Exploratory Data Analysis

Last modified: 2026-11-12 22:02:32 (PST)

Introduction

Before fitting a model, it is good practice to explore and summarize the data. Exploratory data analysis (EDA) serves several purposes:

- understanding the distribution of each variable individually;
- detecting unusual values, outliers, and missing data;
- identifying patterns and relationships among variables;
- motivating choices of model structure, such as transformations;
- providing context for interpreting model results.

This page is adapted from [1, Ch. 2].

The WCGS data

This page illustrates exploratory data analysis using data from the Western Collaborative Group Study (WCGS) [2]. Vittinghoff et al. [1, p. 9] describe the study this way:

> The Western Collaborative Group Study (WCGS) was a large epidemiological study designed to investigate the association between the "type A" behavior pattern and coronary heart disease (CHD).

Exercise 0.1 (Type A behavior)**.** What is "type A" behavior?

Solution

Solution 0.1. From Wikipedia's article "Type A and Type B personality theory":

> The hypothesis describes Type A individuals as outgoing, ambitious, rigidly organized, highly status-conscious, impatient, anxious, proactive, and con-

> cerned with time management….
>
> The hypothesis describes Type B individuals as a contrast to those of Type A. Type B personalities, by definition, are noted to live at lower stress levels. They typically work steadily and may enjoy achievement, although they have a greater tendency to disregard physical or mental stress when they do not achieve.

Study design

The WCGS began in 1960 with 3,524 male volunteers employed by 11 California companies. Participants were 39 to 59 years old and free of heart disease, as determined by electrocardiogram. After the initial screening, various exclusions reduced the study population to 3,154 men and the number of companies to 10. The cohort comprised both blue- and white-collar employees. Average follow-up was 8.5 years, with repeat examinations.

This description paraphrases the help page for the `wcgs` dataset in the `faraway` R package.

At baseline, the study collected:

- socio-demographic characteristics: age, education, marital status, income, and occupation;
- physical and physiological measurements: height, weight, blood pressure, electrocardiogram, and corneal arcus;
- biochemical measurements: cholesterol and lipoprotein fractions;
- medical and family history, and use of medications;
- behavioral data: the Type A interview, smoking, exercise, and alcohol use.

Later surveys added anthropometry, triglycerides, the Jenkins Activity Survey, and caffeine use.

Loading the data

The WCGS data are distributed with Vittinghoff et al. [1] on the book's companion website, as a Stata file that R can read directly:

```
# one unbroken string, so that link checkers test the whole URL:
url <- "https://regression.ucsf.edu/sites/g/files/tkssra16191/files/wysiwyg/home/data/wcgs.d
wcgs <- haven::read_dta(url)
```

The `rmb` R package includes the same file, which these notes use so that rendering does not depend on the website. `haven::as_factor()` converts the Stata value labels to factors:

```r
wcgs <- rmb::wcgs |> haven::as_factor()
wcgs |> head()
#> # A tibble: 6 x 22
#>     age arcus behpat   bmi chd69  chol   dbp dibpat height    id lnsbp lnwght
#>   <dbl> <dbl> <fct>  <dbl> <fct> <dbl> <dbl> <fct>   <dbl> <dbl> <dbl>  <dbl>
#> 1    50     1 A1      31.3 No      249    90 Type A     67  2343  4.88   5.30
#> 2    51     0 A1      25.3 No      194    74 Type A     73  3656  4.79   5.26
#> 3    59     1 A1      28.7 No      258    94 Type A     70  3526  5.06   5.30
#> 4    51     1 A1      22.1 No      173    80 Type A     69 22057  4.84   5.01
#> 5    44     0 A1      22.3 No      214    80 Type A     71 12927  4.84   5.08
#> 6    47     0 A1      27.1 No      206    76 Type A     64 16029  4.75   5.06
#> # i 10 more variables: ncigs <dbl>, sbp <dbl>, smoke <fct>, t1 <dbl>,
#> #   time169 <dbl>, typchd69 <dbl>, uni <dbl>, weight <dbl>, wghtcat <fct>,
#> #   agec <fct>
```

```r
# attach a descriptive label to each variable;
# gtsummary tables print these labels instead of the column names:
wcgs_labels <- c(
  age = "Age (years)",
  chol = "Cholesterol (mg/dL)",
  sbp = "Systolic BP (mmHg)",
  dbp = "Diastolic BP (mmHg)",
  bmi = "BMI (kg/m^2)",
  weight = "Weight (lbs)",
  ncigs = "Cigarettes per day",
  chd69 = "CHD event by 1969",
  smoke = "Current smoker",
  arcus = "Arcus senilis",
  dibpat = "Behavioral pattern (A/B)",
  behpat = "Behavioral pattern (A1/A2/B3/B4)",
  wghtcat = "Weight category",
  agec = "Age group"
)
for (var in names(wcgs_labels)) {
  attr(wcgs[[var]], "label") <- wcgs_labels[[var]]
}

wcgs_labels
#>                                 age                                chol
#>                       "Age (years)"              "Cholesterol (mg/dL)"
#>                                 sbp                                 dbp
#>                "Systolic BP (mmHg)"               "Diastolic BP (mmHg)"
```

```
#>                                 bmi                                 weight
#>                       "BMI (kg/m^2)"                       "Weight (lbs)"
#>                               ncigs                                chd69
#>                 "Cigarettes per day"                  "CHD event by 1969"
#>                               smoke                                arcus
#>                     "Current smoker"                     "Arcus senilis"
#>                              dibpat                              behpat
#>           "Behavioral pattern (A/B)" "Behavioral pattern (A1/A2/B3/B4)"
#>                             wghtcat                                agec
#>                    "Weight category"                           "Age group"
```

The dataset has one row per participant:

```
dplyr::glimpse(wcgs)
#> Rows: 3,154
#> Columns: 22
#> $ age      <dbl> 50, 51, 59, 51, 44, 47, 40, 41, 50, 43, 59, 54, 48, 39, 49, 5~
#> $ arcus    <dbl> 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1~
#> $ behpat   <fct> A1, A1, A1, A1, A1, A1, A1, A1, A1, A1, A1, A1, A1, A1, A1, A~
#> $ bmi      <dbl> 31.3210, 25.3286, 28.6939, 22.1487, 22.3130, 27.1177, 23.2420~
#> $ chd69    <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, Yes, No, ~
#> $ chol     <dbl> 249, 194, 258, 173, 214, 206, 190, 212, 130, 233, 181, 214, 2~
#> $ dbp      <dbl> 90, 74, 94, 80, 80, 76, 78, 84, 70, 80, 86, 76, 78, 74, 80, 7~
#> $ dibpat   <fct> Type A, Type A, Type A, Type A, Type A, Type A, Type A, Type ~
#> $ height   <dbl> 67, 73, 70, 69, 71, 64, 70, 70, 71, 68, 72, 67, 71, 70, 73, 7~
#> $ id       <dbl> 2343, 3656, 3526, 22057, 12927, 16029, 3894, 11389, 12681, 10~
#> $ lnsbp    <dbl> 4.88280, 4.78749, 5.06259, 4.83628, 4.83628, 4.75359, 4.80402~
#> $ lnwght   <dbl> 5.29832, 5.25750, 5.29832, 5.01064, 5.07517, 5.06259, 5.08760~
#> $ ncigs    <dbl> 25, 25, 0, 0, 0, 80, 0, 25, 0, 25, 10, 0, 20, 0, 4, 0, 0, 20,~
#> $ sbp      <dbl> 132, 120, 158, 126, 126, 116, 122, 130, 112, 120, 130, 118, 1~
#> $ smoke    <fct> Yes, Yes, No, No, No, Yes, No, Yes, No, Yes, Yes, No, Yes, No~
#> $ t1       <dbl> -1.633353, -4.063366, 0.639729, 1.121768, 2.425011, -0.787520~
#> $ time169  <dbl> 1367, 2991, 2960, 3069, 3081, 2114, 2929, 3010, 3104, 2861, 2~
#> $ typchd69 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
#> $ uni      <dbl> 0.4860738, 0.1859543, 0.7277991, 0.6244636, 0.3789776, 0.7355~
#> $ weight   <dbl> 200, 192, 200, 150, 160, 158, 162, 160, 195, 187, 206, 152, 1~
#> $ wghtcat  <fct> 170-200, 170-200, 170-200, 140-170, 140-170, 140-170, 140-170~
#> $ agec     <fct> 46-50, 51-55, 56-60, 51-55, 41-45, 46-50, 35-40, 41-45, 46-50~
```

Populations and samples

Definition 0.1 (Population)**.** A **population** is the entire collection of units (people, animals, objects, or events) that a study aims to learn about.
Source: Wikipedia contributors [3].

Definition 0.2 (Population distribution)**.** The **population distribution** of a variable X is the distribution of X for a unit chosen at random from a population (Definition 0.1): the probability that X falls in a set A is

$$\mathbb{P}(A) \stackrel{\text{def}}{=} \Pr(X \in A),$$

where the probability refers to the random choice of the unit.

Remark

Remark. A population may be finite or infinite. For a finite population U of N units, with $X(u)$ the value of X for unit u, the random choice gives each unit probability $1/N$, so

$$\mathbb{P}(A) = \frac{|\{u \in U : X(u) \in A\}|}{N}.$$

An infinite population, such as all the coin flips a coin could ever produce, is described directly by a distribution. The population distribution can equally be described by its cumulative distribution function $\mathbb{F}(t) = \mathbb{P}((-\infty, t])$ or, when it has one, by its probability mass or density function.

Definition 0.3 (Observation)**.** An **observation** is the set of values recorded for the variables on one unit of a population (Definition 0.1). Each recorded value is a realization of the corresponding variable.

Example 0.1 (One observation in the WCGS)**.** Each row of the `wcgs` data frame is one observation: the values recorded for one participant. The first participant's observation has age 50 years and cholesterol 249 mg/dL.

Definition 0.4 (Sample)**.** A **sample** of size n is the data $x_1, \ldots, x_n$ observed on n units drawn from a population (Definition 0.1).

Definition 0.5 (Random sample)**.** A sample (Definition 0.4) is a **random sample** when the units are chosen by a chance mechanism such that the observations are independent and identically distributed draws from the population distribution (Definition 0.2):

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathbb{P},$$

where $\mathbb{P}$ is the population distribution.
Source: Wikipedia contributors [4].

Remark

Remark. Drawing without replacement from a finite population gives observations that are only approximately independent, and nearly so when the population is much larger than the sample. Few real studies draw a perfectly random sample, so conclusions about the population rest on how well the sampling mechanism approximates one.

Example 0.2 (Population and sample in the WCGS)**.** In 1960, the WCGS recruited employed men aged 39 to 59 years in California. Take that group of men as the study's population (Definition 0.1), so the population distribution (Definition 0.2) of cholesterol is its distribution among those men. The `wcgs` data frame is the sample (Definition 0.4): it holds 3154 of the men in that population. The men volunteered, through 10 companies, so the sample is not a random sample (Definition 0.5) of the population. Its sample mean cholesterol, 226.4 mg/dL, describes these 3154 men exactly, but it estimates the population mean well only to the extent that the volunteers resemble the rest of the population.

Definition 0.6 (Population parameter)**.** A **population parameter** is a number computed from the population distribution (Definition 0.2), such as its population mean.

Example 0.3 (A population parameter in the WCGS)**.** In Example 0.2, the population mean of cholesterol is a population parameter. It describes all the men in the study's population (Definition 0.1). We know only the mean cholesterol of the sample (Definition 0.4) of the `wcgs` data, 226.4 mg/dL.

Remark (Population parameters are fixed and usually unknown)

Remark 0.1 (Population parameters are fixed and usually unknown). A population parameter is fixed. It is usually unknown.

Summarizing a single variable

Measures of center

Definition 0.1 (Sample mean)**.** The **sample mean** of n observations $x_1, \ldots, x_n$ is:

$$\bar{x} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} x_i$$

Example 0.1 (Sample mean cholesterol in the WCGS)**.** Total cholesterol (`chol`) is missing for 12 of the 3154 WCGS participants. The sample mean of the remaining 3142 values is:

```
mean(wcgs$chol, na.rm = TRUE)
#> [1] 226.372
```

Example 0.2 (Sample mean versus population mean of cholesterol)**.** The population mean of cholesterol is the expectation of the population distribution (Definition 0.2). It is a fixed number that we do not know. The sample mean (Definition 0.1) of the `wcgs` data, 226.4 mg/dL, is computed from the observed data. A different sample from the same population would give a different sample mean, while the population mean would not change.

Definition 0.2 (Sample median)**.** The **sample median** is the middle value when the observations are sorted in increasing order. In terms of the order statistics $x_{(1)} \leq \cdots \leq x_{(n)}$:

- if n is odd, the median is $x_{((n+1)/2)}$;
- if n is even, the median is $\frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right)$.

Source: Wikipedia contributors [5].

Example 0.3 (Median cholesterol in the WCGS)**.**

```
median(wcgs$chol, na.rm = TRUE)
#> [1] 223
```

The median cholesterol is a little lower than the sample mean (Example 0.1). Cholesterol has a longer right tail than left tail (the maximum is 645 mg/dL), and the large values

in that tail pull the sample mean upward but barely change which value is in the middle. A median is less sensitive to extreme values than a sample mean.

Measures of spread

Definition 0.3 (Sample variance)**.** The **sample variance** of $n \geq 2$ observations is:

$$s^2 \stackrel{\text{def}}{=} \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Source: Wikipedia contributors [6].

Definition 0.4 (Sample standard deviation)**.** The **sample standard deviation** is the square root of the sample variance:

$$s \stackrel{\text{def}}{=} \sqrt{s^2}$$

Source: Wikipedia contributors [7].

Remark (Units of the standard deviation)

Remark 0.1 (Units of the standard deviation). The standard deviation has the same units as the observations, which makes it easier to interpret than the variance, whose units are squared.

Example 0.4 (Variance and standard deviation of cholesterol in the WCGS)**.**

```
var(wcgs$chol, na.rm = TRUE)
#> [1] 1885.33
sd(wcgs$chol, na.rm = TRUE)
#> [1] 43.4204
```

The sample variance is in $(\text{mg/dL})^2$, and the sample standard deviation is in mg/dL, like the cholesterol values themselves.

Definition 0.5 (Quartiles (first quartile, third quartile, 25th percentile, 75th percentile))**.** The **quartiles** are the **first quartile** Q_1 and the **third quartile** Q_3, the sample 0.25 and 0.75 quantiles (sample quantile). The first quartile is also called the **25th percentile**, and the third quartile is also called the **75th percentile**.

Source: Wikipedia contributors [8].

Remark (Quartiles in software)

Remark 0.2 (Quartiles in software). With an odd number of observations, the sample 0.5 quantile equals the sample median; with an even number they can differ, because the median averages the two middle order statistics. Software packages compute quartiles with different quantile definitions, so they can disagree slightly for the same data (quantile conventions); R's `quantile(type = 1)` matches this definition, and R's default, `type = 7`, interpolates.

Definition 0.6 (Interquartile range)**.** The **interquartile range** (IQR) is the difference between the third and first quartiles:

$$\mathrm{IQR} \stackrel{\text{def}}{=} Q_3 - Q_1$$

Source: Wikipedia contributors [9].

Example 0.5 (Quartiles and IQR of cholesterol in the WCGS)**.**

```
quantile(wcgs$chol, probs = c(0.25, 0.75), na.rm = TRUE, type = 1)
#> 25% 75%
#> 197 253
IQR(wcgs$chol, na.rm = TRUE, type = 1)
#> [1] 56
```

The middle half of the cholesterol values span 56 mg/dL. Like the median, the IQR does not depend on the most extreme values, so it is less sensitive to outliers than the standard deviation.

Summary statistics in R

The `summary()` function reports the minimum, quartiles, mean, maximum, and number of missing values in one call; its quartiles use R's default interpolating quantile rule (`type = 7`), so they can differ slightly from Definition 0.5:

```
summary(wcgs$chol)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NAs
#>     103     197     223     226     253     645      12
```

Table 1: WCGS: descriptive statistics for continuous variables

```
wcgs |>
  dplyr::select(age, chol, sbp, dbp, bmi, weight) |>
  gtsummary::tbl_summary(
    statistic = list(
      gtsummary::all_continuous() ~
        "{mean} ({sd}); {median} [{p25}, {p75}]"
    ),
    digits = gtsummary::all_continuous() ~ 1
  )
```

Characteristic	**N = 3,154**[1]
Age (years)	46.3 (5.5); 45.0 [42.0, 50.0]
Cholesterol (mg/dL)	226.4 (43.4); 223.0 [197.0, 253.0]
Unknown	12
Systolic BP (mmHg)	128.6 (15.1); 126.0 [120.0, 136.0]
Diastolic BP (mmHg)	82.0 (9.7); 80.0 [76.0, 86.0]
BMI (kg/m^2)	24.5 (2.6); 24.4 [23.0, 25.8]
Weight (lbs)	170.0 (21.1); 170.0 [155.0, 182.0]

[1]Mean (SD); Median [Q1, Q3]

For a formatted table of several variables at once, `gtsummary::tbl_summary()` is useful (Table 1).

Binary and categorical variables

Definition 0.7 (Sample proportion)**.** For a binary variable with values coded 1 ("success") and 0, the **sample proportion** of successes is:

$$\hat{p} \overset{\text{def}}{=} \frac{k}{n}$$

where k is the number of successes and n is the number of observations.

Example 0.6 (Proportion of WCGS participants with a CHD event)**.**

Table 2: Frequency table for categorical variables in the WCGS dataset

```
wcgs |>
  dplyr::select(chd69, smoke, dibpat, behpat, wghtcat) |>
  gtsummary::tbl_summary()
```

Characteristic	**N = 3,154**[1]
CHD event by 1969	257 (8.1%)
Current smoker	1,502 (48%)
Behavioral pattern (A/B)	
Type B	1,565 (50%)
Type A	1,589 (50%)
Behavioral pattern (A1/A2/B3/B4)	
A1	264 (8.4%)
A2	1,325 (42%)
B3	1,216 (39%)
B4	349 (11%)
Weight category	
< 140	232 (7.4%)
140-170	1,538 (49%)
170-200	1,171 (37%)
> 200	213 (6.8%)

[1] n (%)

```
table(wcgs$chd69)
#>
#>   No  Yes
#> 2897  257
mean(wcgs$chd69 == "Yes")
#> [1] 0.0814838
```

257 of the 3154 participants had a CHD event by 1969, a sample proportion of 0.081.

For categorical variables with more than two levels, the natural descriptive statistics are the count of observations in each category and the corresponding proportions (Table 2).

Read

- Hogg et al. [10, Sec. 6.1], pp. 233–244 covers descriptive statistics.
- Rice [11, Secs. 10.4–10.5], pp. 392–401 covers measures of location and dispersion.

Graphical methods

Graphs can reveal features of a distribution that summary statistics miss, such as skewness, multiple peaks, and outliers. Different types of graph suit different types of variable.

Histograms

Definition 0.1 (Histogram)**.** A **histogram** displays the distribution of a continuous variable by dividing the variable's range into intervals and drawing a bar over each interval whose height is the number (or proportion) of observations in that interval.
Source: Wikipedia contributors [12].

Definition 0.2 (Bin)**.** A **bin** of a histogram is one of its intervals.

Figure 1 shows a single-peaked distribution with a longer right tail than left tail: 97% of the values lie between 150 and 350 mg/dL, but the largest value is 645 mg/dL.

Density plots

Definition 0.3 (Density plot)**.** A **density plot** draws a smooth curve that estimates the probability density of a continuous variable, scaled so that the area under the curve is 1.
Source: Wikipedia contributors [13].

Box plots

Definition 0.4 (Box plot (box-and-whisker plot))**.** A **box plot** (or **box-and-whisker plot**) summarizes the distribution of a continuous variable:

- the box spans the first to the third quartile, so its length is the IQR;
- a line inside the box marks the median;

```r
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = chol) +
  ggplot2::geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  ggplot2::labs(x = "Cholesterol (mg/dL)", y = "Count")
```

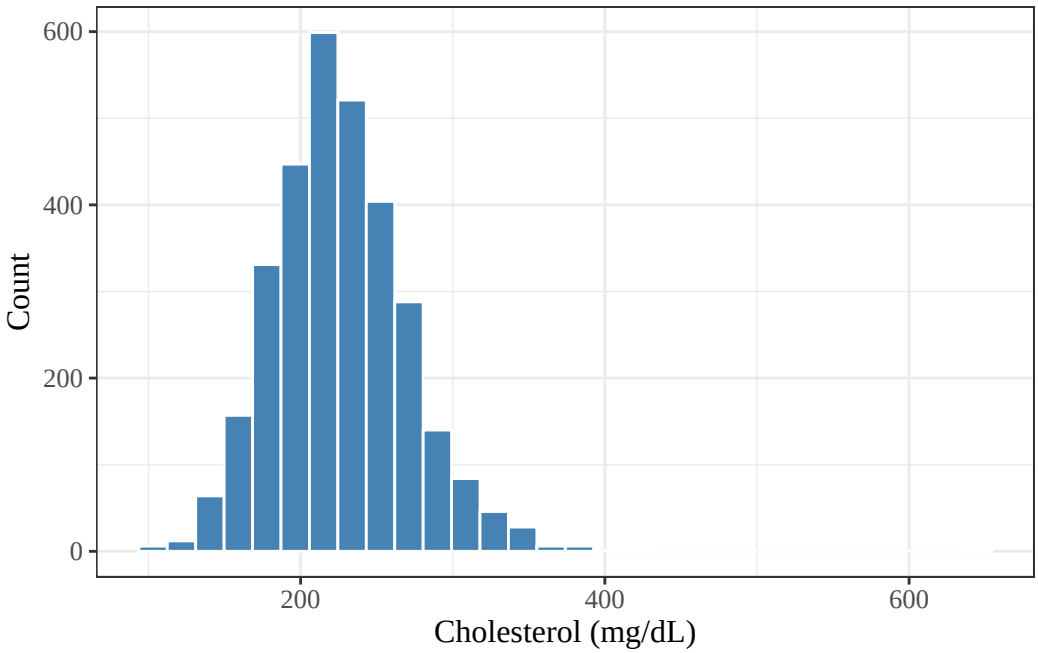


Figure 1: Histogram of total cholesterol in the WCGS dataset

```r
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = chol) +
  ggplot2::geom_density(fill = "steelblue", alpha = 0.5) +
  ggplot2::labs(x = "Cholesterol (mg/dL)", y = "Density")
```

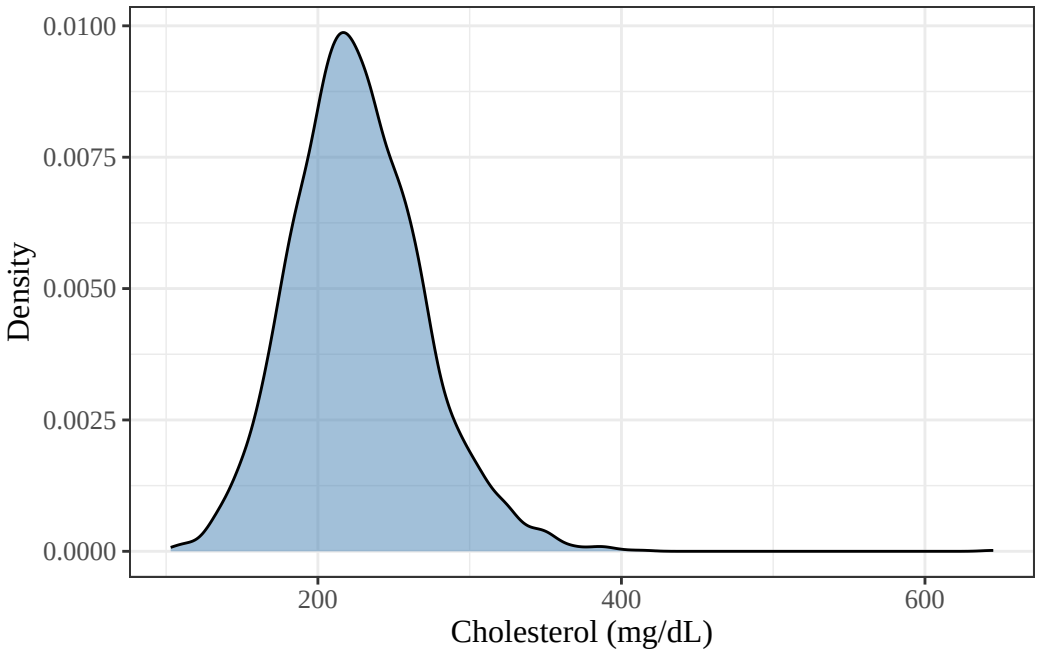


Figure 2: Density plot of total cholesterol in the WCGS dataset

- the whiskers extend from the box to the most extreme observations within 1.5×IQR of the box;
- observations beyond the whiskers are plotted individually, as potential outliers.

Source: Wikipedia contributors [14].

Remark (Whisker rules)

Remark 0.1 (Whisker rules). The $1.5 \times \text{IQR}$ whisker rule is the default in R's `boxplot()` and `ggplot2::geom_boxplot()`; other software and authors use other rules, such as whiskers that run to the minimum and maximum.

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(y = chol) +
  ggplot2::geom_boxplot(fill = "steelblue") +
  ggplot2::labs(y = "Cholesterol (mg/dL)") +
  ggplot2::theme(axis.text.x = ggplot2::element_blank())
```

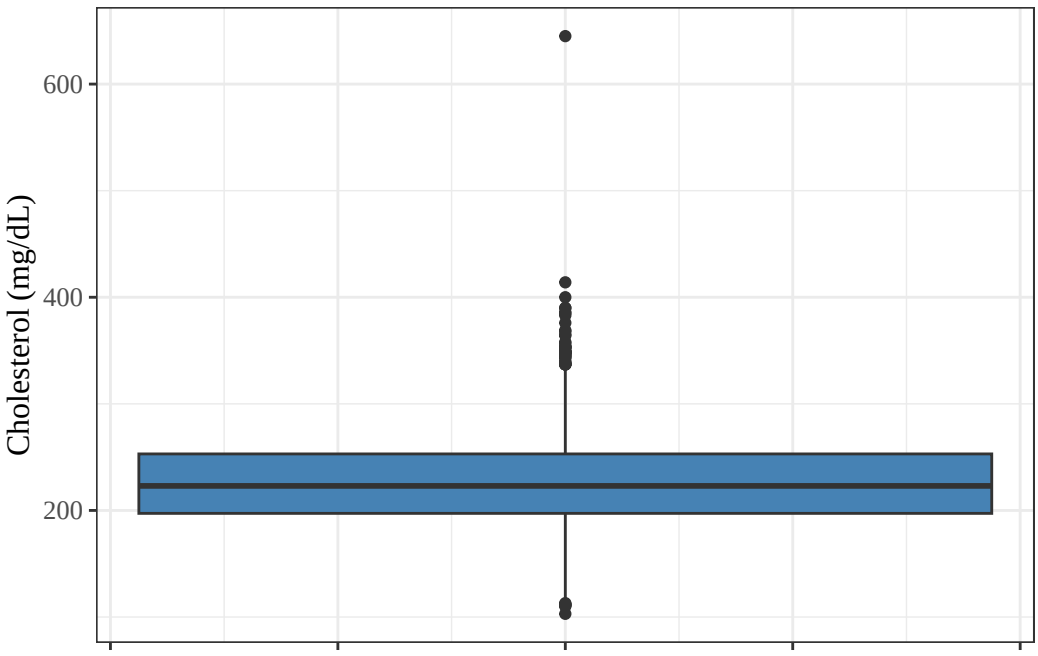


Figure 3: Box plot of total cholesterol in the WCGS dataset

Bar charts

Definition 0.5 (Bar chart)**.** A **bar chart** displays the number or proportion of observations in each category of a categorical variable, as one bar per category.
Source: Wikipedia contributors [15].

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = behpat) +
  ggplot2::geom_bar(fill = "steelblue") +
  ggplot2::labs(x = "Behavioral pattern", y = "Count")
```

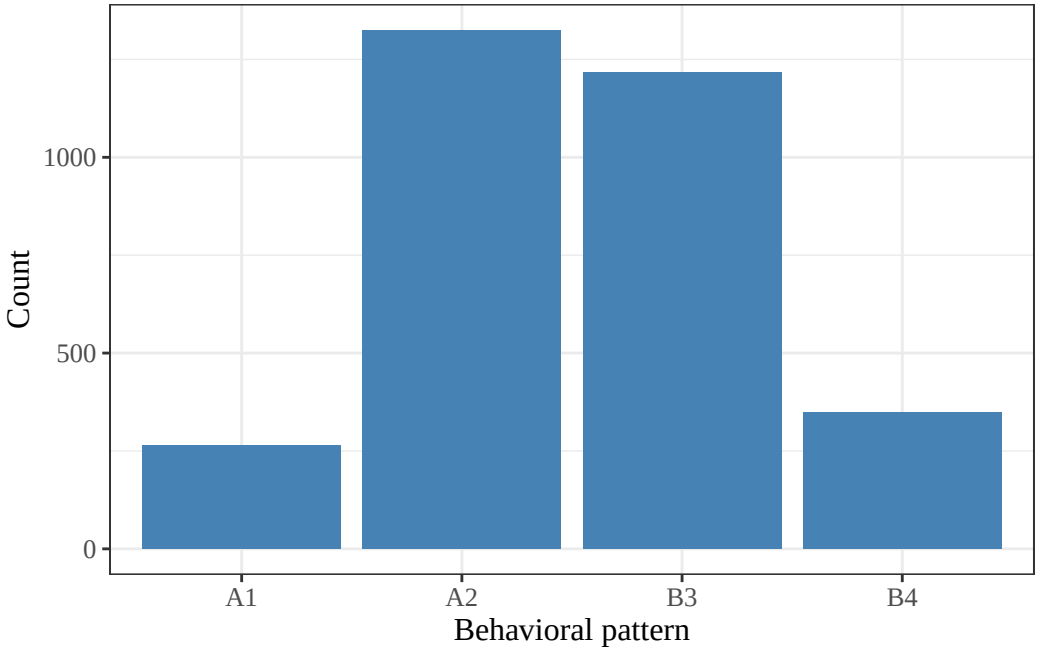


Figure 4: Bar chart of behavioral pattern in the WCGS dataset

Normal quantile-quantile plots

Definition 0.6 (Normal quantile-quantile plot)**.** A **normal quantile-quantile (Q-Q) plot** plots the sorted observations (order statistics) against the corresponding quantiles of a standard Gaussian distribution. If the variable is approximately Gaussian, the points fall close to a straight line.

Source: Wikipedia contributors [16].

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(sample = chol) +
  ggplot2::stat_qq() +
  ggplot2::stat_qq_line(color = "red") +
  ggplot2::labs(x = "Standard Gaussian quantiles", y = "Cholesterol (mg/dL)")
```

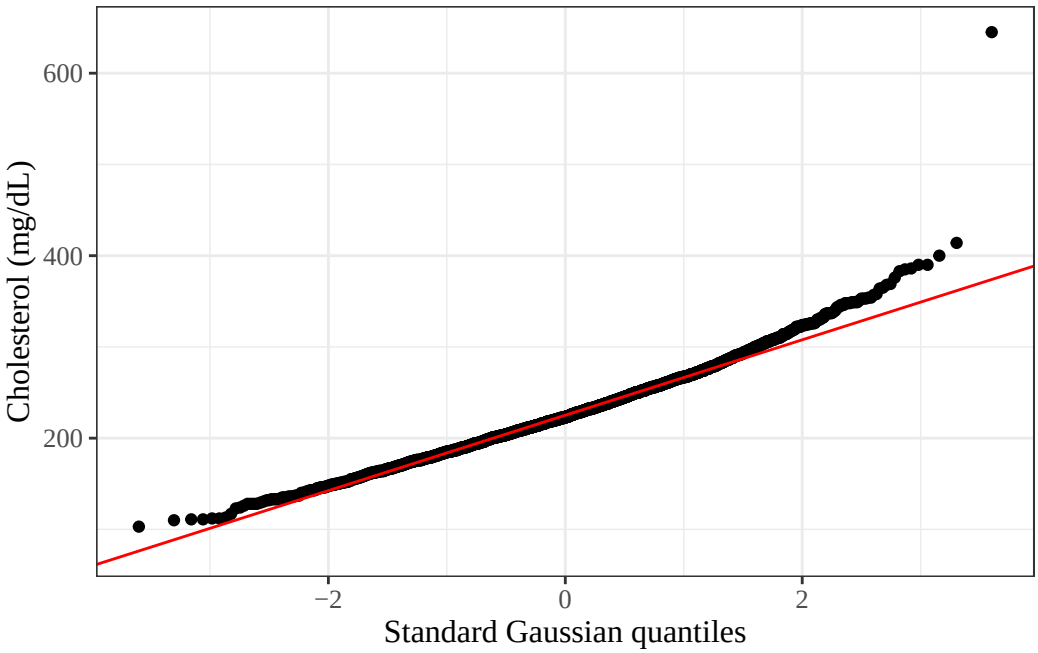


Figure 5: Normal Q-Q plot for total cholesterol in the WCGS dataset

In Figure 5, the points curve above the reference line at the right, which matches the long right tail in Figure 1.

Read

- Hogg et al. [10, Sec. 6.2], pp. 245–255 covers exploratory data analysis.
- Rice [11, Sec. 10.2.3], pp. 385–388 covers quantile-quantile plots.
- Rice [11, Sec. 10.3], pp. 389–391 covers histograms, density curves and stem-and-leaf plots.
- Rice [11, Sec. 10.6], pp. 402–403 covers boxplots.

Bivariate relationships

Two continuous variables

Definition 0.1 (Scatter plot)**.** A **scatter plot** displays the joint distribution of two continuous variables by plotting each observation as a point, with one variable on each axis.
Source: Wikipedia contributors [17].

Definition 0.2 (Pearson correlation coefficient)**.** The **Pearson correlation coefficient** of n paired observations $(x_1, y_1), \dots, (x_n, y_n)$ is:

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

It is defined when neither variable is constant, so that both sums of squares are positive.
Source: Wikipedia contributors [18].

ℹ Remark (The Pearson correlation coefficient estimates the population correlation)

Remark 0.1 (The Pearson correlation coefficient estimates the population correlation). The Pearson correlation coefficient is a sample estimate of the population correlation.

Theorem 0.1 (Range of the correlation coefficient)**.** *The Pearson correlation coefficient satisfies $-1 \le r \le 1$. Moreover, $r = 1$ if and only if the points (x_i, y_i) lie on a straight line with positive slope, and $r = -1$ if and only if they lie on a straight line with negative slope.*

ℹ Proof

Proof. Write $a_i \stackrel{\text{def}}{=} x_i - \bar{x}$ and $b_i \stackrel{\text{def}}{=} y_i - \bar{y}$, so that $r = \sum_{i=1}^{n} a_i b_i / \left(\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}\right)$. The Cauchy–Schwarz inequality states that $\left|\sum_{i=1}^{n} a_i b_i\right| \le \sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}$, with equality if and only if one of the vectors $(a_1, \dots, a_n)$ and $(b_1, \dots, b_n)$ is a scalar multiple of the other. Dividing both sides by the (positive) right-hand side gives $|r| \le 1$. Equality $|r| = 1$ therefore holds if and only if $b_i = c\, a_i$ for some constant $c \neq 0$ and every i, that is, $y_i = \bar{y} + c(x_i - \bar{x})$: the points lie on a line with slope c. Then $\sum_{i=1}^{n} a_i b_i = c \sum_{i=1}^{n} a_i^2$ has the sign of c, so $r = 1$ when $c > 0$ and $r = -1$ when $c < 0$. □

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = sbp, y = chol) +
  ggplot2::geom_point(alpha = 0.3) +
  ggplot2::geom_smooth(
    method = "lm",
    formula = y ~ x,
    se = TRUE,
    color = "red"
  ) +
  ggplot2::labs(x = "Systolic BP (mmHg)", y = "Cholesterol (mg/dL)")
```

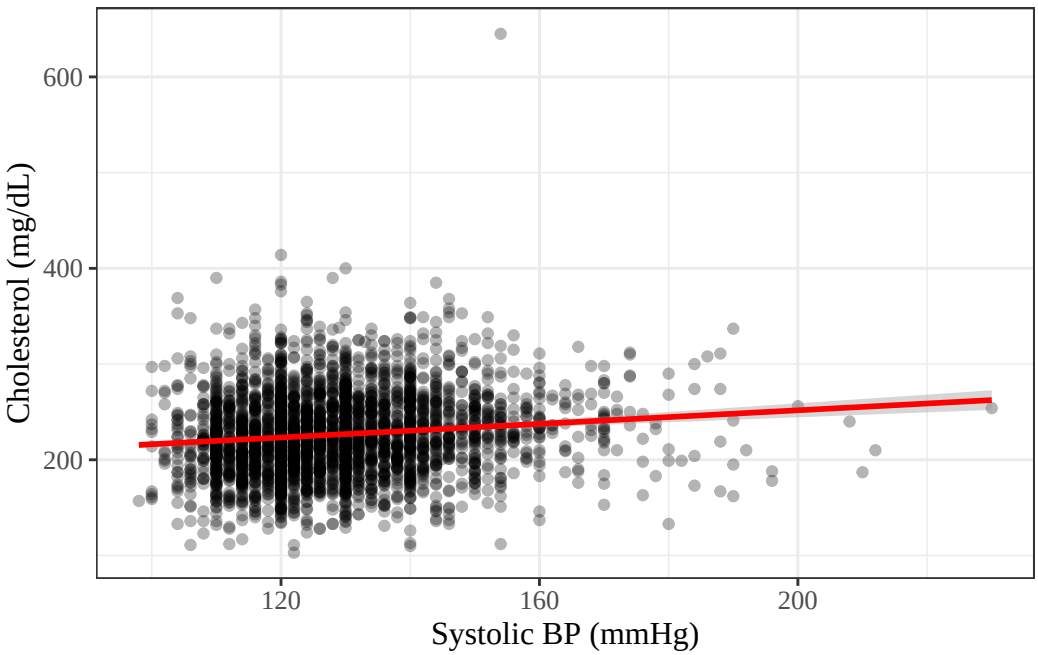


Figure 6: Cholesterol versus systolic blood pressure in the WCGS dataset, with a least-squares line

Example 0.1 (Correlation between cholesterol and blood pressure in the WCGS)**.**

```
cor(wcgs$chol, wcgs$sbp, use = "complete.obs")
#> [1] 0.123061
```

The correlation is positive but small: cholesterol tends to be slightly higher in participants with higher systolic blood pressure, consistent with the shallow slope and wide scatter in Figure 6.

Correlation measures linear association only

Two variables can be strongly related and still have a correlation near zero, if the relationship is not linear. For example, if $y_i = x_i^2$, the x_i are symmetric around 0, and the $|x_i|$ are not all equal, then $r = 0$, although y is a function of x. Correlation also does not imply causation.

Read

Rice [11, Sec. 10.7], pp. 404–406 covers scatterplots.

A continuous variable by a categorical variable

Side-by-side box plots compare the distribution of a continuous variable across the groups defined by a categorical variable (Figure 7 and Figure 8).

Table 3 computes summary statistics separately for each group.

Two categorical variables

Definition 0.3 (Contingency table (cross-tabulation))**.** A **contingency table** (or **cross-tabulation**) displays the joint frequencies of two categorical variables: each cell counts the observations with one combination of categories. For two binary variables, the contingency table is a 2×2 table with cells a, b, c, and d:

	Outcome $= 1$	Outcome $= 0$	Total
Exposure $= 1$	a	b	$a + b$
Exposure $= 0$	c	d	$c + d$
Total	$a + c$	$b + d$	n

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = smoke, y = chol, fill = smoke) +
  ggplot2::geom_boxplot() +
  ggplot2::labs(x = "Current smoker", y = "Cholesterol (mg/dL)") +
  ggplot2::theme(legend.position = "none")
```

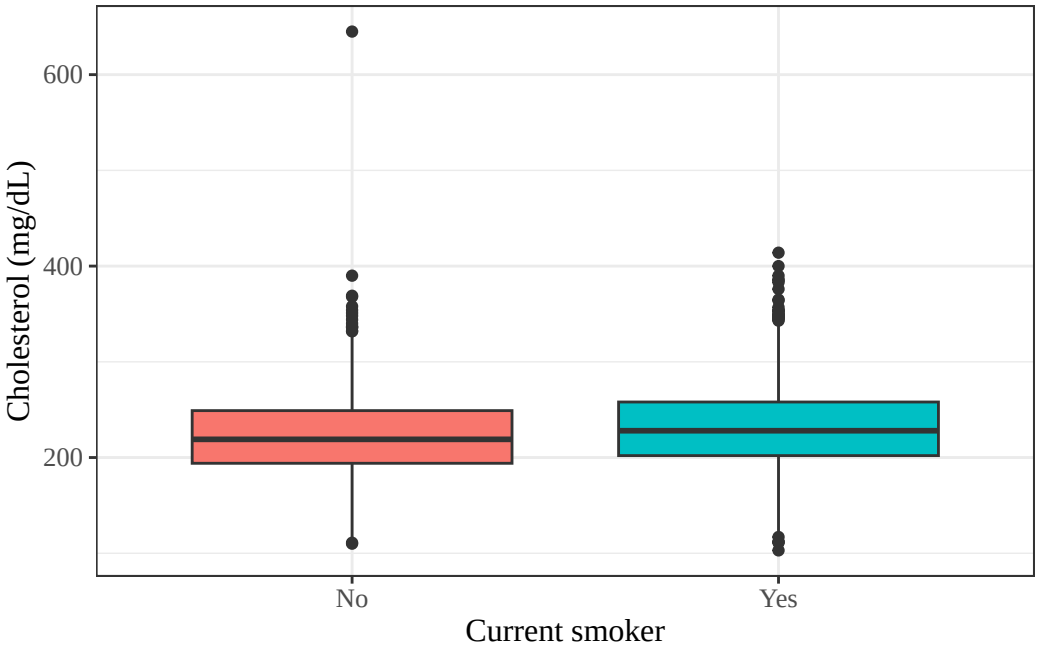


Figure 7: Box plots of cholesterol by smoking status in the WCGS dataset

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = behpat, y = chol, fill = behpat) +
  ggplot2::geom_boxplot() +
  ggplot2::labs(x = "Behavioral pattern", y = "Cholesterol (mg/dL)") +
  ggplot2::theme(legend.position = "none")
```

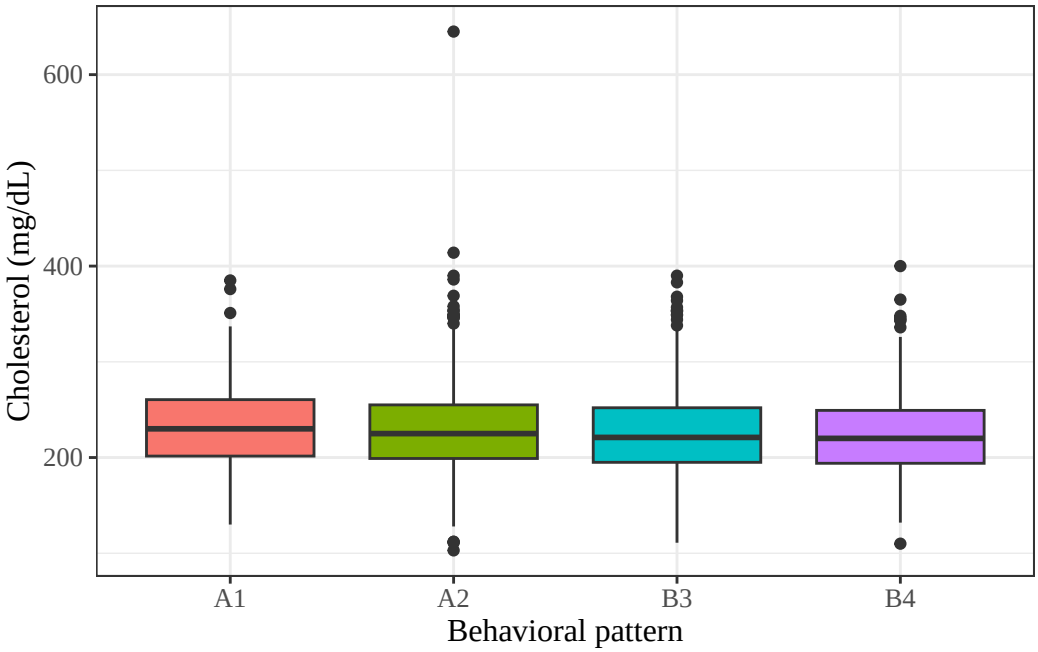


Figure 8: Box plots of cholesterol by behavioral pattern in the WCGS dataset

Table 3: Cholesterol, systolic blood pressure, and BMI by CHD status in the WCGS

```
wcgs |>
  dplyr::select(chol, sbp, bmi, chd69) |>
  gtsummary::tbl_summary(
    by = chd69,
    statistic = list(
      gtsummary::all_continuous() ~
        "{mean} ({sd})"
    ),
    digits = gtsummary::all_continuous() ~ 1
  ) |>
  gtsummary::add_overall()
```

Characteristic	**Overall** N = 3,154[1]	**No** N = 2,897[1]	**Yes** N = 257[1]
Cholesterol (mg/dL)	226.4 (43.4)	224.3 (42.2)	250.1 (49.4)
Unknown	12	12	0
Systolic BP (mmHg)	128.6 (15.1)	128.0 (14.7)	135.4 (17.5)
BMI (kg/m^2)	24.5 (2.6)	24.5 (2.6)	25.1 (2.6)

[1]Mean (SD)

Source: Wikipedia contributors [19].

Example 0.2 (Smoking and CHD in the WCGS)**.**

```
table(Smoking = wcgs$smoke, CHD = wcgs$chd69)
#>         CHD
#> Smoking   No  Yes
#>     No  1554   98
#>     Yes 1343  159
```

Row proportions give the distribution of CHD status within each smoking group:

```
table(Smoking = wcgs$smoke, CHD = wcgs$chd69) |>
  prop.table(margin = 1)
#>         CHD
#> Smoking       No      Yes
#>     No  0.940678 0.059322
#>     Yes 0.894141 0.105859
```

Table 4 shows the same counts as a formatted table.

Table 4: Smoking status and CHD event in the WCGS

```
wcgs |>
  dplyr::select(smoke, chd69) |>
  gtsummary::tbl_summary(by = chd69) |>
  gtsummary::add_overall()
```

Characteristic	**Overall** N = 3,154[1]	**No** N = 2,897[1]	**Yes** N = 257[1]
Current smoker	1,502 (48%)	1,343 (46%)	159 (62%)

[1] n (%)

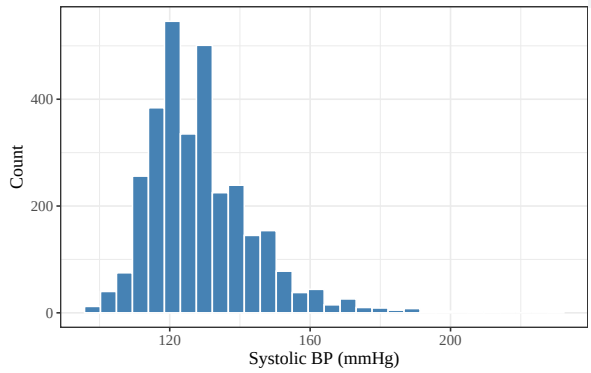
Data transformations

When a continuous variable has a right-skewed distribution (a longer tail to the right than to the left), a logarithmic transformation often makes the distribution more symmetric. A log-transformed variable also has a multiplicative interpretation: an increase of 1 in $\log(x)$ corresponds to multiplying x by $e \approx 2.72$, because $\log(x) + 1 = \log(e \cdot x)$.

The WCGS dataset already contains log-transformed versions of two variables:

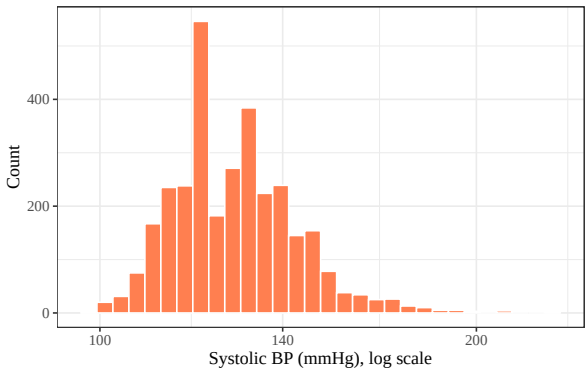
- `lnsbp`: log(SBP);
- `lnwght`: log(weight).

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = sbp) +
  ggplot2::geom_histogram(bins = 30, f
  ggplot2::labs(x = "Systolic BP (mmHg
```



(a) Raw scale

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = sbp) +
  ggplot2::geom_histogram(bins = 30, fill = "coral", c
  ggplot2::scale_x_log10(breaks = c(100, 140, 200)) +
  ggplot2::labs(x = "Systolic BP (mmHg), log scale", y
```



(b) Log scale

Figure 9: Distribution of systolic blood pressure in the WCGS, on two scales. Both panels label the axis in mmHg; the right panel spaces the values logarithmically, so its histogram is the histogram of `lnsbp`.

```
# sample skewness: mean cubed deviation, divided by the cubed SD
skewness <- function(x) {
  x <- x[!is.na(x)]
  mean((x - mean(x))^3) / sd(x)^3
}
sbp_skew <- c(
  raw = skewness(wcgs$sbp),
  log = skewness(wcgs$lnsbp)
)
sbp_skew
#>      raw      log
#> 1.203824 0.739911
```

The log-transformed SBP is less skewed than the raw SBP (sample skewness 0.74 versus 1.2),

but still not symmetric. Whether to transform a variable in a regression model depends on the assumptions of that model and on the scientific question.

Rescaling a variable

Exercise 0.1 (A standardized cholesterol value)**.** In the WCGS data, cholesterol (`chol`) is measured in mg/dL.

1. Compute the sample mean and sample standard deviation of `chol`.
2. A man has `chol` equal to 250 mg/dL. By how many standard deviations does his value differ from the sample mean, and in which direction?
3. What are the units of your answer to part 2?

Solution

Solution 0.1.

```
chol_mean <- mean(wcgs$chol, na.rm = TRUE)
chol_sd <- sd(wcgs$chol, na.rm = TRUE)
chol_z <- (250 - chol_mean) / chol_sd
c(mean = chol_mean, sd = chol_sd, z_250 = chol_z)
#>       mean         sd      z_250
#> 226.372374  43.420426   0.544159
```

1. The sample mean is 226.4 mg/dL and the sample standard deviation is 43.4 mg/dL.
2. $(250 - \bar{x})/s$ is 0.54: his cholesterol is 0.54 standard deviations above the sample mean.
3. The numerator is in mg/dL and the denominator is in mg/dL, so the answer has no units.

Definition 0.1 (Standardization (z-scoring, standard scaling))**.** Let $x_1, \ldots, x_n$ be $n \geq 2$ observations of a variable with sample mean $\bar{x}$ and sample standard deviation $s > 0$. The **standardized value** of x_i is

$$z_i \stackrel{\text{def}}{=} \frac{x_i - \bar{x}}{s}$$

Standardization is also called **z-scoring** or **standard scaling**.
Source: Wikipedia contributors [20].

Remark (Source: Standardization)

Remark 0.1 (Source: Standardization). James et al. [21, Sec. 4.7.6], p. 183. Géron [22, Ch. 2] describes the same operation but says it divides by the variance; the standard deviation is used here, as in James et al. [21].

Remark (Standardized values have no units)

Remark 0.2 (Standardized values have no units). In Exercise 0.1, the standardized value of a cholesterol of 250 mg/dL is the number of sample standard deviations between 250 and the sample mean. Standardizing puts variables measured in different units on a comparable scale. For example, a salary in dollars and an age in years both become unitless numbers of standard deviations.

Theorem 0.1 (Standardized values have sample mean 0 and sample standard deviation 1)**.** *For $n \geq 2$ observations with $s > 0$, the standardized values $z_1, \ldots, z_n$ have sample mean 0 and sample standard deviation 1.*

Proof

Proof. The sample mean of the z_i is

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s} &= \frac{1}{s}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}\right) \\ &= \frac{1}{s}(\bar{x} - \bar{x}) \\ &= 0. \end{aligned}$$

Since the sample mean of the z_i is 0, their sample variance is

$$\begin{aligned} \frac{1}{n-1}\sum_{i=1}^{n} z_i^2 &= \frac{1}{s^2}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{s^2}{s^2} \\ &= 1, \end{aligned}$$

so their sample standard deviation is $\sqrt{1} = 1$. □

Remark (Source: Standardized values have sample mean 0 and sample standard deviation 1)

Remark 0.3 (Source: Standardized values have sample mean 0 and sample standard deviation 1). Stated for the standardized Caravan variables in James et al. [21, Sec. 4.7.6], p. 183; the proof is the elementary calculation.

Example 0.1 (Standardizing cholesterol in R)**.** In R, `scale()` returns the standardized values of a column, as a one-column matrix. For the WCGS cholesterol values:

```
z <- scale(wcgs$chol)[, 1]
mean(z, na.rm = TRUE)
#> [1] 2.79926e-16
sd(z, na.rm = TRUE)
#> [1] 1
```

The sample mean is 0 up to rounding error, and the sample standard deviation is 1, as Theorem 0.1 says. The `na.rm = TRUE` is needed because 12 of the cholesterol values are missing.

Definition 0.2 (Min-max scaling)**.** Let $x_1, \ldots, x_n$ be observations of a variable with minimum $x_{\min}$ and maximum $x_{\max} > x_{\min}$. The **min-max scaled value** of x_i is

$$\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

Source: Géron [22, Ch. 2]; Wikipedia contributors [23].

Remark (Min-max scaling and outliers)

Remark 0.4 (Min-max scaling and outliers). Every min-max scaled value lies in $[0, 1]$: the minimum maps to 0 and the maximum maps to 1. For example, if median incomes range from 0 to 15, an income of 0 maps to 0, and an income of 15 maps to 1.
Min-max scaling is sensitive to outliers. For example, suppose one income of 100 is recorded by mistake. The maximum is then 100, so min-max scaling maps all the other incomes into $[0, 0.15]$. Standardization changes them much less [22, Ch. 2], "Feature Scaling".

Example 0.2 (Min-max scaling of cholesterol in the WCGS)**.**

```r
chol_range <- range(wcgs$chol, na.rm = TRUE)
chol_minmax <- (wcgs$chol - chol_range[1]) / diff(chol_range)
range(chol_minmax, na.rm = TRUE)
#> [1] 0 1
```

The smallest cholesterol value maps to 0 and the largest maps to 1. The scaled values keep the shape of the distribution but have no units.

An exploratory data analysis workflow

A typical exploratory data analysis proceeds as follows:

1. **Understand the dataset**: the number of observations and variables, and each variable's type.
2. **Examine each variable individually**:
   - continuous: histogram, box plot, mean, SD, median, IQR;
   - categorical: bar chart, frequency table;
   - note missing values, unusual values, and outliers.
3. **Examine relationships between variables**:
   - two continuous variables: scatter plot, correlation;
   - continuous by categorical: side-by-side box plots, group means;
   - two categorical variables: contingency table.
4. **Consider transformations** for skewed continuous variables.
5. **Summarize the findings** to guide model-building decisions.

Table 5 summarizes selected WCGS variables in one table.

Further reading

Plotting a Binary Outcome shows how to plot a 0/1 outcome against a continuous feature, where scatterplot dots would pile into two stripes. It draws a rug and two histograms of the WCGS data, in R and Python.

Table 5: Summary of selected WCGS variables

```
wcgs |>
  dplyr::select(
    age, chol, sbp, dbp, bmi, weight, ncigs,
    chd69, smoke, dibpat, behpat
  ) |>
  gtsummary::tbl_summary()
```

Characteristic	**N = 3,154**[1]
Age (years)	45.0 (42.0, 50.0)
Cholesterol (mg/dL)	223 (197, 253)
Unknown	12
Systolic BP (mmHg)	126 (120, 136)
Diastolic BP (mmHg)	80 (76, 86)
BMI (kg/m^2)	24.39 (22.96, 25.84)
Weight (lbs)	170 (155, 182)
Cigarettes per day	0 (0, 20)
CHD event by 1969	257 (8.1%)
Current smoker	1,502 (48%)
Behavioral pattern (A/B)	
Type B	1,565 (50%)
Type A	1,589 (50%)
Behavioral pattern (A1/A2/B3/B4)	
A1	264 (8.4%)
A2	1,325 (42%)
B3	1,216 (39%)
B4	349 (11%)

[1]Median (Q1, Q3); n (%)

References

[1] E. Vittinghoff, D. V. Glidden, S. C. Shiboski, and C. E. McCulloch, *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models*, 2nd ed. Springer, 2012. doi: 10.1007/978-1-4614-1353-0.

[2] R. H. Rosenman, R. J. Brand, C. D. Jenkins, M. Friedman, R. Straus, and M. Wurm, "Coronary heart disease in the Western Collaborative Group Study: Final follow-up experience of 8 1/2 years," *JAMA*, vol. 233, no. 8, pp. 872–877, 1975, doi: 10.1001/jama.1975.03260080034016.

[3] Wikipedia contributors, "Statistical population — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Statistical_population

[4] Wikipedia contributors, "Simple random sample — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Simple_random_sample

[5] Wikipedia contributors, "Median — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Median

[6] Wikipedia contributors, "Variance — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Variance

[7] Wikipedia contributors, "Standard deviation — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Standard_deviation

[8] Wikipedia contributors, "Quartile — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Quartile

[9] Wikipedia contributors, "Interquartile range — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Interquartile_range

[10] R. V. Hogg, E. A. Tanis, and D. L. Zimmerman, *Probability and statistical inference*, Tenth edition. Boston: Pearson, 2019.

[11] J. A. Rice, *Mathematical statistics and data analysis*, 3rd ed. Belmont, CA: Cengage Learning, 2007.

[12] Wikipedia contributors, "Histogram — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Histogram

[13] Wikipedia contributors, "Kernel density estimation — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Kernel_density_estimation

[14] Wikipedia contributors, "Box plot — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Box_plot

[15] Wikipedia contributors, "Bar chart — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Bar_chart

[16] Wikipedia contributors, "Q–q plot — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

[17] Wikipedia contributors, "Scatter plot — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Scatter_plot

[18] Wikipedia contributors, "Pearson correlation coefficient — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[19] Wikipedia contributors, "Contingency table — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Contingency_table

[20] Wikipedia contributors, "Standard score — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Standard_score

[21] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: With applications in R*, 2nd ed. Springer, 2021. doi: 10.1007/978-1-0716-1418-1.

[22] A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow*, 1st ed. O'Reilly Media, 2017.

[23] Wikipedia contributors, "Feature scaling — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Feature_scaling