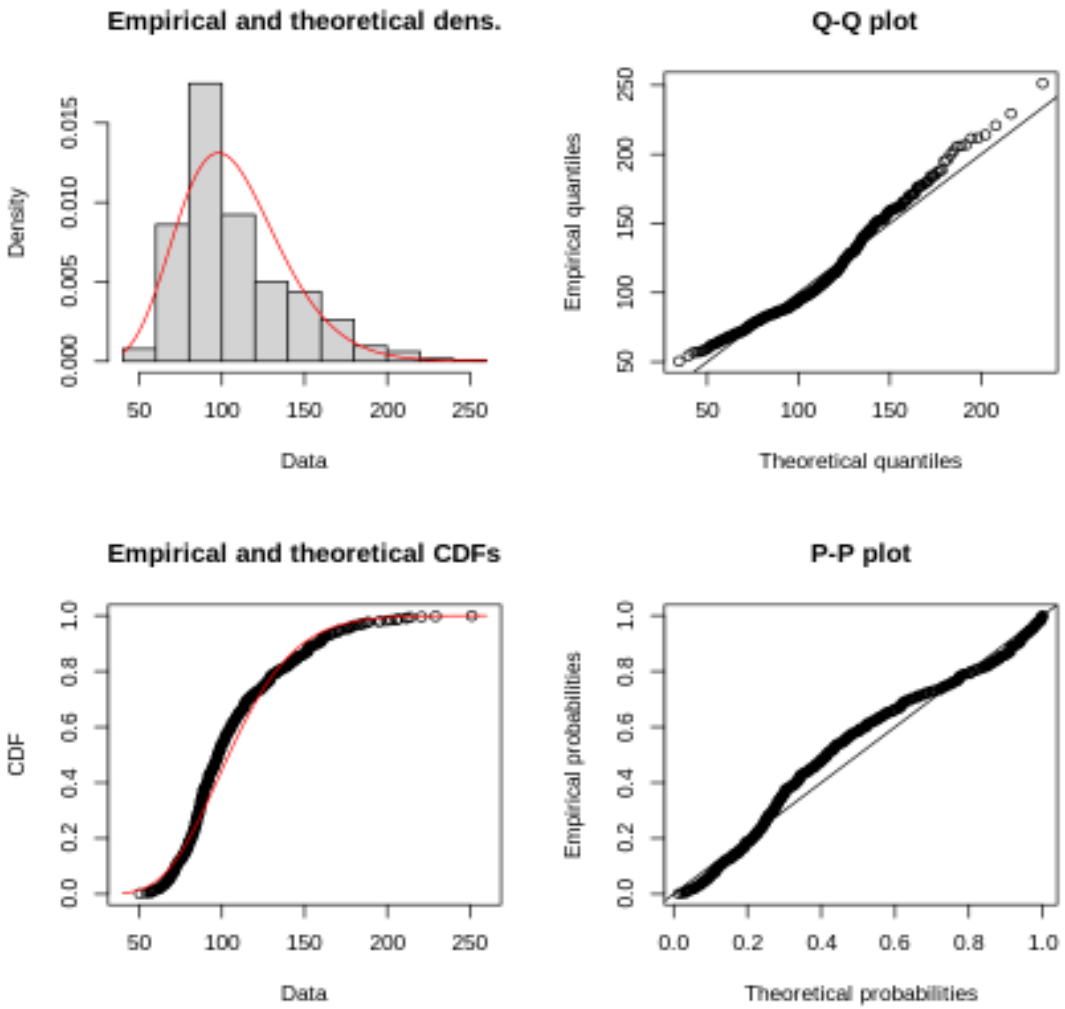
Observa-se que a os valores da variável *radius-mean* estão concentrados no intervalo entre 10 e 15, já para a variável *texture-mean*, a maioria dos dados encontra-se entre o intervalo 15 e 20. Concluir que existe uma relação entre esses valores e o diagnóstico não é viável, tanto pela necessidade de uma análise mais completa de todos os dados quanto pela lacuna de conhecimento médico. Entretanto, é possível afirmar que a criação de histogramas pode indicar certos padrões e/ou tendências que auxiliam na elaboração de importantes *insights*.

Fez-se, também, o ajuste de distribuição das probabilidades para os dados da variável *perimeter-worst* utilizando as distribuições normal e gama:

Distribuição normal da variável *perimeter-worst*



Distribuição gama da variável *perimeter-worst*

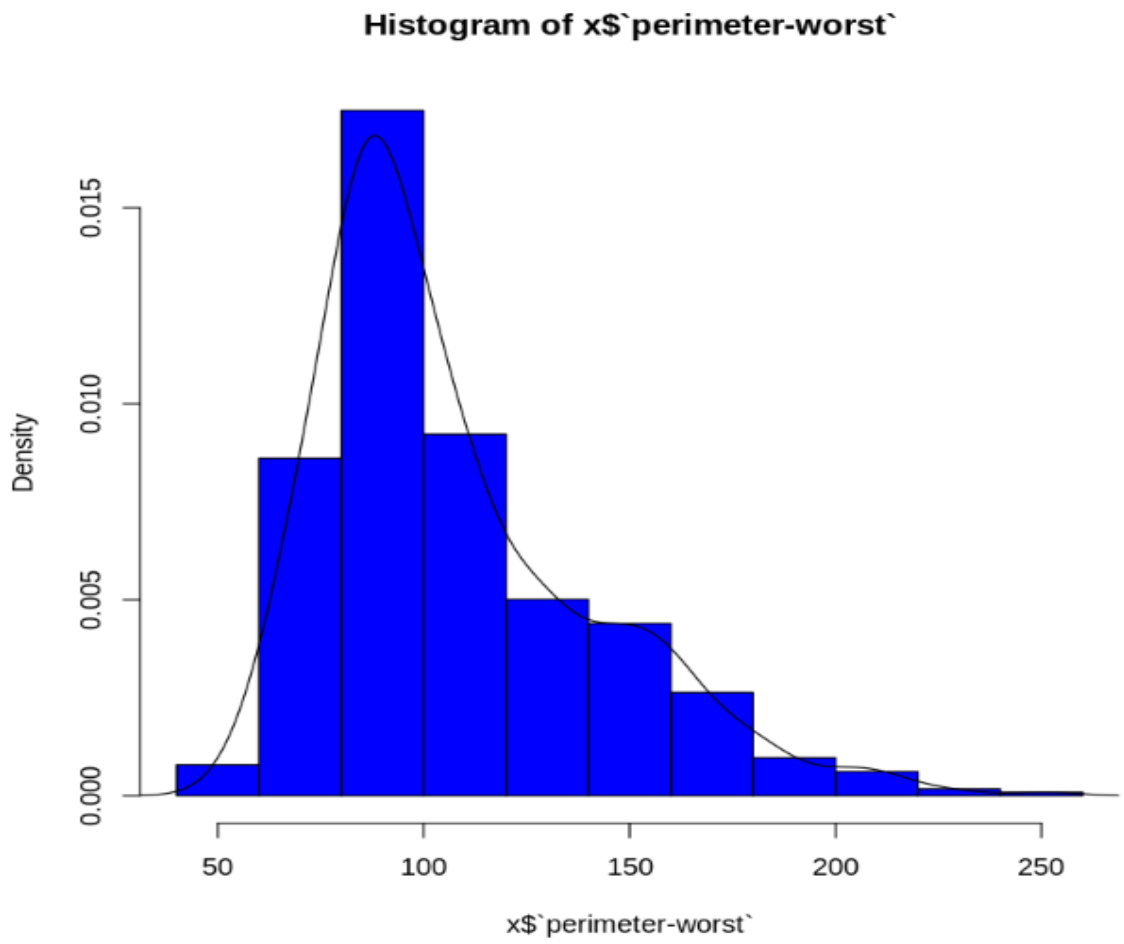



Por fim, criou-se um histograma da variável *perimeter-worst* com a curva de densidade, a qual foi inserida no gráfico utilizando a função *density( )*. Durante a análise do dataset, conclui-se que a curva de densidade é a que melhor se adapta aos dados.

Vale ressaltar, também, que este histograma aceita o parâmetro

probabilidade como sendo verdadeira, ou seja, em vez das contagens absolutas, ele considera as probabilidades para a plotagem.

Histograma da variável *perimeter-worst* com a curva de densidade:

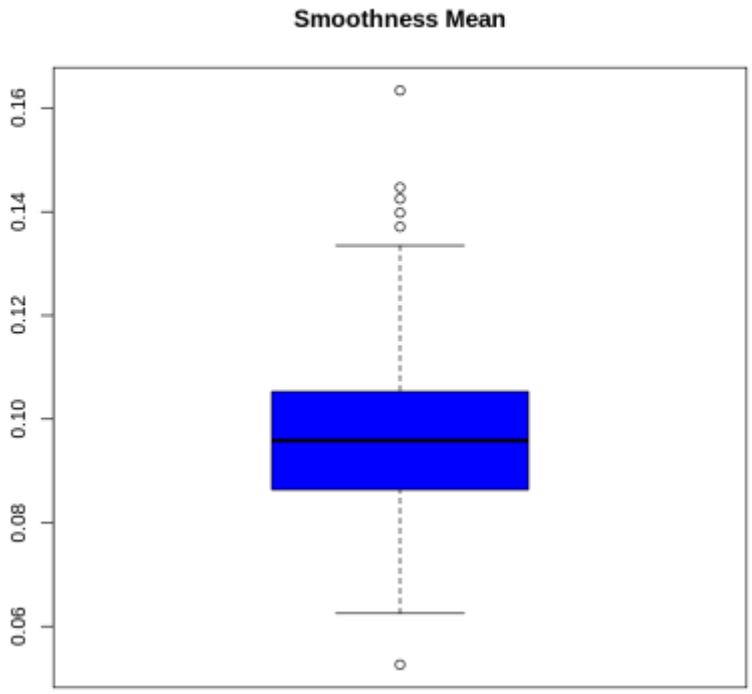



- **Boxplot**

Boxplots são formas convenientes de exibir graficamente uma variável, já que conseguem transmitir informações sobre a dispersão e simetria dos dados e evidenciar possíveis outliers. Além disso, caso necessário, facilitam a comparação entre os conjuntos de dados.

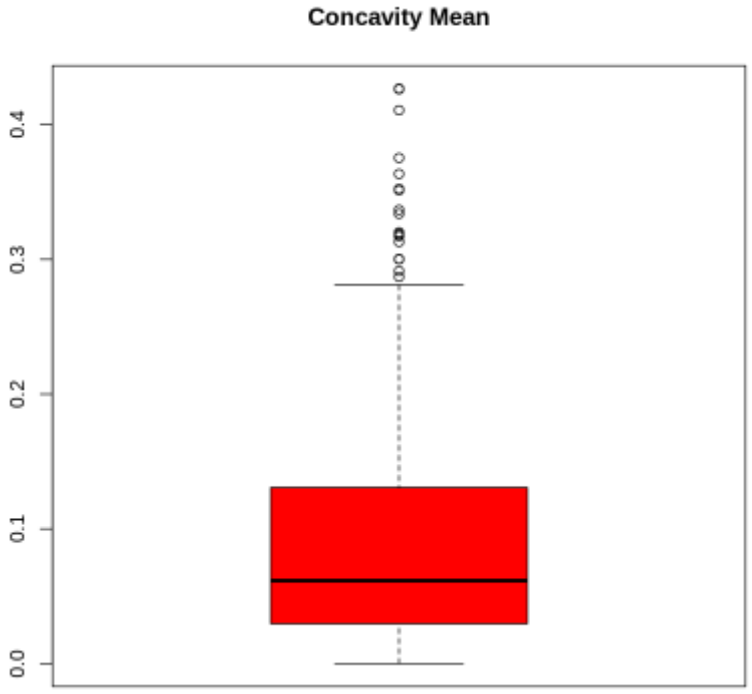
Foram gerados boxplots para as seguintes variáveis: *smoothness-mean, concavity-mean* e *compactness-mean:*

Boxplot da variável *smoothness-mean*



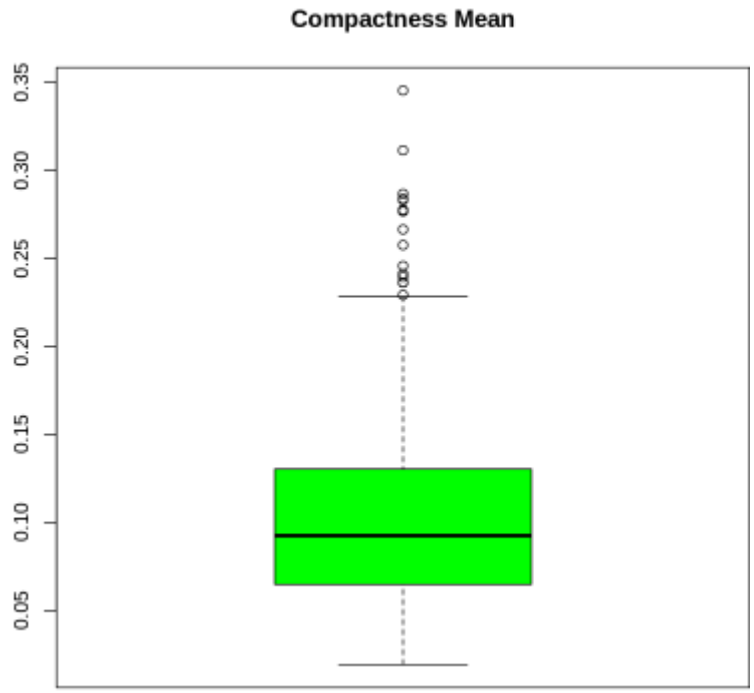

Com mediana em torno de 0.10, Ele apresenta menor quantidade de outliers e menos dispersão. Como *smoothness-mean* mede pequenas variações no raio da superfície do tumor, tende a ser mais estável, tanto em tumores benignos quanto malignos.

Boxplot da variável *concavity-mean*



Com a mediana em torno de 0.06, possui muitos outliers acima de 0.25, indicando que alguns tumores têm *concavity-mean* bem mais alta que a maioria, isso implica em como concavidade está ligada a irregularidades no contorno do tumor, valores altos podem estar associados a tumores malignos.

Boxplot da variável *compactness-mean*




Mediana próxima de 0.09, existem diversos outliers acima de 0.23, a compactação mede a solidez do tumor; valores altos sugerem formas mais anômalas, ligadas a malignidade.

- **Tabela cruzada**

Seguindo as orientações do Professor, o P-valor foi desconsiderado e apenas o X-squared = 223.77 foi utilizado e, com isso, implica-se que as variáveis *concavity-mean* e *diagnosis* são correlacionadas.

Tabela Cruzada:

| | **Benigno** | **Maligno** |
|---|---|---|
| **Concavidade Pequena** | 349 | 93 |
| **Concavidade Média** | 5 | 104 |
| **Concavidade Grande** | 3 | 15 |

- **Teste de Correlação e gráfico de dispersão**

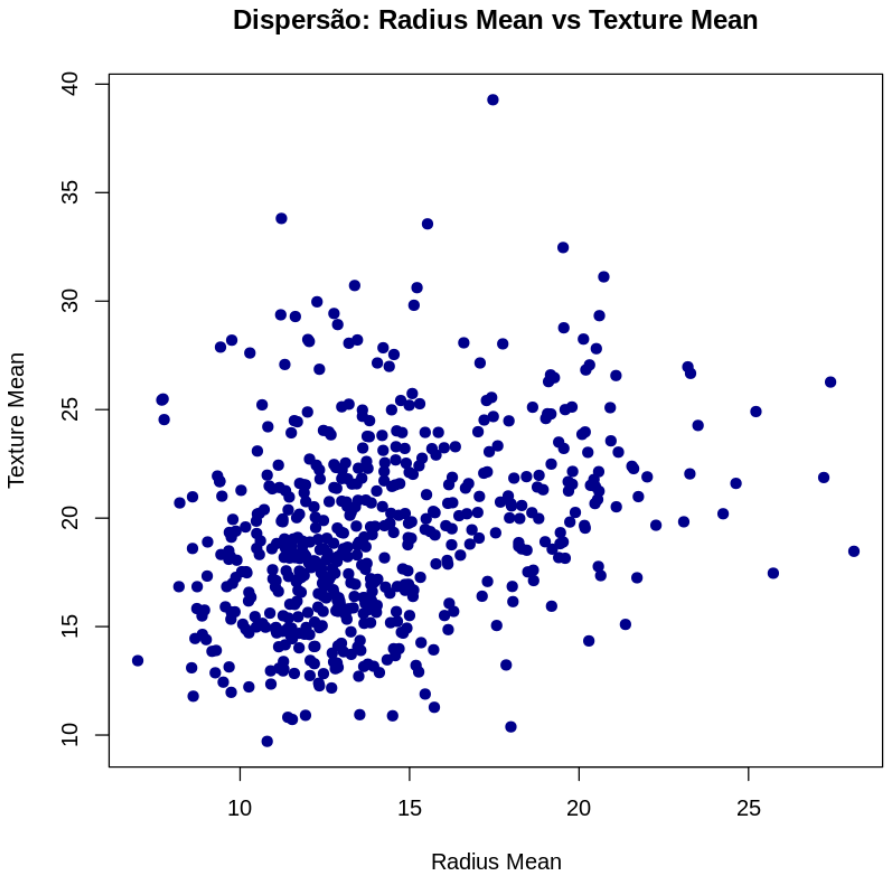
O dataset não possui variáveis quantitativas discretas, portanto, o teste de correlação foi realizado para dados quantitativos contínuos, sendo utilizadas as colunas t*exture-mean* e *radius-mean*.

```
        Pearson's product-moment correlation

data:  x$`radius-mean` and x$`texture-mean`
t = 8.1488, df = 567, p-value = 2.36e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2481897 0.3954548
sample estimates:
      cor
0.3237819
```

Os resultados indicam uma correlação positiva moderada entre as variáveis *radius-mean* e *texture-mean* (r ≈ 0,32), sugerindo que valores maiores de raio médio tendem a estar associados a valores maiores de textura média. O teste de significância apresentou p-valor extremamente baixo (p < 0,001), o que confirma que essa correlação dificilmente ocorreu por acaso. O intervalo de confiança de 95% (0,25 a 0,39) está inteiramente acima de zero, reforçando a existência de uma relação positiva consistente entre as variáveis. Na prática, isso significa que há associação entre o tamanho médio das células e sua textura, embora a relação não seja forte, indicando que outros fatores também influenciam a textura.

Gráfico de Dispersão:




O gráfico de dispersão mostra que existe uma tendência positiva entre as variáveis *radius-mean* e *texture-mean*, ou seja, à medida que o raio médio aumenta, a textura média também tende a aumentar. Apesar de os pontos estarem relativamente espalhados, é possível perceber esse padrão de crescimento, o que confirma a correlação positiva moderada observada anteriormente ($r \approx 0{,}32$). A maior concentração de dados ocorre na faixa de raio médio entre 10 e 15 e textura média entre 15 e 25, indicando que a maioria das observações se encontra nesse intervalo.

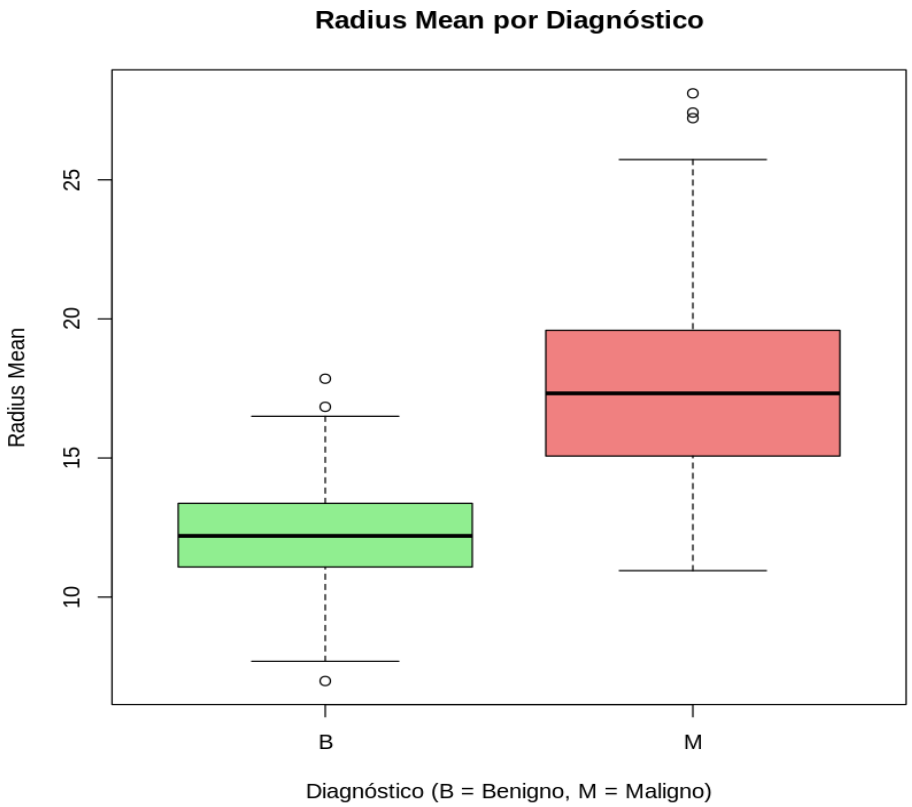
- **Tapply e Boxplot para quantitativa por qualitativa**

```
$B
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  6.981  11.080  12.200  12.147  13.370  17.850 

$M
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  10.95   15.07   17.32   17.46   19.59   28.11 

        B         M 
 3.170222 10.265431 
  B   M 
357 212 
```

A análise descritiva do *radius-mean* mostra que, em média, os tumores malignos apresentam raio médio consideravelmente maior (por volta de 17) do que os tumores benignos (por volta de 12). Além disso, a variância dos casos malignos é mais alta, indicando maior dispersão dos valores. A amostra é composta por 357 casos benignos e 212 malignos, e os valores mínimos e máximos confirmam que tumores malignos tendem a ter raios maiores em relação aos benignos.




O boxplot evidencia que o radius-mean tende a ser significativamente maior em tumores malignos do que em tumores benignos, reforçando que essa variável é potencialmente útil para distinguir entre os dois diagnósticos. Apesar de alguma sobreposição entre os grupos, a separação entre as medianas e a maior amplitude dos valores malignos indicam que o tamanho médio das células é uma característica relevante para a classificação.

- **Referências**


Zwitter, M.; Soklic, **M. Breast Cancer (Diagnostic) Dataset.** UCI Machine Learning Repository, 1988. DOI. Disponível em: https://doi.org/10.24432/C51P4M.

**BREAST-CANCER-EDA-AND-MODELING-R [repositório].** GitHub. Disponível em: https://github.com/fertavares/Breast-Cancer-EDA-and-Modeling-R.

R CORE TEAM. The R Stats Package. **Vienna: R Foundation for Statistical Computing, 2023.** Disponível em: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html.

WICKHAM, H.; HELLMUND, B. **readr: Read Rectangular Text Data.** R package version 2.1.4. RStudio, 2023. Disponível em: https://readr.tidyverse.org/.DELIGNETTE-MULLER, M. L.

DUTANG, C. **fitdistrplus: Help to Fit of a Parametric Distribution to Non-Censored or Censored Data.** R package version 1.1-11. Vienna: R Foundation for Statistical Computing, 2015. Disponível em: https://cran.r-project.org/package=fitdistrplus.