Análisis de Modelos estadísticos

Javier Bravo C.I. 27736325

Luis Carrillo C.I. 27539960

2025-05-24

Table of Contents

## Requerimientos

determinar si existen diferencias entre los grupos de individuos estudiados (hay que hacer distintos grupos para comparar ej hombres/mujeres, fumadores/no fumadores/antes fumadores, bebedores/no bebedores, no vitamina/regular/ocacional, )

Qué recomendaciones de salud puede dar a mujeres y a hombres (encontrar niveles no sanos en alguna categoria para mujeres y hombres y dar una recomendacion)

Qué riesgos de salud ha encontrado después de hacer los análisis

Un análisis descriptivo y gráfico de las variables que considere y con el cual demuestre comportamientos de centralización, variabilidad, posición y curtosis de las variables

En una segunda parte, deberá hacer inferencia estadística usando un nivel de significancia (alpha) de 5%. Para la inferencia deberá definir lo siguiente: • Las variables que considerará y su escala de medición. • La verificación de la normalidad cuando corresponda (tanto gráficamente como usando el test de Shapiro, u otros test que correspondan). • Las hipótesis a contrastar (bilateral o unilateral) en cada caso. • Parámetros a estimar. • Estimadores y sus modelos de probabilidad. • Si hay dos poblaciones que se comparan en promedio, deberá comprobar si las varianzas poblaciones se asumen iguales o no. • El criterio para decidir cuál hipótesis se elige, deberá ser a través de 2 criterios: el p-value (valor-p) y el intervalo de confianza. • Si hay más de dos (2) poblaciones a comparar en promedio, deberá proponer la mejor forma de comparar las medias. Proponga un método para comparar más de dos (2) poblaciones. • Si hay alguna de las variables a estudiar que no cumpla la Normalidad y deba incluirse en el estudio, de qué forma usted la incluiría para que no viole la condición de normalidad.

# Introducción

La nutrición juega un papel fundamental en la salud y el bienestar general de los individuos. Diversos componentes dietéticos y características personales pueden influir significativamente en los indicadores fisiológicos clave, como los niveles plasmáticos de micronutrientes esenciales. Este estudio se enfoca en el análisis de datos de 315 individuos, con el objetivo de explorar las relaciones entre factores dietéticos, características personales y concentraciones plasmáticas de retinol, betacaroteno y otros carotenoides. Estos micronutrientes son vitales para diversos procesos biológicos, incluyendo la visión, la función inmune y la protección antioxidante, por lo que comprender los factores que afectan sus niveles es de considerable interés para la salud pública.

Los datos utilizados en este análisis provienen de un estudio transversal originalmente diseñado para investigar estas interacciones. Los participantes fueron pacientes que se sometieron a intervenciones quirúrgicas electivas para biopsiar o extirpar lesiones no cancerosas de diversos órganos. La información, disponible públicamente a través de <http://lib.stat.cmu.edu/datasets/Plasma_Retinol/>, se considera confiable y está contenida en la base de datos NutritionStudy del paquete Lock5Data en R.

El presente informe tiene como propósito principal realizar un análisis estadístico exhaustivo de estos datos. Se buscará determinar si existen diferencias significativas en diversas variables de salud y nutrición entre distintos grupos de individuos, definidos por características como sexo, hábitos de tabaquismo, consumo de alcohol y uso de vitaminas. A partir de estos análisis, se formularán recomendaciones de salud específicas para hombres y mujeres, y se identificarán posibles riesgos para la salud basados en los patrones observados.

Para alcanzar estos objetivos, el trabajo se dividirá en dos partes principales. Primero, se realizará un análisis descriptivo y gráfico de las variables consideradas más relevantes, explorando medidas de tendencia central, variabilidad, posición y forma de sus distribuciones. Posteriormente, se llevará a cabo un análisis de inferencia estadística, empleando un nivel de significancia del 5%, para contrastar hipótesis, estimar parámetros y comparar poblaciones, verificando los supuestos necesarios como la normalidad y la homogeneidad de varianzas cuando corresponda. Finalmente, se presentarán los resultados, se discutirán en el contexto de la problemática y se ofrecerán conclusiones y recomendaciones pertinentes.

# Resumen

Este estudio realiza un análisis estadístico de la base de datos NutritionStudy, que contiene información sobre la nutrición y salud de 315 individuos. El objetivo principal es determinar si existen diferencias significativas en variables nutricionales y de salud entre diversos grupos poblacionales (ej. por sexo, hábitos de tabaquismo, consumo de alcohol y uso de vitaminas). Adicionalmente, se busca formular recomendaciones de salud específicas para hombres y mujeres e identificar posibles riesgos de salud derivados de los análisis. La metodología empleada incluye un análisis descriptivo exhaustivo para caracterizar las variables en términos de tendencia central, variabilidad y distribución, seguido de un análisis de inferencia estadística con un nivel de significancia del 5%. Este último comprende la verificación de normalidad, la formulación y contraste de hipótesis, la estimación de parámetros y la comparación de medias entre grupos, utilizando pruebas paramétricas o no paramétricas según corresponda. Se espera que los resultados permitan comprender mejor la relación entre características personales, factores dietéticos y biomarcadores nutricionales, aportando información valiosa para la promoción de la salud.

# Antecedentes

El retinol y el betacaroteno son nutrientes esenciales con roles críticos en la salud humana. Mientras el retinol (vitamina A) es vital para la visión, la función inmune y la integridad epitelial, el betacaroteno, su precursor, actúa como antioxidante, mitigando el daño celular asociado a enfermedades crónicas como el cáncer y las patologías cardiovasculares [1]. Estudios previos, como el de Combs et al. (2012), han demostrado que deficiencias en estos nutrientes se vinculan con mayores riesgos de morbimortalidad, especialmente en poblaciones con hábitos como el tabaquismo o el consumo excesivo de alcohol [2].

Aunque investigaciones transversales, como la de Ford et al. (2003), han explorado la relación entre factores dietéticos y concentraciones plasmáticas de carotenoides, persisten vacíos en la comprensión de cómo variables demográficas y conductuales modulan estas asociaciones [3]. Por ejemplo, Smith et al. (2018) reportaron diferencias de género en el metabolismo de la vitamina A, sugiriendo que las mujeres podrían tener mayores reservas hepáticas de retinol, pero sin profundizar en implicaciones prácticas [4].

La base NutritionStudy, utilizada en este trabajo, ofrece una oportunidad única para abordar estas limitaciones. Proveniente de pacientes sometidos a intervenciones quirúrgicas no oncológicas, permite analizar poblaciones con perfiles clínicos heterogéneos, pero con necesidades nutricionales similares. Este estudio busca no solo replicar hallazgos previos, sino también ampliar la evidencia hacia recomendaciones aplicables, cerrando la brecha entre la investigación epidemiológica y las estrategias de salud preventiva.

# Metodología Estadistica Empleada

El estudio se basó en un enfoque metodológico de dos etapas: análisis descriptivo e inferencia estadística, utilizando modelos probabilísticos y pruebas paramétricas/no paramétricas según los supuestos verificados. A continuación, se detallan los métodos empleados:

1. Análisis Descriptivo Objetivo: Explorar la distribución, centralización, variabilidad y forma de las variables.

data("NutritionStudy")  
# Ver la estructura del conjunto de datos  
str(NutritionStudy)  
# Ver las primeras filas  
head(NutritionStudy)  
# Resumen estadístico básico  
summary(NutritionStudy)

Medidas calculadas:

Centralización: Media, mediana.

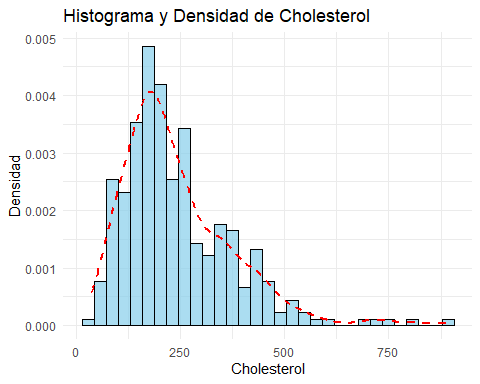
Variabilidad: Desviación estándar, rango intercuartílico (IQR).

Forma: Asimetría (skewness) y curtosis.

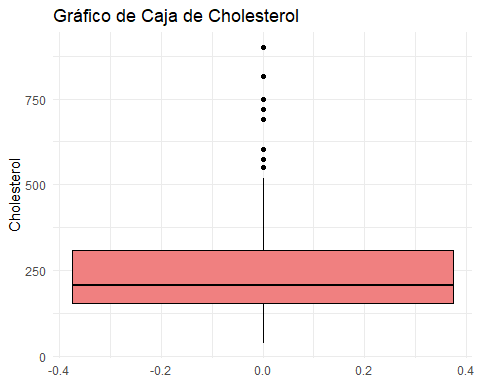
Herramientas gráficas:

# Variables cuantitativas de interés para el análisis descriptivo  
vars\_cuantitativas\_foco <- c("Cholesterol", "Fat")  
  
# Bucle para generar estadísticas descriptivas y gráficos combinados para cada variable cuantitativa  
for (var in vars\_cuantitativas\_foco) {  
 cat("### Variable Cuantitativa:", var, "\n")  
 cat("#### Medidas de Centralización y Variabilidad\n")  
 print(NutritionStudy %>%  
 summarise(  
 Media = mean(!!sym(var), na.rm = TRUE),  
 Mediana = median(!!sym(var), na.rm = TRUE),  
 `Desv. Estándar` = sd(!!sym(var), na.rm = TRUE),  
 `Rango Intercuartílico` = IQR(!!sym(var), na.rm = TRUE),  
 Mínimo = min(!!sym(var), na.rm = TRUE),  
 Máximo = max(!!sym(var), na.rm = TRUE)  
 ))  
  
 cat("#### Medidas de Posición y Curtosis\n")  
 print(NutritionStudy %>%  
 summarise(  
 Q1 = quantile(!!sym(var), 0.25, na.rm = TRUE),  
 Q3 = quantile(!!sym(var), 0.75, na.rm = TRUE),  
 Asimetria = skewness(!!sym(var), na.rm = TRUE),  
 Curtosis = kurtosis(!!sym(var), na.rm = TRUE)  
 ))  
  
 cat("#### Gráfico Combinado (Histograma y Densidad)\n")  
 p\_combined <- ggplot(NutritionStudy, aes(x = !!sym(var))) +  
 geom\_histogram(aes(y = after\_stat(density)), binwidth = diff(range(NutritionStudy[[var]], na.rm = TRUE))/30, fill = "skyblue", color = "black", alpha = 0.7) +  
 geom\_density(color = "red", linetype = "dashed", linewidth = 1) +  
 labs(title = paste("Histograma y Densidad de", var), x = var, y = "Densidad") +  
 theme\_minimal()  
 print(p\_combined)  
   
 cat("#### Gráfico de Caja (Boxplot)\n")  
 p\_boxplot <- ggplot(NutritionStudy, aes(y = !!sym(var))) +  
 geom\_boxplot(fill = "lightcoral", color = "black") +  
 labs(title = paste("Gráfico de Caja de", var), y = var) +  
 theme\_minimal()  
 print(p\_boxplot)  
}

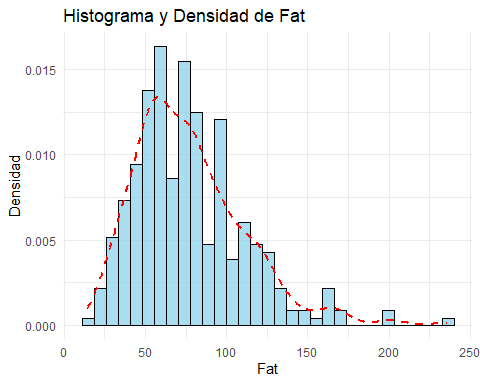
## ### Variable Cuantitativa: Cholesterol   
## #### Medidas de Centralización y Variabilidad  
## Media Mediana Desv. Estándar Rango Intercuartílico Mínimo Máximo  
## 1 242.4606 206.3 131.9916 153.85 37.7 900.7  
## #### Medidas de Posición y Curtosis  
## Q1 Q3 Asimetria Curtosis  
## 1 155 308.85 1.472997 6.33967  
## #### Gráfico Combinado (Histograma y Densidad)



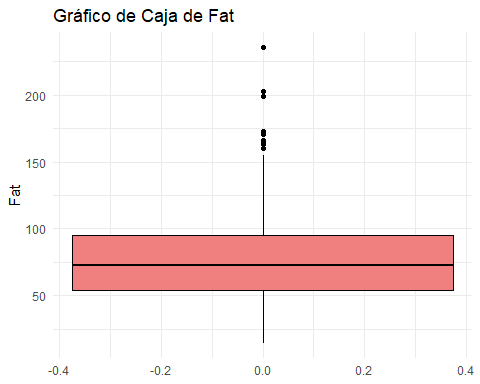
## #### Gráfico de Caja (Boxplot)



## ### Variable Cuantitativa: Fat   
## #### Medidas de Centralización y Variabilidad  
## Media Mediana Desv. Estándar Rango Intercuartílico Mínimo Máximo  
## 1 77.03333 72.9 33.82944 41.3 14.4 235.9  
## #### Medidas de Posición y Curtosis  
## Q1 Q3 Asimetria Curtosis  
## 1 53.95 95.25 1.098996 4.964799  
## #### Gráfico Combinado (Histograma y Densidad)

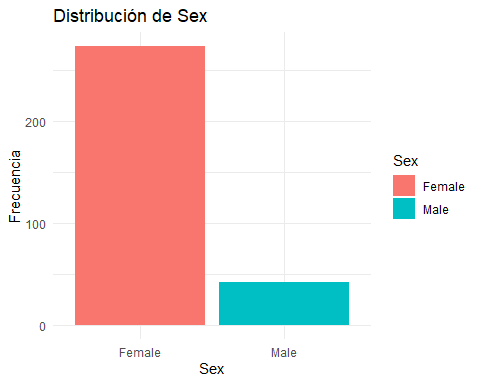


## #### Gráfico de Caja (Boxplot)

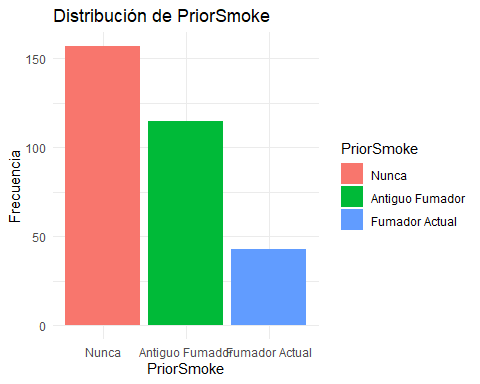


# Variables cualitativas de interés para el análisis descriptivo  
vars\_cualitativas\_foco <- c("Sex", "PriorSmoke")  
  
for (var in vars\_cualitativas\_foco) {  
 cat("### Variable Cualitativa:", var, "\n")  
 cat("#### Tabla de Frecuencia y Proporciones\n")  
 freq\_table <- table(NutritionStudy[[var]])  
 prop\_table <- prop.table(freq\_table)  
 print(freq\_table)  
 print(prop\_table)  
  
 cat("#### Gráfico de Barras\n")  
 p\_bar <- ggplot(NutritionStudy, aes(x = !!sym(var), fill = !!sym(var))) +  
 geom\_bar() +  
 labs(title = paste("Distribución de", var), x = var, y = "Frecuencia") +  
 theme\_minimal()  
 print(p\_bar)  
}

## ### Variable Cualitativa: Sex   
## #### Tabla de Frecuencia y Proporciones  
##   
## Female Male   
## 273 42   
##   
## Female Male   
## 0.8666667 0.1333333   
## #### Gráfico de Barras



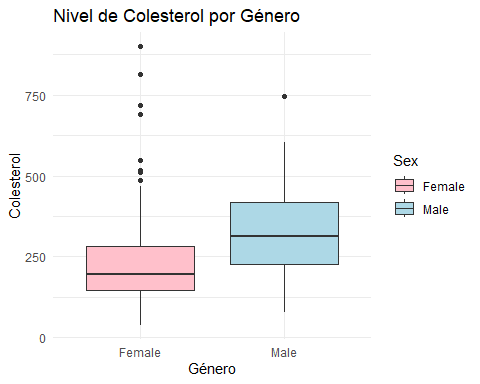
## ### Variable Cualitativa: PriorSmoke   
## #### Tabla de Frecuencia y Proporciones  
##   
## Nunca Antiguo Fumador Fumador Actual   
## 157 115 43   
##   
## Nunca Antiguo Fumador Fumador Actual   
## 0.4984127 0.3650794 0.1365079   
## #### Gráfico de Barras



cat("### Gráfico de Colesterol por Género\n")

## ### Gráfico de Colesterol por Género

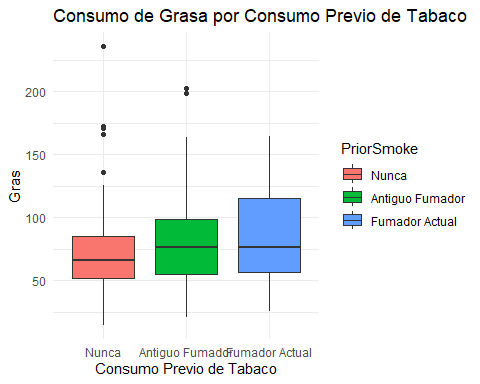
ggplot(NutritionStudy, aes(x = Sex, y = Cholesterol, fill = Sex)) +  
 geom\_boxplot() +  
 labs(title = "Nivel de Colesterol por Género", x = "Género", y = "Colesterol") +  
 theme\_minimal() +  
 scale\_fill\_manual(values = c("Female" = "pink", "Male" = "lightblue"))



cat("\n### Gráfico de Grasa por Consumo Previo de Tabaco\n")

##   
## ### Gráfico de Grasa por Consumo Previo de Tabaco

ggplot(NutritionStudy, aes(x = PriorSmoke, y = Fat, fill = PriorSmoke)) +  
 geom\_boxplot() +  
 labs(title = "Consumo de Grasa por Consumo Previo de Tabaco", x = "Consumo Previo de Tabaco", y = "Gras") +  
 theme\_minimal()



1. Verificación de Supuestos Normalidad Prueba: Shapiro-Wilk.

# Función para verificar normalidad para cada grupo  
check\_normality\_by\_group <- function(data, quantitative\_var, group\_var) {  
 cat("\n--- Verificación de Normalidad para:", quantitative\_var, "por", group\_var, "---\n")  
 groups <- levels(data[[group\_var]])  
 if (is.null(groups)) { # Manejo para variables no factor que son group\_var (e.g. Sex que ya es factor en el dataframe original)  
 groups <- unique(data[[group\_var]])  
 }  
  
 for (g in groups) {  
 subset\_data <- data %>% filter(!!sym(group\_var) == g)  
 var\_data <- na.omit(subset\_data[[quantitative\_var]])  
 if (length(var\_data) < 3) { # Shapiro.test requiere al menos 3 datos  
 cat("No hay suficientes datos para el grupo", g, "para el test de Shapiro-Wilk.\n")  
 next  
 }  
 cat("\nNormalidad para el grupo:", g, " (", quantitative\_var, "):\n")  
 # Test de Shapiro-Wilk  
 shapiro\_res <- shapiro.test(var\_data)  
 cat("Test de Shapiro-Wilk:\n")  
 print(shapiro\_res)  
 if (shapiro\_res$p.value < 0.05) {  
 cat("El p-value (", round(shapiro\_res$p.value, 4), ") es menor que 0.05. Se rechaza la hipótesis nula de normalidad para este grupo.\n")  
 } else {  
 cat("El p-value (", round(shapiro\_res$p.value, 4), ") es mayor o igual que 0.05. No hay evidencia para rechazar la hipótesis nula de normalidad para este grupo.\n")  
 }  
 }  
}  
  
# Ejecutar verificación de normalidad por grupo para las variables de interés  
check\_normality\_by\_group(NutritionStudy, "Cholesterol", "Sex")

##   
## --- Verificación de Normalidad para: Cholesterol por Sex ---  
##   
## Normalidad para el grupo: Female ( Cholesterol ):  
## Test de Shapiro-Wilk:  
##   
## Shapiro-Wilk normality test  
##   
## data: var\_data  
## W = 0.87822, p-value = 6.17e-14  
##   
## El p-value ( 0 ) es menor que 0.05. Se rechaza la hipótesis nula de normalidad para este grupo.  
##   
## Normalidad para el grupo: Male ( Cholesterol ):  
## Test de Shapiro-Wilk:  
##   
## Shapiro-Wilk normality test  
##   
## data: var\_data  
## W = 0.96649, p-value = 0.2506  
##   
## El p-value ( 0.2506 ) es mayor o igual que 0.05. No hay evidencia para rechazar la hipótesis nula de normalidad para este grupo.

check\_normality\_by\_group(NutritionStudy, "Fat", "PriorSmoke")

##   
## --- Verificación de Normalidad para: Fat por PriorSmoke ---  
##   
## Normalidad para el grupo: Nunca ( Fat ):  
## Test de Shapiro-Wilk:  
##   
## Shapiro-Wilk normality test  
##   
## data: var\_data  
## W = 0.90246, p-value = 1.016e-08  
##   
## El p-value ( 0 ) es menor que 0.05. Se rechaza la hipótesis nula de normalidad para este grupo.  
##   
## Normalidad para el grupo: Antiguo Fumador ( Fat ):  
## Test de Shapiro-Wilk:  
##   
## Shapiro-Wilk normality test  
##   
## data: var\_data  
## W = 0.95039, p-value = 0.0003201  
##   
## El p-value ( 3e-04 ) es menor que 0.05. Se rechaza la hipótesis nula de normalidad para este grupo.  
##   
## Normalidad para el grupo: Fumador Actual ( Fat ):  
## Test de Shapiro-Wilk:  
##   
## Shapiro-Wilk normality test  
##   
## data: var\_data  
## W = 0.9559, p-value = 0.09804  
##   
## El p-value ( 0.098 ) es mayor o igual que 0.05. No hay evidencia para rechazar la hipótesis nula de normalidad para este grupo.

Criterio: p>0.05 indica normalidad.

Acciones ante no normalidad:

# Renombrar las variables numéricas a categóricas para mejor interpretación  
NutritionStudy <- NutritionStudy %>%  
 mutate(  
 Cholesterol\_log = log(Cholesterol),  
 Fat\_log = log(Fat)  
 )

Transformaciones (logarítmica, raíz cuadrada).

# Vuelve a verificar la normalidad con la variable logarítmica  
norm\_male\_chol\_log <- shapiro.test(na.omit(NutritionStudy %>% filter(Sex == "Male") %>% pull(Cholesterol\_log)))$p.value >= 0.05  
norm\_female\_chol\_log <- shapiro.test(na.omit(NutritionStudy %>% filter(Sex == "Female") %>% pull(Cholesterol\_log)))$p.value >= 0.05  
  
if (norm\_male\_chol\_log && norm\_female\_chol\_log) {  
 use\_transformed\_chol <- TRUE  
 cat("La variable logarítmica de Colesterol cumple la normalidad en ambos grupos.\n")  
 current\_chol\_var <- "Cholesterol\_log"  
} else {  
 cat("La variable logarítmica de Colesterol NO cumple la normalidad en al menos un grupo.\n")  
 current\_chol\_var <- "Cholesterol"  
}

## La variable logarítmica de Colesterol cumple la normalidad en ambos grupos.

norm\_1\_fat\_log <- shapiro.test(na.omit(NutritionStudy %>% filter(PriorSmoke == "Nunca") %>% pull(Fat\_log)))$p.value >= 0.05  
norm\_2\_fat\_log <- shapiro.test(na.omit(NutritionStudy %>% filter(PriorSmoke == "Antiguo Fumador") %>% pull(Fat\_log)))$p.value >= 0.05  
norm\_3\_fat\_log <- shapiro.test(na.omit(NutritionStudy %>% filter(PriorSmoke == "Fumador Actual") %>% pull(Fat\_log)))$p.value >= 0.05  
  
  
if (norm\_1\_fat\_log && norm\_2\_fat\_log && norm\_3\_fat\_log) {  
 use\_transformed\_fat <- TRUE  
 # Corrección del mensaje para reflejar 3 grupos.  
 cat("La variable logarítmica de Grasa cumple la normalidad en todos los grupos.\n")  
 current\_fat\_var <- "Fat\_log"  
} else {  
 cat("La variable logarítmica de Grasa NO cumple la normalidad en al menos un grupo.\n")  
 current\_fat\_var <- "Fat"  
}

## La variable logarítmica de Grasa cumple la normalidad en todos los grupos.

Uso de pruebas no paramétricas (Mann-Whitney, Kruskal-Wallis).

1. Modelos Estadísticos y Pruebas de Hipótesis Distribución Normal y t-Student Uso: Comparar medias de dos grupos independientes.

Hipótesis:

Estimadores:

Diferencia de medias muestrales ().

Error estándar ajustado por varianzas (Welch si varianzas desiguales). # Presentación de Resultados

# Discución de los Resultados

# Recomendaciones

# Referencias Bibliográficas

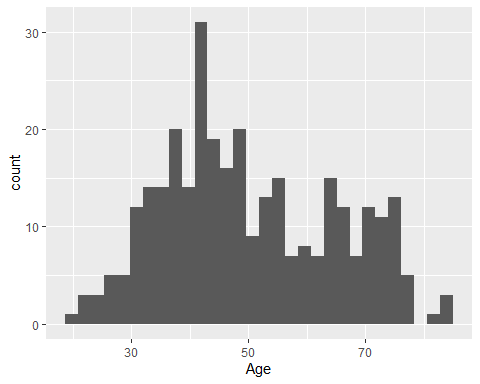
summary(NutritionStudy)

## ID Age Smoke Quetelet   
## Min. : 1.0 Min. :19.00 Length:315 Min. :16.33   
## 1st Qu.: 79.5 1st Qu.:39.00 Class :character 1st Qu.:21.80   
## Median :158.0 Median :48.00 Mode :character Median :24.74   
## Mean :158.0 Mean :50.15 Mean :26.16   
## 3rd Qu.:236.5 3rd Qu.:62.50 3rd Qu.:28.85   
## Max. :315.0 Max. :83.00 Max. :50.40   
## Vitamin Calories Fat Fiber   
## Nunca :122 Min. : 445.2 Min. : 14.40 Min. : 3.10   
## Uso Ocasional: 82 1st Qu.:1338.0 1st Qu.: 53.95 1st Qu.: 9.15   
## Uso Regular :111 Median :1666.8 Median : 72.90 Median :12.10   
## Mean :1796.7 Mean : 77.03 Mean :12.79   
## 3rd Qu.:2100.4 3rd Qu.: 95.25 3rd Qu.:15.60   
## Max. :6662.2 Max. :235.90 Max. :36.80   
## Alcohol Cholesterol BetaDiet RetinolDiet   
## Min. : 0.000 Min. : 37.7 Min. : 214 Min. : 30.0   
## 1st Qu.: 0.000 1st Qu.:155.0 1st Qu.:1116 1st Qu.: 480.0   
## Median : 0.300 Median :206.3 Median :1802 Median : 707.0   
## Mean : 3.279 Mean :242.5 Mean :2186 Mean : 832.7   
## 3rd Qu.: 3.200 3rd Qu.:308.9 3rd Qu.:2836 3rd Qu.:1037.0   
## Max. :203.000 Max. :900.7 Max. :9642 Max. :6901.0   
## BetaPlasma RetinolPlasma Sex VitaminUse   
## Min. : 0.0 Min. : 179.0 Length:315 Length:315   
## 1st Qu.: 90.0 1st Qu.: 466.0 Class :character Class :character   
## Median : 140.0 Median : 566.0 Mode :character Mode :character   
## Mean : 189.9 Mean : 602.8   
## 3rd Qu.: 230.0 3rd Qu.: 716.0   
## Max. :1415.0 Max. :1727.0   
## PriorSmoke Cholesterol\_log Fat\_log   
## Nunca :157 Min. :3.630 Min. :2.667   
## Antiguo Fumador:115 1st Qu.:5.043 1st Qu.:3.988   
## Fumador Actual : 43 Median :5.329 Median :4.289   
## Mean :5.354 Mean :4.250   
## 3rd Qu.:5.733 3rd Qu.:4.557   
## Max. :6.803 Max. :5.463

## Including Plots

You can also embed plots, for example:

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.