Estadística Descriptiva y Exploración de Datos

Análisis Programático con R

Prof. Nicolás Sidicaro
30 de septiembre de 2025

Objetivos de la clase

- Aplicar estadística descriptiva con tidyverse
- Interpretar distribuciones y detectar problemas
- Identificar cuándo transformar variables
- Preparar datos para modelado

¿Por qué estadística en análisis de datos?

Hasta ahora en el curso:

- Manipular datos: `filter()`, `select()`, `mutate()`
- Reorganizar: `pivot_longer()`, `pivot_wider()`
- Integrar: `left_join()`, `inner_join()`
- Manejar fechas: `lubridate`

Pero... ¿cómo sabemos si nuestros datos tienen sentido?

Estadística descriptiva: Primera línea de defensa contra datos problemáticos

Nuestro caso de estudio: Salarios en EEUU

```
# Datos reales: Encuesta salarial 1976
data("wage1")
salarios ← as_tibble(wage1)

glimpse(salarios)
```

```
## Rows: 526
## Columns: 24
## $ wage     <dbl> 3.10, 3.24, 3.00, 6.00, 5.30, 8.75, 11.25, 5.00, 3.60, 18.18, 6.25, 8.…
## $ educ     <int> 11, 12, 11, 8, 12, 16, 18, 12, 12, 17, 16, 13, 12, 12, 12, 16, 12, 13,…
## $ exper    <int> 2, 22, 2, 44, 7, 9, 15, 5, 26, 22, 8, 3, 15, 18, 31, 14, 10, 16, 13, 3…
## $ tenure   <int> 0, 2, 0, 28, 2, 8, 7, 3, 4, 21, 2, 0, 0, 3, 15, 0, 0, 10, 0, 6, 4, 13,…
## $ nonwhite <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ female   <int> 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1…
## $ married  <int> 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0…
## $ numdep   <int> 2, 3, 2, 0, 1, 0, 0, 0, 2, 0, 0, 0, 2, 0, 1, 1, 0, 0, 3, 0, 0, 3, 0, 0…
## $ smsa     <int> 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ northcen <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ south    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ west     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ construc <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ ndurman  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ trcommpu <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ trade    <int> 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
```

Exploración básica con tidyverse

```
salarios %>%
  summarise(
    n_trabajadores = n(),
    salario_promedio = mean(wage),
    salario_mediano = median(wage),
    edad_promedio = mean(educ),
    experiencia_promedio = mean(exper)
  )
```

```
## # A tibble: 1 × 5
##   n_trabajadores salario_promedio salario_mediano edad_promedio experiencia_promedio
##            <int>            <dbl>           <dbl>         <dbl>                <dbl>
## 1            526             5.90            4.65          12.6                 17.0
```

¿Qué nos dicen estos números?

- Tenemos 526 trabajadores
- Salario promedio: $5.90/hora
- Pero... ¿es representativo?

Definiciones formales

Medidas de tendencia central y dispersión

Media aritmética

Definición formal

La **media** (o promedio) es la suma de todos los valores dividida por el número de observaciones:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

donde:

- $\bar{x}$ = media muestral
- n = número de observaciones
- x_i = cada observación individual

En R:

```
# Tres formas de calcular la media
mean(salarios$wage)                    # Función base
salarios %>% summarise(mean(wage))     # Con tidyverse
sum(salarios$wage) / length(salarios$wage)  # Manualmente
```

Media: Propiedades y cuándo usarla

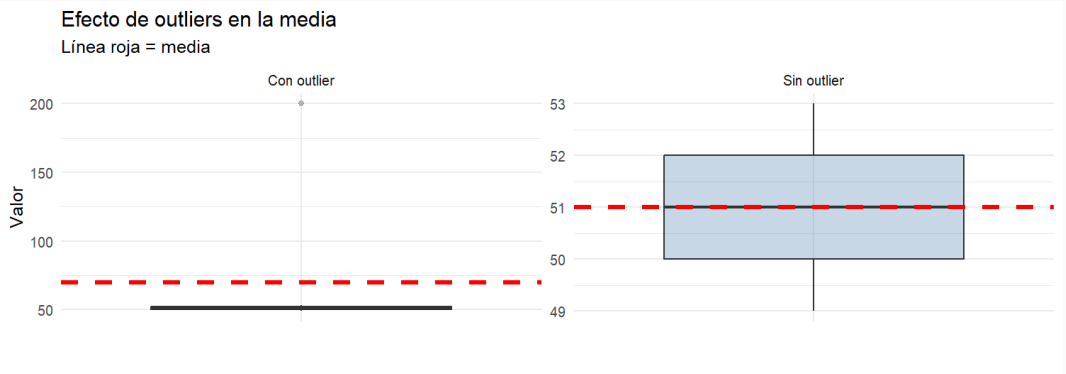
Propiedades importantes:

Ventajas:

- Usa toda la información disponible
- Tiene propiedades matemáticas útiles
- Necesaria para cálculos de varianza
- Base de muchos modelos estadísticos

Sensibilidad:

- **MUY sensible a valores extremos** (outliers)
- Un solo valor muy grande/pequeño puede distorsionarla



Media sin outlier: **51.0**

Media con outlier: **69.6** ⚠️

Media: Aplicación práctica

```r
# Comparando medias por grupo
salarios %>%
  mutate(genero = if_else(female == 1, "Mujer", "Hombre")) %>%
  group_by(genero) %>%
  summarise(
    n = n(),
    salario_promedio = mean(wage),
    salario_total = sum(wage),  # La media nos permite calcular totales
    .groups = "drop"
  )
```

```
## # A tibble: 2 × 4
##   genero     n salario_promedio salario_total
##   <chr>  <int>           <dbl>         <dbl>
## 1 Hombre   274            7.10         1945.
## 2 Mujer    252            4.59         1156.
```

Interpretación:

- La media nos dice cuánto gana en promedio cada grupo
- También nos permite estimar totales: total = media × n
- Pero cuidado: si hay outliers, puede no ser representativa

Mediana

Definición formal

La **mediana** es el valor que divide el conjunto ordenado de datos en dos mitades iguales:

$$\text{Mediana} = \begin{cases} x_{(n+1)/2} & \text{si } n \text{ es impar} \\ \frac{x_{n/2} + x_{(n/2)+1}}{2} & \text{si } n \text{ es par} \end{cases}$$

donde $x_{(i)}$ representa el i-ésimo valor cuando los datos están **ordenados de menor a mayor**.

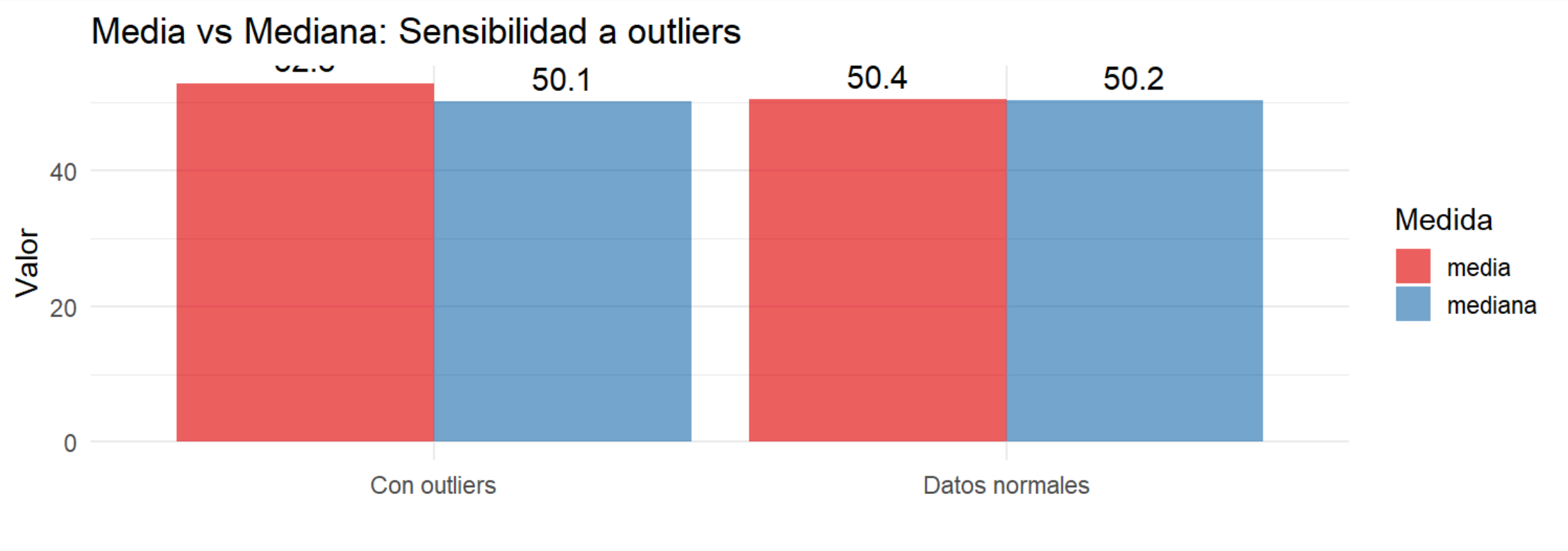
Paso a paso:

1. **Ordenar** los datos de menor a mayor
2. Si n es **impar**: tomar el valor central
3. Si n es **par**: promediar los dos valores centrales

Mediana: Ejemplo manual

```
## Valores ordenados: 45 47 48 49 50 51 52

## n = 7 (impar)

## Posición central: 4

## Mediana: 49

## Valores ordenados: 45 46 47 48 49 50 51 52

## n = 8 (par)

## Posiciones centrales: 4 y 5

## Valores centrales: 48 y 49

## Mediana: 48.5
```

Mediana: Robustez a outliers



Conclusión: La mediana es **robusta** - no se ve afectada por valores extremos

Percentiles y Cuartiles

Definición

Un **percentil** es un valor que deja un determinado porcentaje de observaciones por debajo:

- **Percentil 25** (P_{25} o Q_1): 25% de datos por debajo
- **Percentil 50** (P_{50} o Q_2): 50% de datos por debajo → **Es la mediana**
- **Percentil 75** (P_{75} o Q_3): 75% de datos por debajo

Los **cuartiles** dividen los datos en 4 partes iguales:

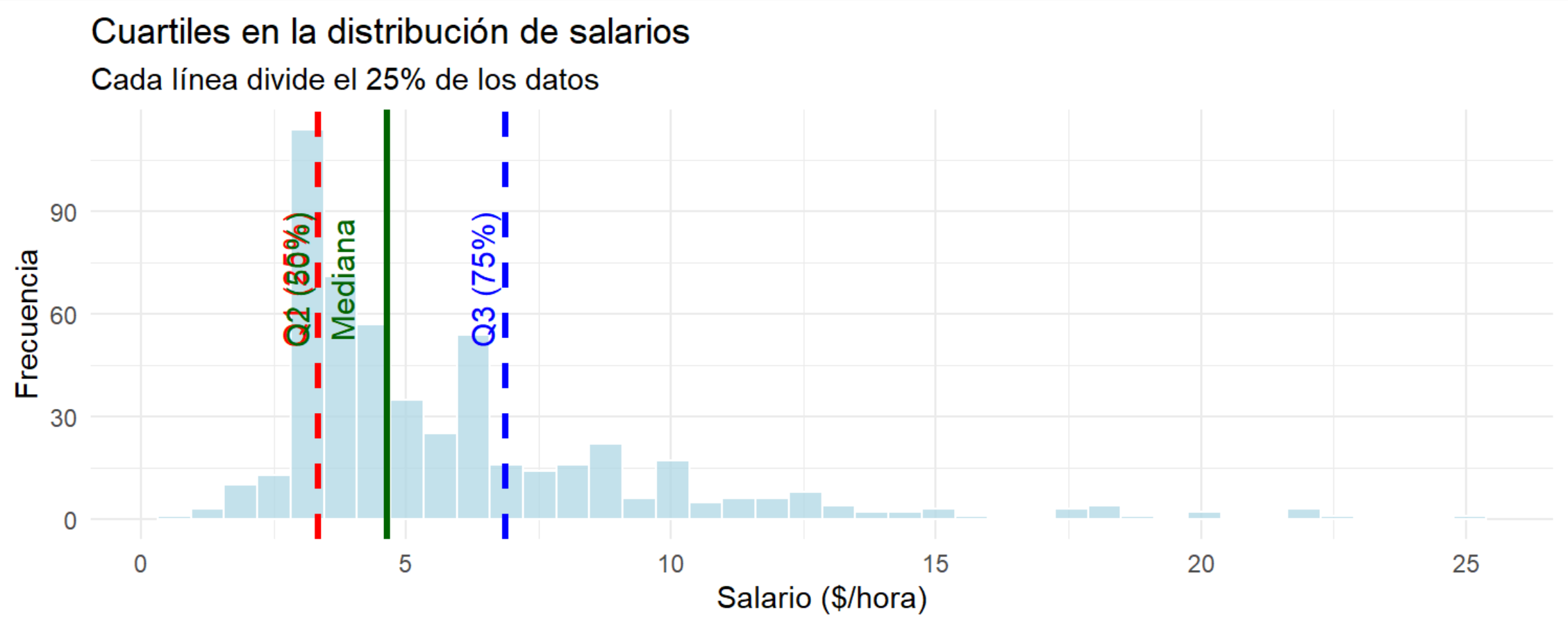
$$Q_1 \quad \leftarrow 25\% \rightarrow \quad Q_2 \text{ (mediana)} \quad \leftarrow 25\% \rightarrow \quad Q_3$$

Percentiles: Interpretación práctica

```r
# Calcular percentiles de salarios
salarios %>%
  summarise(
    minimo = min(wage),
    p10 = quantile(wage, 0.10),   # 10% gana menos
    q1 = quantile(wage, 0.25),    # 25% gana menos (Q1)
    mediana = median(wage),        # 50% gana menos (Q2)
    q3 = quantile(wage, 0.75),    # 75% gana menos (Q3)
    p90 = quantile(wage, 0.90),   # 90% gana menos
    maximo = max(wage)
  ) %>%
  pivot_longer(everything(), names_to = "estadistica", values_to = "valor") %>%
  mutate(valor = round(valor, 2))
```

```
## # A tibble: 7 × 2
##   estadistica valor
##   <chr>       <dbl>
## 1 minimo       0.53
## 2 p10          2.92
## 3 q1           3.33
## 4 mediana      4.65
## 5 q3           6.88
## 6 p90         10
## 7 maximo      25.0
```

Cuartiles: Visualización



Rango

Definición formal

El **rango** es la diferencia entre el valor máximo y el mínimo:

$$\text{Rango} = \max(x) - \min(x)$$

En R:

```
salarios %>%
  summarise(
    minimo = min(wage),
    maximo = max(wage),
    rango = max(wage) - min(wage),
    # Alternativa con función range()
    rango_alt = diff(range(wage))
  )
```

```
## # A tibble: 1 × 4
##   minimo maximo rango rango_alt
##    <dbl>  <dbl> <dbl>     <dbl>
## 1  0.530   25.0  24.4      24.4
```

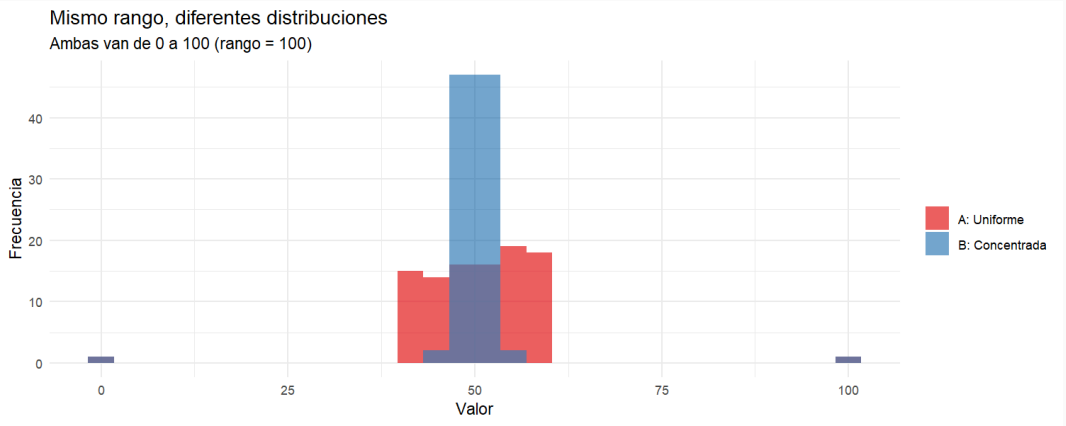
Rango: Limitaciones importantes

Ventajas:

- Muy fácil de calcular
- Intuitivo de entender
- Muestra el "espacio" de los datos

Desventajas:

- **Solo usa 2 valores** (min y max)
- **Muy sensible a outliers**
- **Ignora toda la información intermedia**
- No dice nada sobre concentración



Ambas tienen **rango = 100**,
pero son muy diferentes

Rango Intercuartílico (IQR)

Definición formal

El **Rango Intercuartílico** es la diferencia entre el tercer y primer cuartil:

$$\text{IQR} = Q_3 - Q_1$$

Representa el **rango del 50% central** de los datos (entre percentiles 25 y 75).

En R:

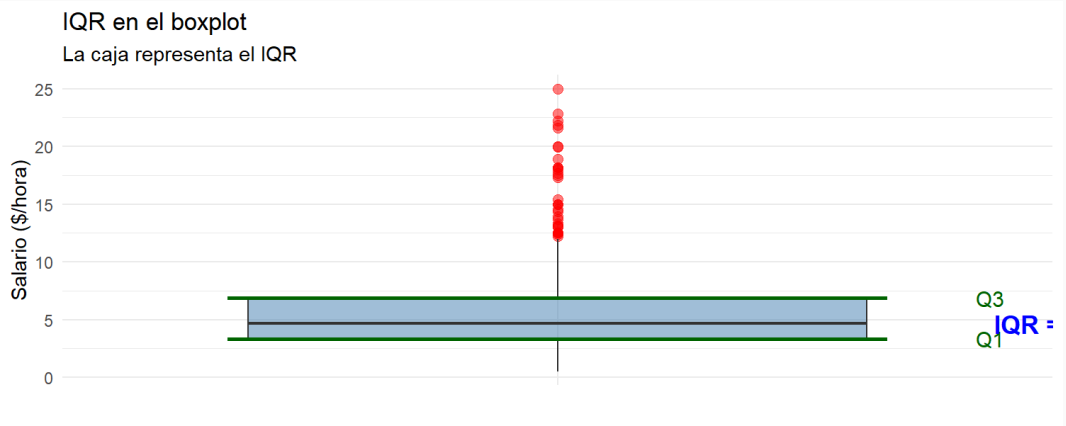
```
salarios %>%
  summarise(
    q1 = quantile(wage, 0.25),
    q3 = quantile(wage, 0.75),
    iqr_manual = q3 - q1,
    iqr_funcion = IQR(wage)  # Función directa
  )
```

```
## # A tibble: 1 × 4
##      q1    q3 iqr_manual iqr_funcion
##   <dbl> <dbl>      <dbl>       <dbl>
## 1  3.33  6.88       3.55        3.55
```

IQR: Ventajas sobre el rango simple

¿Por qué es mejor?

1. **Robusto a outliers**
   - No se ve afectado por valores extremos
   - Se enfoca en el "núcleo" de los datos
2. **Usa más información**
   - No solo 2 puntos, sino toda la distribución central
3. **Base para detectar outliers**
   - Regla: outlier si está fuera de $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$



IQR: Detección de outliers

```r
# Método formal para detectar outliers
salarios %>%
  mutate(
    q1 = quantile(wage, 0.25),
    q3 = quantile(wage, 0.75),
    iqr = IQR(wage),
    limite_inf = q1 - 1.5 * iqr,
    limite_sup = q3 + 1.5 * iqr,
    es_outlier = wage < limite_inf | wage > limite_sup,
    tipo_outlier = case_when(
      wage < limite_inf ~ "Outlier bajo",
      wage > limite_sup ~ "Outlier alto",
      TRUE ~ "Normal"
    )
  ) %>%
  count(tipo_outlier) %>%
  mutate(porcentaje = round(100 * n / sum(n), 1))
```

```
## # A tibble: 2 × 3
##   tipo_outlier     n porcentaje
##   <chr>        <int>      <dbl>
## 1 Normal         490       93.2
## 2 Outlier alto    36        6.8
```

Comparación: Rango vs IQR

```
salarios %>%
  summarise(
    # Rango simple
    rango_total = max(wage) - min(wage),
    # IQR
    iqr = IQR(wage),
    # Comparación
    ratio = rango_total / iqr,
    # Interpretación
    descripcion = paste0(
      "El rango total es ", round(ratio, 1),
      " veces más grande que el IQR"
    )
  ) %>%
  select(rango_total, iqr, descripcion)
```

```
## # A tibble: 1 × 3
##   rango_total   iqr descripcion
##         <dbl> <dbl> <chr>
## 1        24.4  3.55 "El rango total es 6.9 veces m\u00e1s grande que el IQR"
```

Moraleja: Si rango >> IQR, probablemente hay outliers importantes

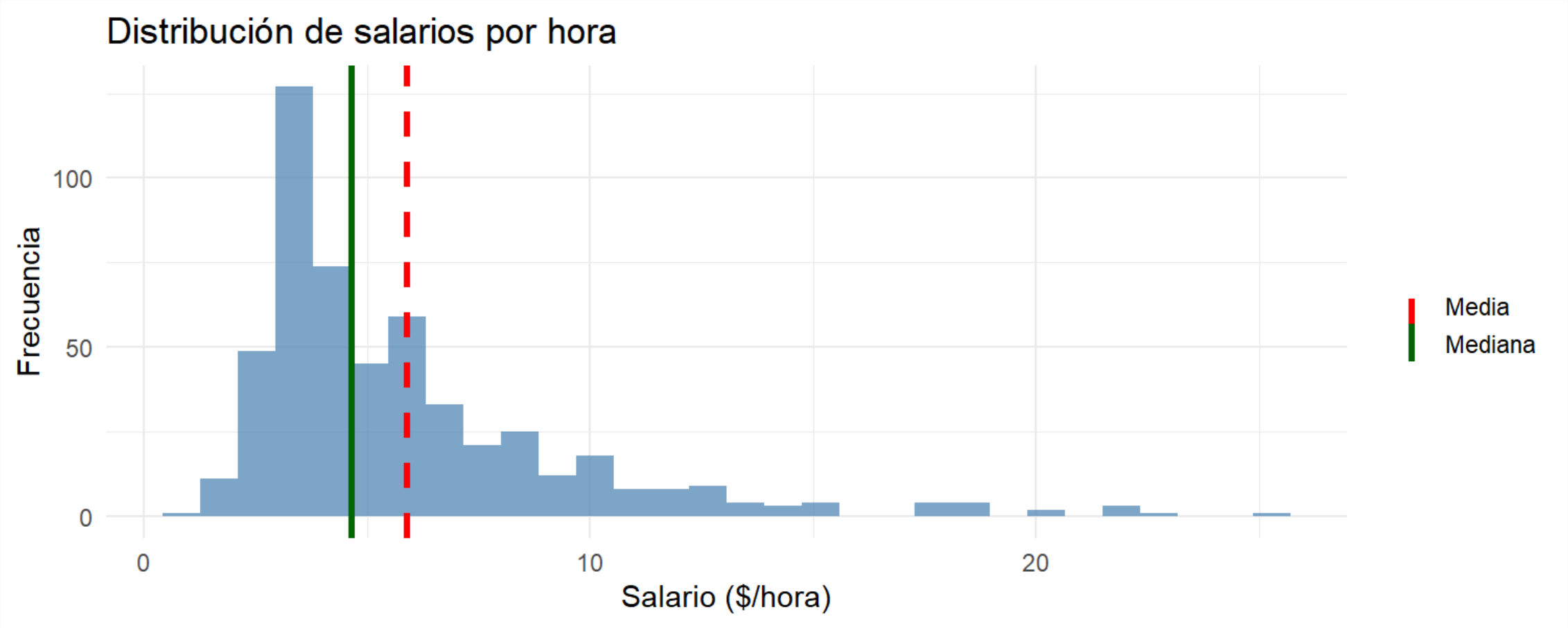
Resumen de medidas: ¿Cuál usar?

Medida	Ventaja principal	Desventaja principal	Cuándo usarla
Media	Usa toda la info	Sensible a outliers	Datos simétricos sin outliers
Mediana	Robusta a outliers	Ignora magnitudes	Datos asimétricos o con outliers
Rango	Simple e intuitivo	Solo usa 2 valores	Exploración inicial rápida
IQR	Robusto, informativo	Ignora colas	Detectar outliers, datos asimétricos

Regla práctica:

- Reporta **media Y mediana** juntas
- Si son muy diferentes → investigar por qué
- Complementa con **IQR** para entender dispersión
- Usa **visualizaciones** (histograma + boxplot)

Media vs Mediana: ¿Cuál usar?



Observación: Media > Mediana → Distribución asimétrica hacia la derecha

¿Por qué importa la asimetría?

En economía es frecuente:

Asimetría positiva (cola derecha):

- Ingresos personales
- Precios de viviendas
- Tamaño de empresas
- Riqueza

→ Media **sobreestima** valor típico

→ **Usar mediana** para reportar

Asimetría negativa (cola izquierda):

- Edad de jubilación
- Tiempo hasta conseguir empleo
- Días de stock

→ Media **subestima** valor típico

→ **Usar mediana** para reportar

Calculando asimetría con tidyverse

```
library(moments)

salarios %>%
  summarise(
    media = mean(wage),
    mediana = median(wage),
    diferencia = media - mediana,
    asimetria = skewness(wage),
    curtosis = kurtosis(wage)
  )
```

```
## # A tibble: 1 × 5
##   media mediana diferencia asimetria curtosis
##   <dbl>   <dbl>      <dbl>     <dbl>    <dbl>
## 1  5.90    4.65       1.25      2.01     7.97
```

Interpretación:

- Asimetría = 2.01 → **Fuerte asimetría positiva**
- Curtosis = 11.7 → **Colas pesadas** (más extremos que lo normal)

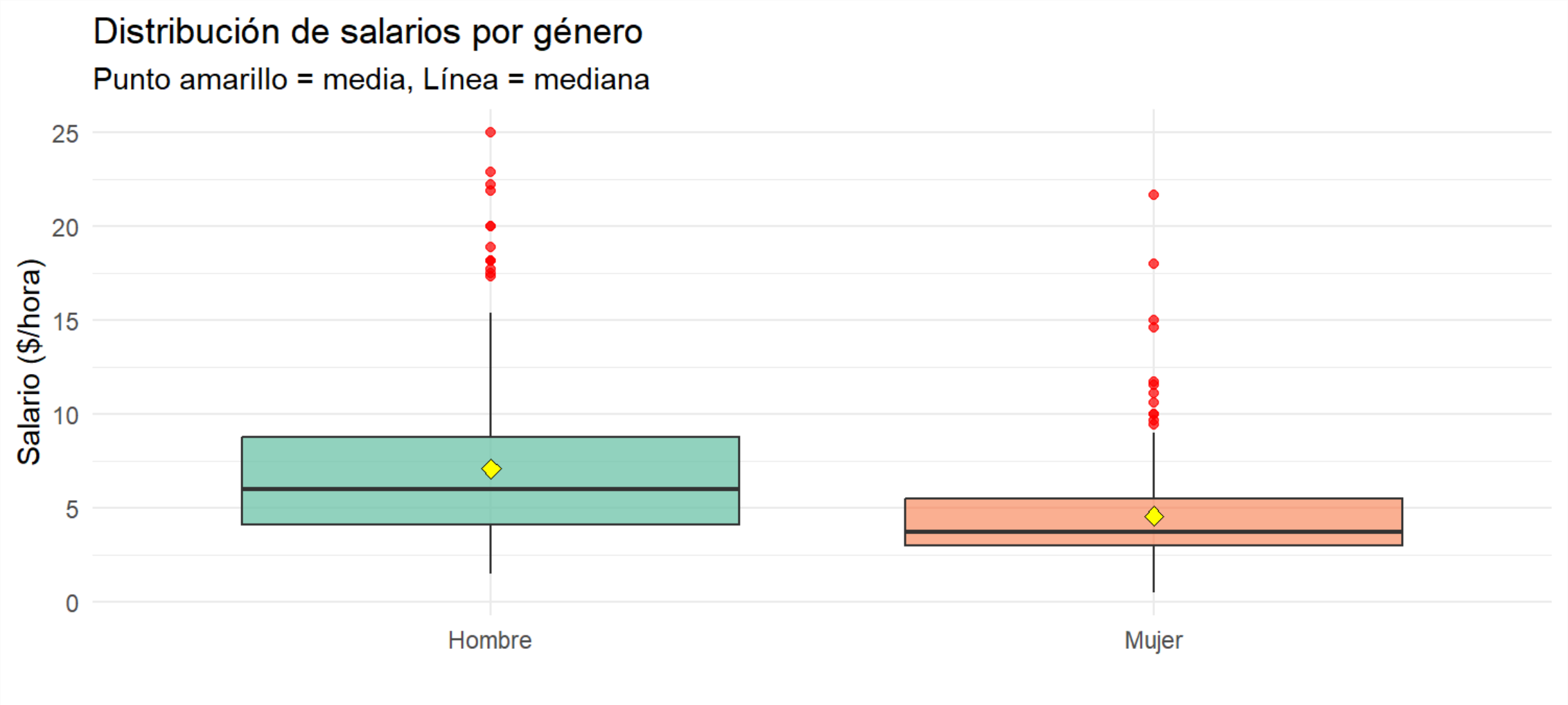
Medidas de dispersión: ¿Qué tan variable son los

```
salarios %>%
  summarise(
    media = mean(wage),
    desv_std = sd(wage),
    coef_var = sd(wage) / mean(wage) * 100,
    rango = max(wage) - min(wage),
    iqr = IQR(wage),
    percentil_25 = quantile(wage, 0.25),
    percentil_75 = quantile(wage, 0.75)
  ) %>%
  mutate(across(where(is.numeric), ~round(.x, 2)))
```

```
## # A tibble: 1 × 7
##   media desv_std coef_var rango   iqr percentil_25 percentil_75
##   <dbl>    <dbl>    <dbl> <dbl> <dbl>        <dbl>        <dbl>
## 1   5.9     3.69     62.6  24.4  3.55         3.33         6.88
```

Coeficiente de variación = 60.4% → Alta variabilidad relativa

Visualización de dispersión: Boxplot



Análisis por grupos con group_by()

```
## # A tibble: 4 × 6
##   nivel_educativo            n salario_mediano salario_promedio desv_std    cv
##   <chr>                  <int>           <dbl>            <dbl>    <dbl> <dbl>
## 1 Universitario+            99             7.5             8.95     4.75  53.1
## 2 Universidad incompleta   113             5               6.03     3.32  55.1
## 3 Secundario               198             4.5             5.37     3.09  57.6
## 4 Sin secundario           116             3.30            4.06     1.99  49.0
```

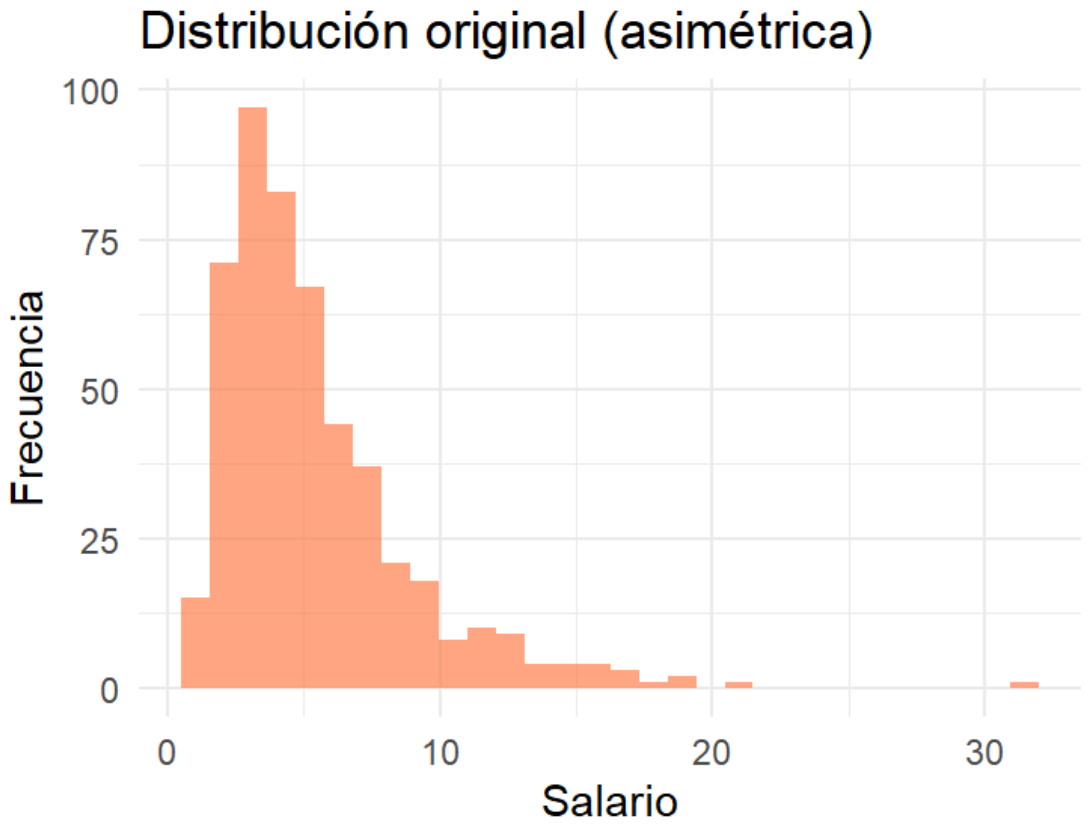
Insight: Retorno educativo claro, pero aumenta dispersión con educación

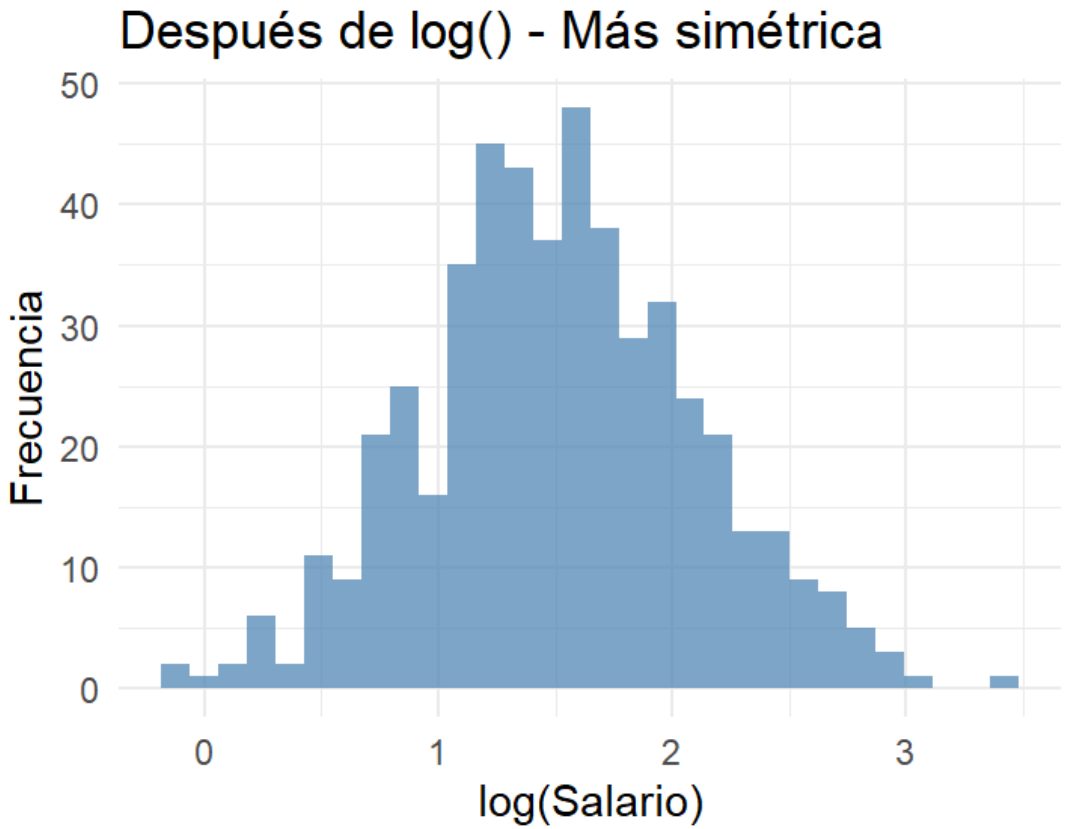
Detectando outliers: Método del IQR

```
## # A tibble: 10 × 4
##     wage  educ exper es_outlier
##    <dbl> <int> <int> <lgl>
##  1  25.0    18    29 TRUE
##  2  22.9    16    16 TRUE
##  3  22.2    12    31 TRUE
##  4  21.9    12    24 TRUE
##  5  21.6    18     8 TRUE
##  6  20      12    22 TRUE
##  7  20.0    14    26 TRUE
##  8  18.9    17    26 TRUE
##  9  18.2    17    22 TRUE
## 10  18.2    16    29 TRUE
```

¿Outliers o datos válidos? → Requiere juicio experto

El problema: ¿Deberían estos datos verse así?





Próxima clase: Transformaciones de variables

Introducción a probabilidad: ¿Por qué nos importa?

En análisis de datos usamos probabilidad para:

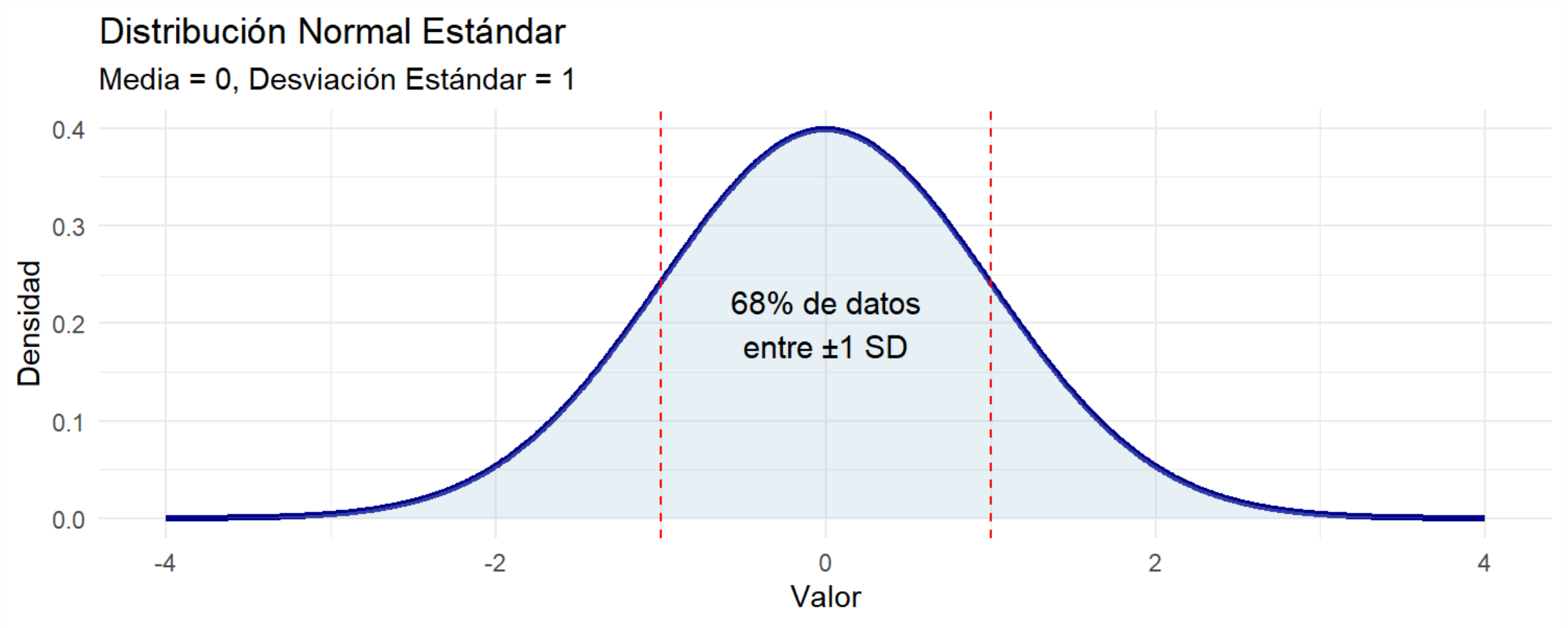
1. **Modelar incertidumbre**: "¿Cuál es la probabilidad de default?"
2. **Hacer inferencia**: "¿Este efecto es estadísticamente significativo?"
3. **Evaluar modelos**: "¿Qué tan probable es este resultado bajo mi modelo?"

Concepto clave: Distribución

Una **distribución de probabilidad** describe qué valores puede tomar una variable y cuán probables son.

En R: Muchas funciones trabajan con distribuciones (rnorm, dnorm, pnorm, qnorm)

La distribución más importante: Normal



Regla 68-95-99.7: 68% dentro de ±1σ, 95% dentro de ±1.96σ, 99.7% dentro de ±3σ

¿Por qué la Normal es importante?

1. Teorema Central del Límite (TCL)

Idea intuitiva: Si sumas/promedias muchas variables aleatorias, el resultado tiende a ser Normal

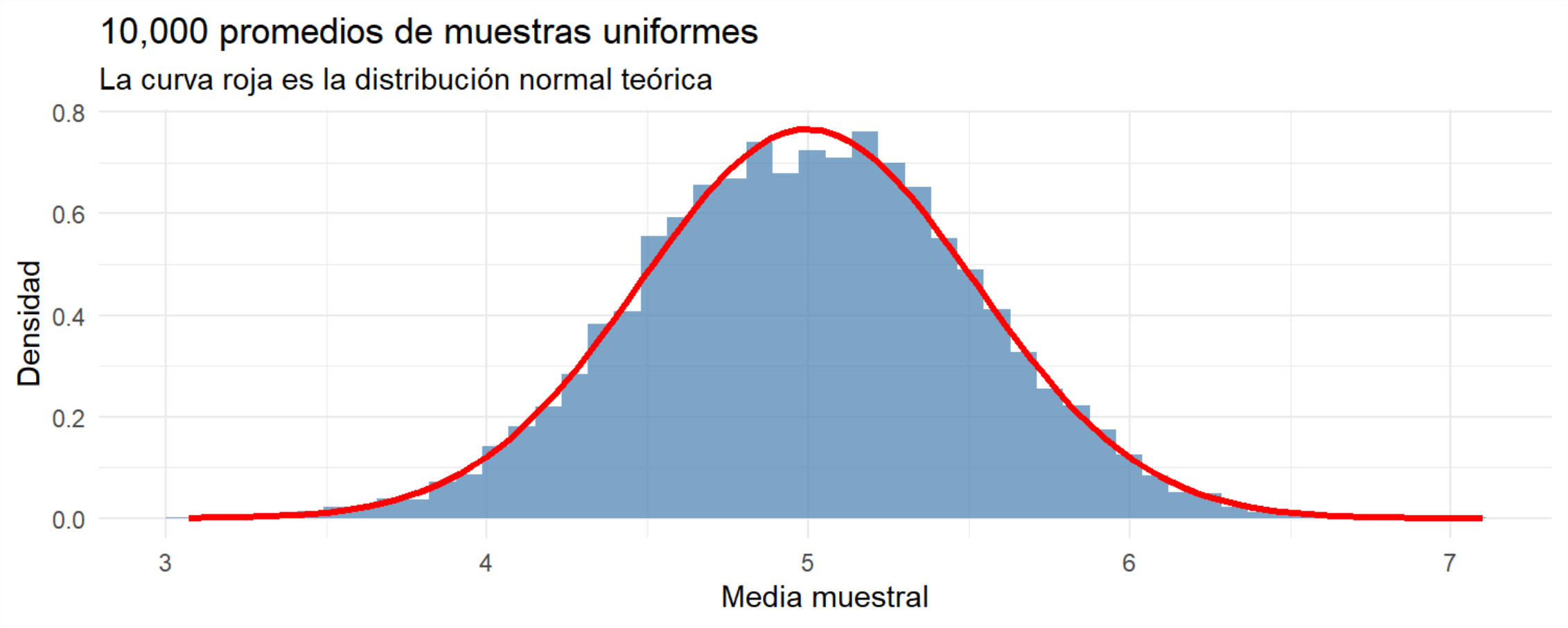
Implicación práctica:

- Los promedios muestrales son aproximadamente normales
- Los coeficientes de regresión son aproximadamente normales
- Podemos hacer inferencia sin conocer la distribución original

2. Muchos fenómenos naturales son aproximadamente normales

- Errores de medición
- Alturas, pesos (dentro de poblaciones homogéneas)
- Scores en tests estandarizados

Ejemplo del TCL: Simulación



Aplicación: Intervalos de confianza

```
# Salario promedio con intervalo de confianza al 95%
salarios %>%
  summarise(
    n = n(),
    media = mean(wage),
    error_std = sd(wage) / sqrt(n),  # Error estándar
    limite_inferior = media - 1.96 * error_std,
    limite_superior = media + 1.96 * error_std
  ) %>%
  mutate(across(c(media, error_std, limite_inferior, limite_superior),
                ~round(.x, 2)))
```

```
## # A tibble: 1 × 5
##       n media error_std limite_inferior limite_superior
##   <int> <dbl>     <dbl>           <dbl>           <dbl>
## 1   526   5.9      0.16            5.58            6.21
```

Interpretación: Estamos 95% seguros de que el salario promedio poblacional está entre $5.59 y $6.21

Esto funciona gracias al TCL → Las medias muestrales son aproximadamente normales

Otras distribuciones importantes (solo las esenciales)

t-Student

Cuándo: Muestras pequeñas (n < 30)

Por qué: Colas más pesadas que la Normal

En R: `t.test()`

```
# Test t para salarios por género
t.test(wage ~ female,
       data = salarios)
```

Chi-cuadrado

Cuándo: Variables categóricas

Por qué: Tests de independencia

En R: `chisq.test()`

```
# ¿El género afecta el sector?
tabla ← table(salarios$female,
              salarios$smsa)
chisq.test(tabla)
```

Conectando con el resto del curso

Lo que viene:

Clase 14: Tests estadísticos

- ¿Cómo saber si dos grupos son diferentes?
- Aplicaremos lo visto hoy sobre distribución Normal

Clase 15: Transformaciones de variables

- ¿Qué hacer con asimetría y curtosis?
- log(), sqrt(), escalado

Clases 18-19: Regresión lineal

- Los coeficientes son normales (gracias al TCL)
- La inferencia depende de supuestos sobre distribuciones

Pasos recomendados para análisis exploratorio

```r
# 1. Cargar y ver estructura
datos %>% glimpse()

# 2. Estadísticas descriptivas por grupos
datos %>%
  group_by(variable_categorica) %>%
  summarise(across(where(is.numeric),
                   list(mean = mean, median = median, sd = sd)))

# 3. Visualizar distribuciones
datos %>%
  ggplot(aes(x = variable_continua)) +
  geom_histogram() +
  facet_wrap(~variable_categorica)

# 4. Detectar problemas
# - Asimetría fuerte → Considerar transformación
# - Outliers → Investigar
# - Missing values → Revisar patrones

# 5. Documentar hallazgos
# - ¿Qué transformaciones aplicar?
# - ¿Qué variables son problemáticas?
# - ¿Qué preguntas surgen?
```

Checklist: ¿Tu análisis descriptivo está completo?

- [ ] **Estructura**: `glimpse()`, dimensiones, tipos de variables
- [ ] **Completitud**: ¿Hay NAs? ¿Cuántos? ¿Patrón?
- [ ] **Tendencia central**: Media y mediana calculadas
- [ ] **Dispersión**: Desviación estándar, IQR
- [ ] **Forma**: Asimetría, curtosis, outliers identificados
- [ ] **Visualización**: Histograma/densidad + boxplot
- [ ] **Por grupos**: Si aplica, comparar subpoblaciones
- [ ] **Decisiones**: ¿Transformar? ¿Eliminar outliers? ¿Imputar?

Puntos clave

1. Estadística descriptiva es **diagnóstico**, no solo resumen
2. Media vs Mediana importa cuando hay asimetría
3. Asimetría y curtosis afectan qué métodos usar después
4. La Normal es central por el Teorema Central del Límite
5. Siempre visualizar antes de modelar

Próxima clase

Tests estadísticos aplicados

- Comparación de medias: t-test
- Comparación de proporciones: chi-cuadrado
- Correlaciones y su significancia
- **Todo con tidyverse y datos económicos reales**

Preparación recomendada

- Revisar material de tidyverse (repaso)
- Experimentar con `gapminder` o datasets propios
- Pensar en preguntas de investigación para sus proyectos

¿Preguntas?

La estadística no es solo fórmulas, es entender tus datos

Referencias

- Wickham & Grolemund (2017). *R for Data Science*
- Ismay & Kim (2020). *Statistical Inference via Data Science: A ModernDive*
- Wooldridge, J.M. (2019). *Introductory Econometrics*

Recursos online

- **Tidyverse**: https://www.tidyverse.org
- **R for Data Science**: https://r4ds.had.co.nz
- **ModernDive**: https://moderndive.com

Datos usados

- `wooldridge::wage1` - Encuesta salarial EEUU 1976
- `gapminder` - Indicadores socioeconómicos 1952-2007