Aula #06 - Associação, Correlação e Regressão: examinando relações entre variáveis

Métodos Quantitativos 101
Prof. ***Ricardo Ceneviva*** (PhD)
ricardo.ceneviva@ufabc.edu.br

UFABC
14 de outubno de 2025

Objetivos da aula

- Entender associação em dados categóricos (proporções condicionais; contraste).
- Descrever relações em dados quantitativos (dispersão; direção; força).
- Interpretar correlação de Pearson e reconhecer limitações.
- Ajustar e interpretar regressão linear simples; analisar resíduos.
- Evitar extrapolação; identificar outliers e pontos influentes.
- Reconhecer variáveis de confusão
- **Distinguir associação de causalidade.**

Table 3.1 Frequencies for Food Type and Pesticide Status

The row totals and the column totals are the frequencies for the categories of each variable. The counts inside the table give information about the association.

	Pesticide Residues		
Food Type	**Present**	**Not Present**	**Total**
Organic	29	98	**127**
Conventional	19,485	7,086	**26,571**
Total	**19,514**	**7,184**	**26,698**

Questions to Explore

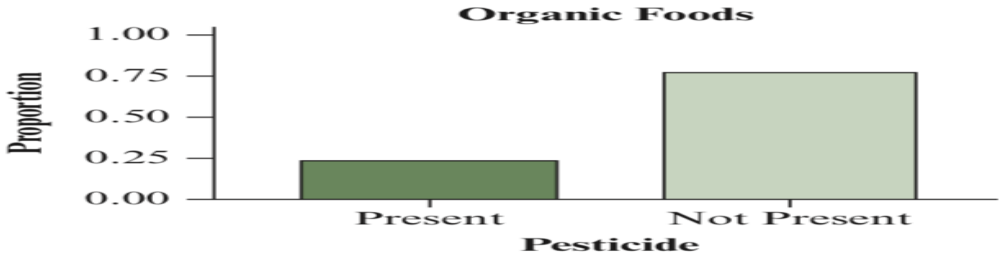
a. What is the response variable, and what is the explanatory variable?

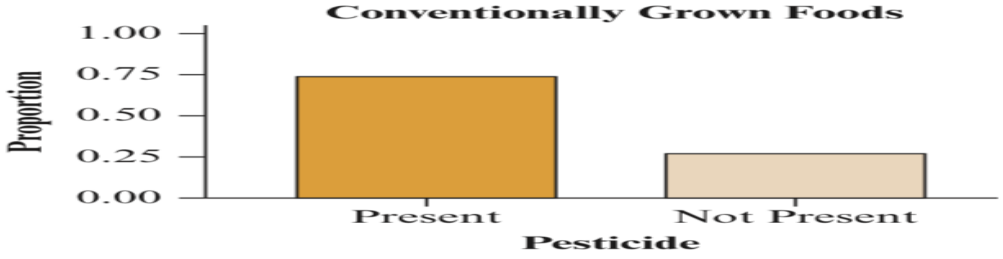
b. Only 127 organic food types were sampled compared to 26,571 conventional ones. Despite this huge imbalance, can we still make a fair comparison?

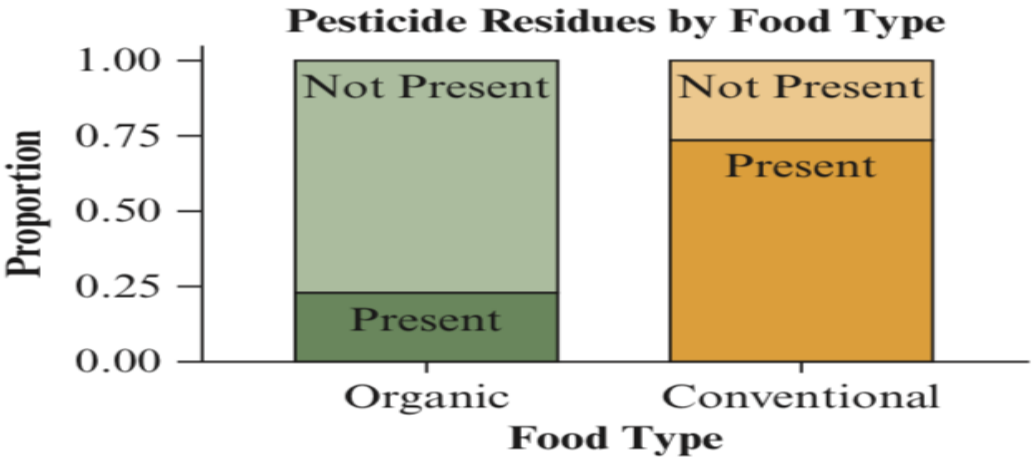
Table 3.2 Conditional Proportions on Pesticide Status for Two Food Types

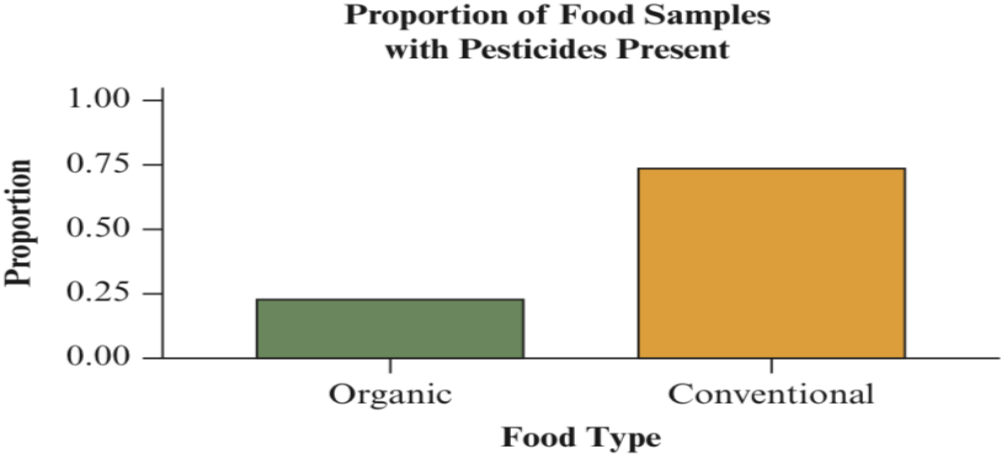
These conditional proportions (using two decimal places) treat pesticide status as the response variable. The sample size n in the last column shows the total on which the conditional proportions are based.

	Pesticide Residues			
Food Type	**Present**	**Not Present**	**Total**	***n***
Organic	0.23	0.77	**1.00**	127
Conventional	0.73	0.27	**1.00**	26,571

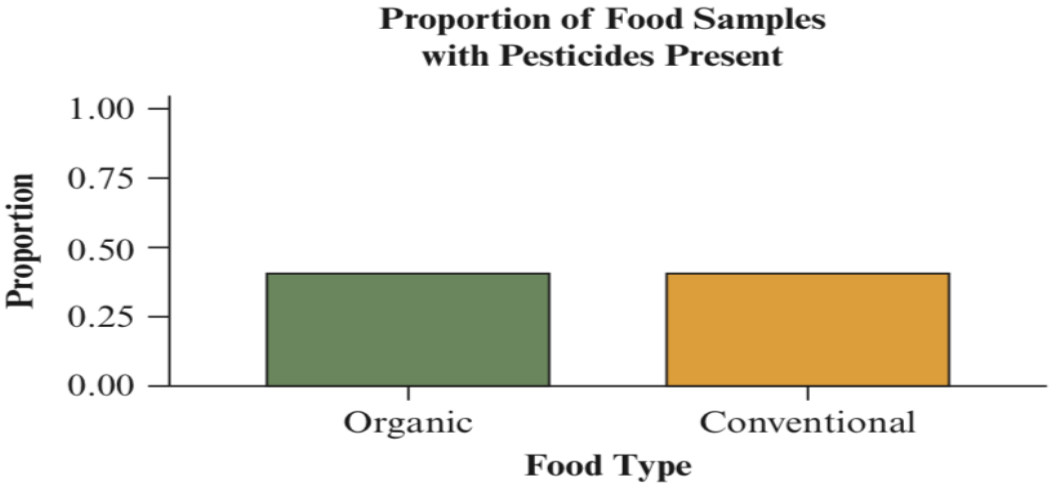

Organic Foods
Proportion
1.00
0.75
0.50
0.25
0.00
Present
Not Present
Pesticide


Conventionally Grown Foods
Proportion
1.00
0.75
0.50
0.25
0.00
Present
Not Present
Pesticide


Pesticide Residues by Food Type
Proportion
1.00
0.75
0.50
0.25
0.00
Not Present
Present
Not Present
Present
Organic
Conventional
Food Type


Proportion of Food Samples
with Pesticides Present
Proportion
1.00
0.75
0.50
0.25
0.00
Organic
Conventional
Food Type

Situação Hipotética: nesse caso, há associação entre as variáveis?



Comparando duas proporções condicionais

Tabelas 2×2: marginais e condicionais — notação

- Considere contagens n_{ij} na célula (linha i, coluna j), totais de linha n_{i+}, totais de coluna n_{+j} e total n.
- **Proporções marginais** (usam as *margens*):

$$p_{i+} = \frac{n_{i+}}{n} \quad \text{(marginal de linhas)}, \qquad p_{+j} = \frac{n_{+j}}{n} \quad \text{(marginal de colunas)}.$$

- **Proporções condicionais por linha** (distribuição de Y *dado* $X = i$):

$$p_{j|i} = \frac{n_{ij}}{n_{i+}}.$$

- **Proporções condicionais por coluna** (distribuição de X *dado* $Y = j$):

$$p_{i|j} = \frac{n_{ij}}{n_{+j}}.$$

Por que "linhas = condicionais" e "colunas = marginais"?

- **Convenção**: variável explicativa nas linhas (X) e resposta nas colunas (Y).
- Para estudar efeito de X sobre Y: comparar $Y \mid X = i \Rightarrow$ **percentuais por linha** (condicionais).
- Para descrever Y no conjunto: usar **marginal de** $Y \Rightarrow p_{+j}$ (totais de coluna sobre n).
- **Não é regra estrutural**: também há condicionais por coluna ($p_{i|j}$) e marginais de linhas (p_{i+}).
- **Boa prática**: sempre declarar qual variável é a condicionante e qual é a resposta.

Exemplo 2×2 (fumante $\times$ óbito): marginais vs. condicionais

	Óbito	**Vivo**	**Total (linha)**
Fuma	40	60	100
Não fuma	20	80	100
Total (coluna)	60	140	200

- **Condicionais por linha** (distribuição de óbito dado hábito):

$$P(\text{óbito} \mid \text{fuma}) = \frac{40}{100} = 0{,}40, \quad P(\text{óbito} \mid \text{não fuma}) = \frac{20}{100} = 0{,}20.$$

- **Marginal de colunas** (distribuição de óbito no total):

$$P(\text{óbito}) = \frac{60}{200} = 0{,}30, \quad P(\text{vivo}) = \frac{140}{200} = 0{,}70.$$

- **Mensagem**: marginais usam *margens* (totais); condicionais comparam *dentro* de cada linha/coluna (dado a outra variável).

Treating COVID-19 Infections

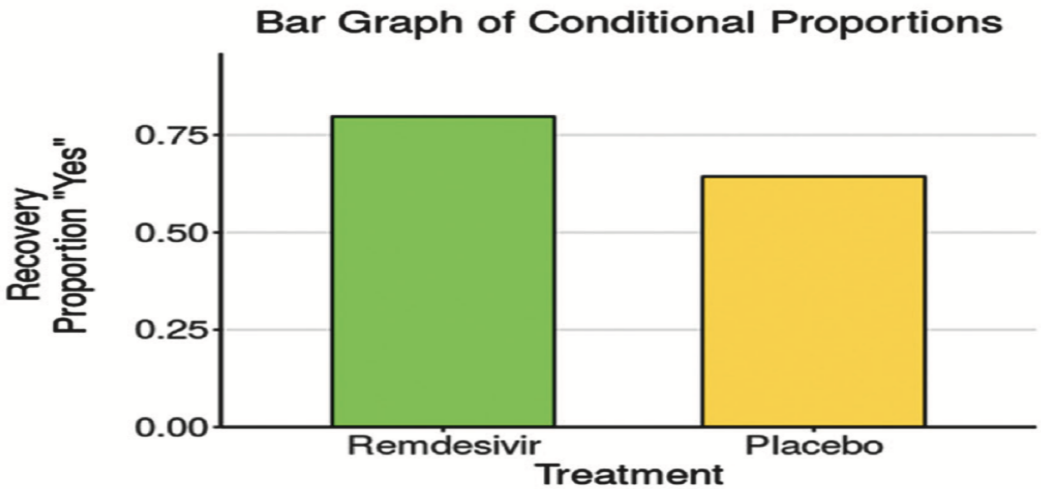
	Recovery		
Treatment	**Yes**	**No**	**Total**
Remdesivir	177	45	**222**
Placebo	128	71	**199**

Contingency Table (Observed Counts):

Treatment	Recovery: Yes	No	Total
Remdesivir	177	45	222
Placebo	128	71	199
Total	305	116	421

Conditional Sample Proportions:

	Recovery		
Treatment	Yes	No	Total
Remdesivir	0.7973	0.2027	1
Placebo	0.6432	0.3568	1
Total	0.7245	0.2755	1


Bar Graph of Conditional Proportions
Recovery
Proportion "Yes"
0.75
0.50
0.25
0.00
Remdesivir
Placebo
Treatment

Descriptive Statistics:

Proportion "Yes" for "Remdesivir"	Proportion "Yes" for "Placebo"	Ratio of Proportions
0.797	0.643	1.24

Categóricas: tabela e barras (base R × tidyverse)

base R

```
d <- read.csv("assets/data/pesticides.csv")
tab  <- with(d, table(food_type,
↪  pesticide_status))
prop <- prop.table(tab, margin = 1)  #
↪  P(pesticida | tipo)
barplot(t(prop), beside = TRUE, legend =
↪  TRUE,
        xlab = "Tipo de alimento", ylab =
        ↪  "Proporção condicional")
```

tidyverse

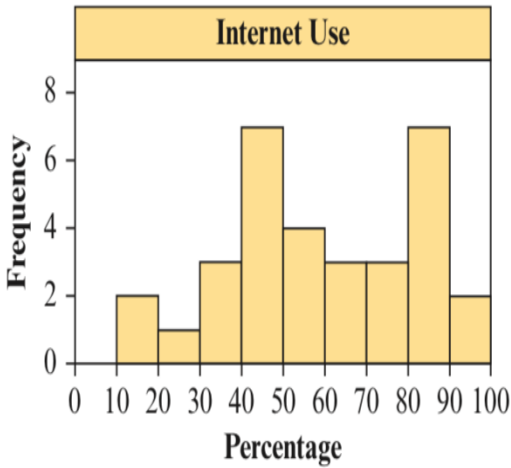
```
library(tidyverse)
d <- read_csv("assets/data/pesticides.csv")
prop <- d |>
  count(food_type, pesticide_status) |>
  group_by(food_type) |>
  mutate(prop = n/sum(n))
ggplot(prop, aes(food_type, prop, fill =
↪  pesticide_status)) +
  geom_col(position = "dodge") +
  labs(x = "Tipo de alimento", y = "Proporção
  ↪  condicional")
```

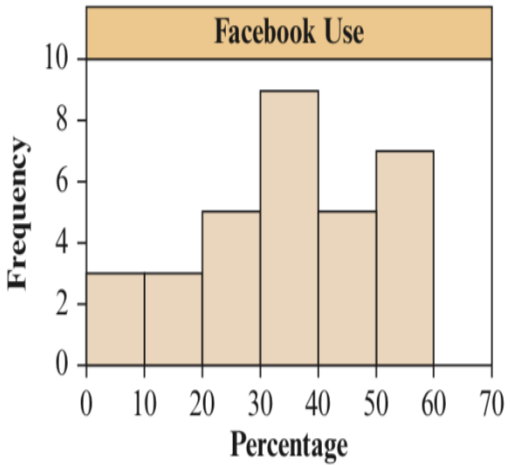
Associação entre duas variáveis quantitativas

Table 3.4 Internet and Facebook Penetration Rates for 32 Countries

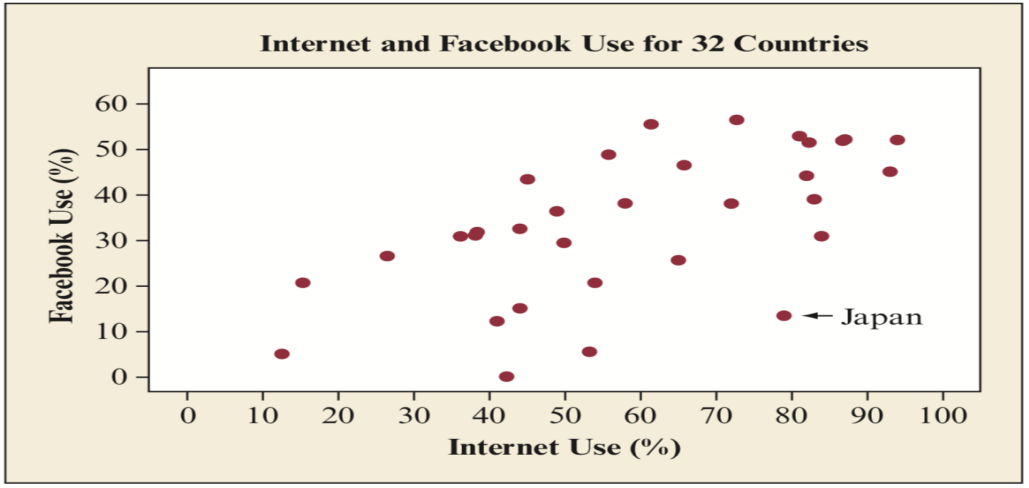
Country	Internet Penetration	Facebook Penetration
Argentina	55.8%	48.8%
Australia	82.4%	51.5%
Belgium	82.0%	44.2%
Brazil	49.9%	29.5%
Canada	86.8%	51.9%
Chile	61.4%	55.5%
China	42.3%	0.1%
Colombia	49.0%	36.3%
Egypt	44.1%	15.1%
France	83.0%	39.0%
Germany	84.0%	30.9%
Hong Kong	72.8%	56.4%
India	12.6%	5.1%

(*Continued*)


Internet Use
Frequency
0
2
4
6
8
0 10 20 30 40 50 60 70 80 90 100
Percentage


Facebook Use
Frequency
0
2
4
6
8
10
0 10 20 30 40 50 60 70
Percentage

Internet x Facebook



Como é calculado r — covariância padronizada

- Ideia: medir como x e y variam *juntas*, padronizando pelas dispersões.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Passo a passo:
  1. Calcule as médias $\bar{x}$ e $\bar{y}$.
  2. Obtenha os desvios: $(x_i - \bar{x})$ e $(y_i - \bar{y})$.
  3. Multiplique os desvios par a par e some: $\sum(x_i - \bar{x})(y_i - \bar{y})$.
  4. Some os quadrados dos desvios de x e de y e extraia as raízes.
  5. Divida o numerador pelo produto dos termos do passo 4.
- Forma compacta: $r = \dfrac{\mathrm{Cov}(X, Y)}{s_X\, s_Y}$ (os fatores $1/(n-1)$ se cancelam).

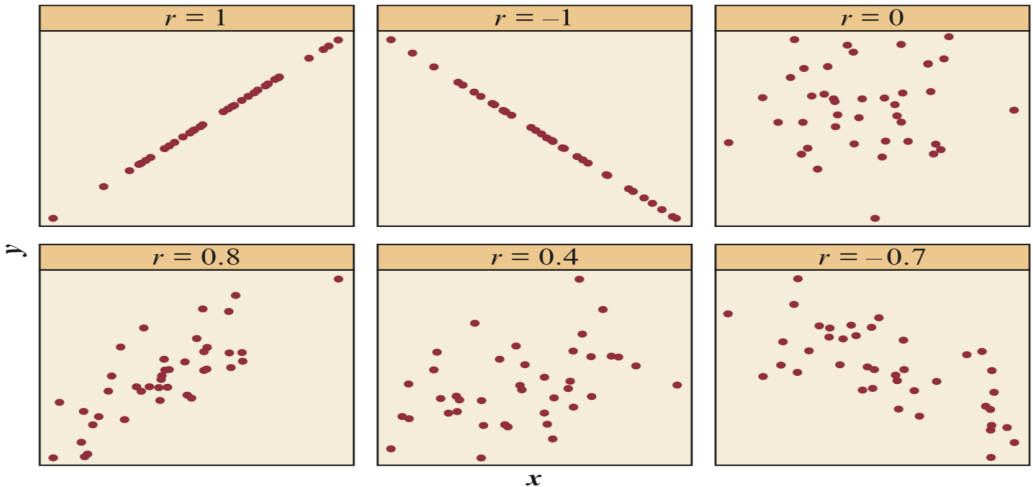
Como é calculado r — escores-z e interpretação

- Transforme para escores-z: $z_{x,i} = \dfrac{x_i - \bar{x}}{s_X}, \quad z_{y,i} = \dfrac{y_i - \bar{y}}{s_Y}.$

$$r = \frac{1}{n-1} \sum_{i=1}^{n} z_{x,i}\, z_{y,i}$$

- Leitura: r é a **média** dos produtos $z_x \times z_y$.
- Se x e y tendem a ficar acima da média ao mesmo tempo $\Rightarrow$ produtos positivos $\Rightarrow$ $r > 0$.
- Se um sobe quando o outro desce $\Rightarrow$ produtos negativos $\Rightarrow$ $r < 0$.
- Relação curvilínea: produtos positivos e negativos se anulam $\Rightarrow$ $r \approx 0$ mesmo havendo padrão.
- Resumo operacional:
  1. Centralize/padronize x e y.
  2. Some os produtos dos desvios (ou escores-z).
  3. Normalize por s_X e s_Y (ou use a fórmula acima).
  4. Interprete $r \in [-1, 1]$: sinal = direção; módulo = força *linear*.

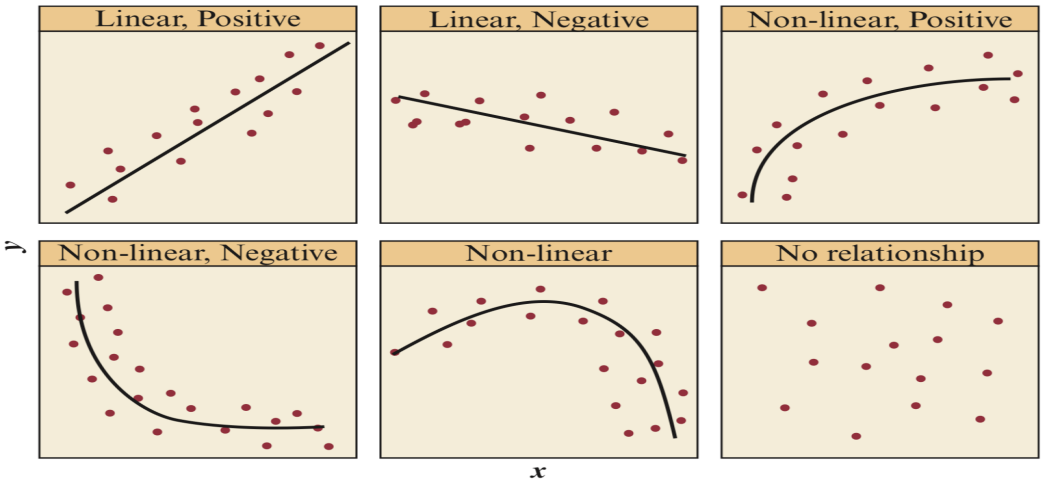
Dispersão e valores de r



Correlação de Pearson (r) - Resumo (1/2)

- **Interpretação**: r mede *força* e *direção* da associação **linear** entre duas variáveis quantitativas.
  - Sinal: $r > 0$ (quando x aumenta, y tende a aumentar); $r < 0$ (quando x aumenta, y tende a diminuir).
  - Magnitude: quanto mais perto de $|1|$, mais forte a associação linear; próximo de 0 indica fraca associação linear.
- **Escala**: $r \in [-1, 1]$; *sem unidade* (padronização por desvios-padrão); *simétrico* em (x, y), isto é, $r(x, y) = r(y, x)$.
- **Linearidade**: r captura padrão **linear**.
  - Em relações **curvilíneas**, r pode ser ≈ 0 mesmo havendo relação (vide Fig-3.9).

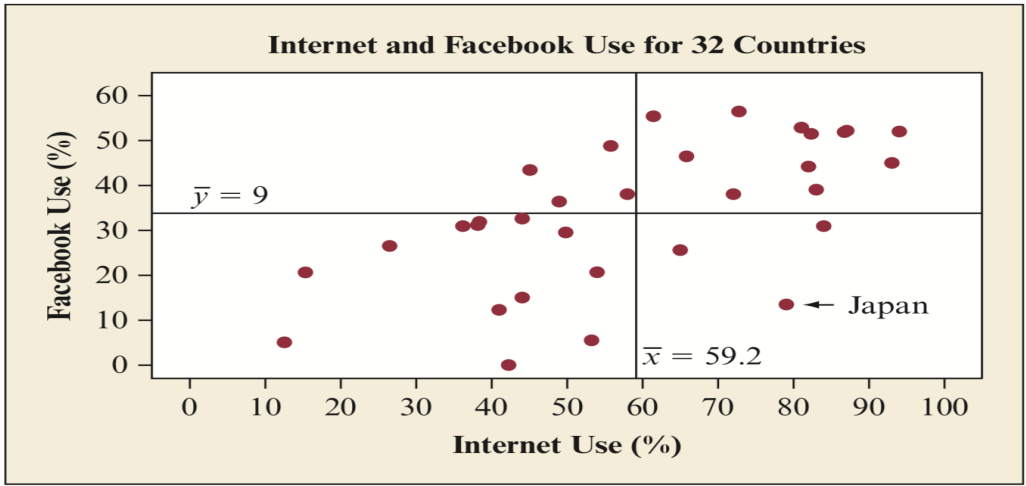
Padrões de associação: quanti x quanti



Correlação de Pearson (r) - Resumo (2/2)

- **Sensível a outliers**: um ponto extremo pode alterar muito r (até inverter o sinal).
  - Regra prática: *sempre* inspecione o *scatterplot* antes de citar r.
- **Condições de uso (checagem mínima)**:
  - Variáveis **quantitativas**; relação **aproximadamente linear**.
  - **Sem grandes outliers** ou, se houver, tratar/justificar.
  - **Checar marginais** (histogramas): assimetrias/caudas podem sugerir transformação ou estratificação antes de calcular r.
- **Mensagem-chave**: reporte r **com** visualização (Fig-3.9). O gráfico revela linearidade/curvatura e outliers; r sozinho pode induzir a erro.

Internet × Facebook



Dispersão, correlação e reta (base R × tidyverse)

base R

```
df <-
↪  read.csv("assets/data/facebook_data.csv")
plot(df$internet_pct, df$facebook_pct,
     xlab = "Uso de Internet (%)", ylab =
     ↪  "Uso de Facebook (%)")
abline(lm(facebook_pct ~ internet_pct, data =
↪  df), col = "gray")
cor(df$internet_pct, df$facebook_pct, use =
↪  "complete.obs")
```

tidyverse

```
library(tidyverse)
df <-
↪  read_csv("assets/data/facebook_data.csv")
ggplot(df, aes(internet_pct, facebook_pct)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Uso de Internet (%)", y = "Uso de
  ↪  Facebook (%)")
df |> summarize(r = cor(internet_pct,
↪  facebook_pct, use = "complete.obs"))
```