

# STATS8: Introduction to Biostatistics

## Data Exploration

Babak Shahbaba  
Department of Statistics, UCI

# Introduction

- After clearly defining the scientific problem, selecting a set of representative members from the population of interest, and collecting data (either through passively observing events or experiments), we usually begin our analysis with data exploration.
- We start by focusing on data exploration techniques for one variable at a time.
- Our objective is to develop a high-level understanding of the data, learn about the possible values for each characteristic, and find out how a characteristic varies among individuals in our sample.
- In short, we want to learn about the **distribution** of variables.

## Variable types

- The visualization techniques and summary statistics we use for a variable depend on its type.
- Based on the values a variable can take, we can classify them into two general groups: **numerical** variables and **categorical** variables.
- Consider the `Pima.tr` data available from the MASS library.
- Variables `npreg`, `age`, and `bmi` in this data set are numerical variables since they take numerical values, and the numbers they take have their usual meaning.
- The `type` variable in this data set, on the other hand, is categorical since the set of values it can take consists of a finite number of categories.

## Variable types

- Some numerical variables are **count** variables. For example, number of pregnancies and number of physician visits.
- For categorical variables, we typically use numerical codings.
- Categorical variables are either **ordinal** or **nominal** depending on the extent of information the numerical coding provides.
- For nominal variables, such as type, the numbers are simply labels, which are chosen arbitrarily.
- For ordinal variables, such as disease severity, although the numbers do not have their usual meaning, they preserve a rank ordering.

## Frequency and relative frequency

- The number of times a specific category is observed is called **frequency**. We denote the frequency for category  $c$  by  $n_c$ .
- The relative frequency is the sample proportion for each possible category. It is obtained by dividing the frequencies  $n_c$  by the total number of observations  $n$ :

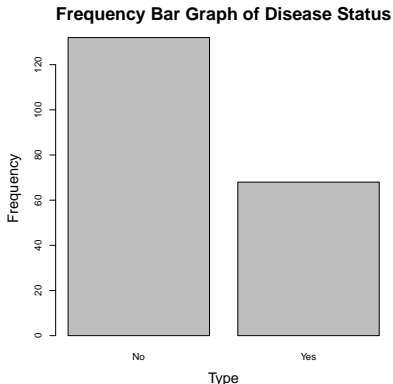
$$p_c = \frac{n_c}{n}$$

- Relative frequencies are sometimes presented as percentages after multiplying proportions  $p_c$  by 100.
- For a categorical variable, the mode of is the most common value, i.e., the value with the highest frequency.

## Bar graph

- For categorical variables, **bar graphs** are one of the simplest ways of visualizing the data.
- Using a bar graph, we can visualize the possible values (categories) a categorical variable can take, as well as the number of times each category has been observed in our sample.
- The height of each bar in this graph shows the number of times the corresponding category has been observed.

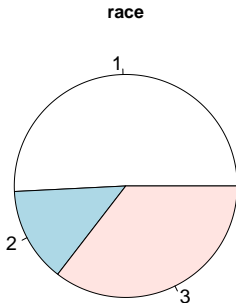
## Bar graphs and frequencies



**Figure:** Using R-Commander to create and view a frequency bar graph for `type` in the `Pima.tr` data set. The heights of the bars sum to the sample size  $n$ . Overall, bar graphs show us how the observed values of a categorical variable in our sample are distributed

## Pie chart

- We can use a pie chart to visualize the relative frequencies of different categories for a categorical variable.
- In a pie chart, the area of a circle is divided into sectors, each representing one of the possible categories of the variable.
- The area of each sector  $c$  is proportional to its frequency.

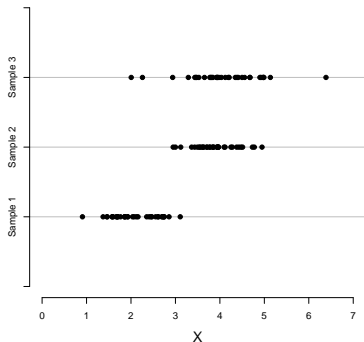




## Exploring Numerical Variables

- For numerical variables, we are especially interested in two key aspects of the distribution: its **location** and its **spread**.
- The location of a distribution refers to the *central tendency* of values, that is, the point around which most values are gathered.
- The spread of a distribution refers to the *dispersion* of possible values, that is, how scattered the values are around the location.

# Exploring Numerical Variables

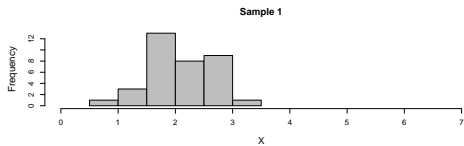
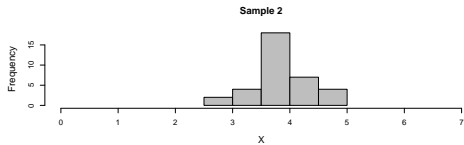
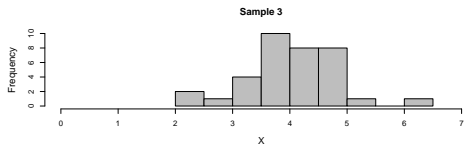


**Figure:** Three separate samples for variable  $X$ . Observations in Sample 1 are gathered around 2, whereas observations in Sample 2 and Sample 3 are gathered around 4. Observations in Sample 3 are more dispersed compared to those in Sample 1 and Sample 2

# Histograms

- **Histograms** are commonly used to visualize numerical variables.
- A histogram is similar to a bar graph after the values of the variable are grouped (binned) into a finite number of intervals (bins).
- For each interval, the bar height corresponds to the frequency (count) of observation in that interval.

# Histograms



# Histograms

- The bar height for each interval could be set to its relative frequency  $p_c = n_c/n$ , or the percentage  $p_c \times 100$ , of observations that fall into that interval.
- For histograms, however, it is more common to use the **density** instead of the relative frequency or percentage.
- The density is the relative frequency for a unit interval. It is obtained by dividing the relative frequency by the interval width:

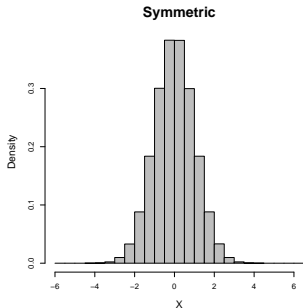
$$f_c = \frac{p_c}{w_c}.$$

Here,  $p_c = n_c/n$  is the relative frequency with  $n_c$  as the frequency of interval  $c$  and  $n$  as the total sample size.

- The width of interval  $c$  is denoted  $w_c$ .

## Shapes of histograms

- Besides the location and spread of a distribution, the shape of a histogram also shows us how the observed values spread around the location.
- We say the following histogram is **symmetric** around its location (here, zero) since the densities are the [almost] same for any two intervals that are equally distant from the center.

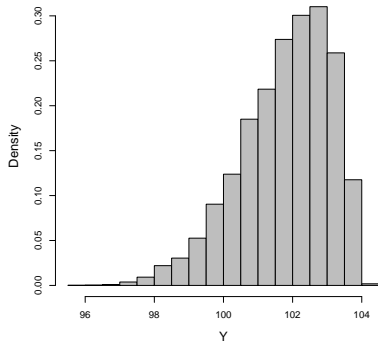


## Skewed histograms

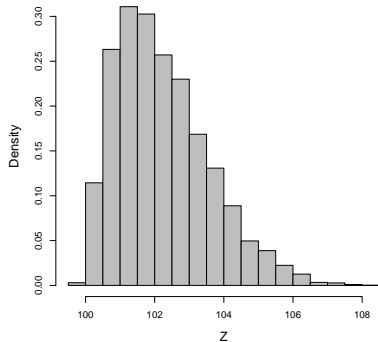
- In many situations, we find that a histogram is stretched to the left or right.
- We call such histograms **skewed**.
- More specifically, we call them **left-skewed** if they are stretched to the left, or **right-skewed** if they are stretched to the right.

# Skewed histograms

## Left-Skewed



## Right-Skewed

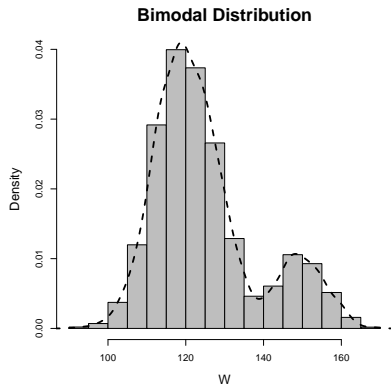




## Unimodal vs. bimodal

- The above histograms, whether symmetric or skewed, have one thing in common: they all have one *peak* (or mode).
- We call such histograms (and their corresponding distributions) **unimodal**.
- Sometimes histograms have multiple modes.
- The bimodal histogram appears to be a combination of two unimodal histograms.
- Indeed, in many situations bimodal histograms (and multimodal histograms in general) indicate that the underlying population is not *homogeneous* and may include two (or more in case of multimodal histograms) subpopulations.

# Unimodal vs. bimodal



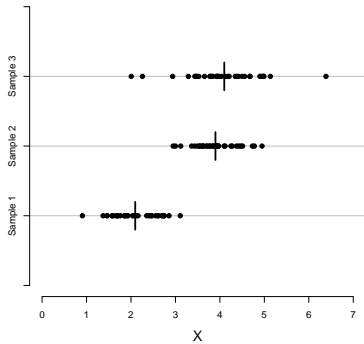
## Sample mean

- Histograms are useful for visualizing numerical data and identifying their location and spread. However, we typically use summary statistics for more precise specification of the central tendency and dispersion of observed values.
- A common summary statistic for location is the **sample mean**.
- The sample mean is simply the average of the observed values. For observed values  $x_1, \dots, x_n$ , we denote the sample mean as  $\bar{x}$  and calculate it by

$$\bar{x} = \frac{\sum_i x_i}{n},$$

where  $x_i$  is the  $i$ th observed value of  $X$ , and  $n$  is the sample size.

# Sample mean



## Sample mean

- Sample mean is sensitive to very large or very small values, which might be outliers (unusual values).
- For instance, suppose that we have measured the resting heart rate (in beats per minute) for five people.

$$x = \{74, 80, 79, 85, 81\}, \quad \bar{x} = \frac{74 + 80 + 79 + 85 + 81}{5} = 79.8.$$

- In this case, the sample mean is 79.8, which seems to be a good representative of the data.
- Now suppose that the heart rate for the first individual is recorded as 47 instead of 74.

$$x = \{47, 80, 79, 85, 81\}, \quad \bar{x} = \frac{47 + 80 + 79 + 85 + 81}{5} = 74.4.$$

- Now, the sample mean does not capture the central tendency.

## Sample median

- The **sample median** is an alternative measure of location, which is less sensitive to outliers.
- For observed values  $x_1, \dots, x_n$ , the median is denoted  $\tilde{x}$  and is calculated by first sorting the observed values (i.e., ordering them from the lowest to the highest value) and selecting the middle one.
- If the sample size  $n$  is odd, the median is the number at the middle of the sorted observations. If the sample size is even, the median is the average of the two middle numbers.
- The sample medians for the above two scenarios are

$$x = \{74, 79, 80, 81, 85\}, \quad \tilde{x} = 80;$$

$$x = \{47, 79, 80, 81, 85\}, \quad \tilde{x} = 80.$$

## Variance and standard deviation

- While summary statistics such as mean and median provide insights into the central tendency of values for a variable, they are rarely enough to fully describe a distribution.
- We need other summary statistics that capture the dispersion of the distribution.
- Consider the following measurements of blood pressure (in mmHg) for two patients:

$$\text{Patient A: } x = \{95, 98, 96, 95, 96\}, \quad \bar{x} = 96, \quad \tilde{x} = 96.$$

$$\text{Patient B: } y = \{85, 106, 88, 105, 96\}, \quad \bar{y} = 96, \quad \tilde{y} = 96.$$

- While the mean and median for both patients are 96, the readings are more dispersed for Patient B.

## Variance and standard deviation

- Two common summary statistics for measuring dispersion are the **sample variance** and **sample standard deviation**.
- These two summary statistics are based on the **deviation** of observed values from the mean as the center of the distribution.
- For each observation, the deviation from the mean is calculated as  $x_j - \bar{x}$ .



## Variance and standard deviation

- The sample variance is a common measure of dispersion based on the squared deviations

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

- The square root of the variance is called the sample standard deviation

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

## Variance and standard deviation

Patient A			Patient B		
$x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
95	-1	1	85	-11	121
98	2	4	106	10	100
96	0	0	88	-8	65
95	-1	1	105	9	81
96	0	0	96	0	0
$\Sigma$	0	6	$\Sigma$	0	366

$s^2 = 6/4 = 1.5$	$s^2 = 366/4 = 91.5$
$s = \sqrt{1.5} = 1.22$	$s = \sqrt{91.5} = 9.56$

## Quantiles

- Informally, the sample median could be interpreted as the point that divides the ordered values of the variable into two equal parts.
- That is, the median is the point that is greater than or equal to at least half of the values and smaller than or equal to at least half of the values.
- The median is called the 0.5 **quantile**.
- Similarly, the 0.25 quantile is the point that is greater than or equal to at least 25% of the values and smaller than or equal to at least 75% of the values.
- In general, the  $q$  quantile is the point that is greater than or equal to at least  $100q\%$  of the values and smaller than or equal to at least  $100(1 - q)\%$  of the values.

## Quartiles

- We can divide the ordered values of a variable into four equal parts using 0.25, 0.5, and 0.75 quantiles.
- The corresponding points are denoted  $Q_1$ ,  $Q_2$ , and  $Q_3$ , respectively.
- We refer to these three points as **quartiles**, of which  $Q_1$  is called the *first quartile* or the *lower quartile*,  $Q_2$  (i.e., median) is called the *second quartile*, and  $Q_3$  is called the *third quartile* or *upper quartile*.
- The interval from  $Q_1$  (0.25 quantile) to  $Q_3$  (0.75 quantile) covers the middle 50% of the ordered data.

## Five-number summary and boxplot

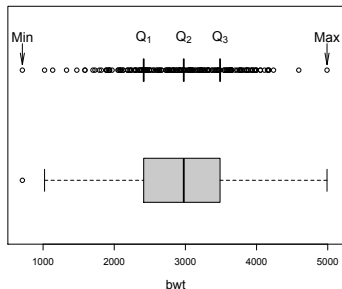
- The **minimum** (min), which is the smallest value of the variable in our sample, is in fact the 0 quantile.
- On the other hand, the **maximum** (max), which is the largest value of the variable in our sample, is the 1 quantile.
- The minimum and maximum along with quartiles ( $Q_1$ ,  $Q_2$ , and  $Q_3$ ) are known as **five-number summary**.
- These are usually presented in the increasing order: min, first quartile, median, third quartile, max.
- This way, the five-number summary provides 0, 0.25, 0.50, 0.75, and 1 quantiles.

## Five-number summary and boxplot

- The five-number summary can be used to derive two measures of dispersion: the **range** and the **interquartile range**.
- The range is the difference between the maximum observed value and the minimum observed value.
- The interquartile range (IQR) is the difference between the third quartile ( $Q_3$ ) and the first quartile ( $Q_1$ ).

## Five-number summary and boxplot

- To visualize the five-number summary, the range and the IQR, we often use a **boxplot** (a.k.a. **box and whisker** plot).



- Very often, boxplots are drawn vertically.

## Five-number summary and boxplot

- The thick line at the middle of the “box” shows the median.
- The left side of the box shows the lower quartile.
- Likewise, the right side of the box is the upper quartile.
- The dashed lines are known as the **whiskers**.
- The whisker on the right of the box extends to the largest observed value or  $Q_3 + 1.5 \times \text{IQR}$ , whichever it reaches first.
- The whisker on the left extends to the lowest value or  $Q_1 - 1.5 \times \text{IQR}$ , whichever it reaches first.
- Data points beyond the whiskers are shown as circles and considered as possible outliers.

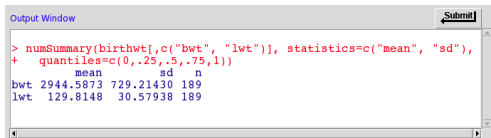


## Data preprocessing

- Data we collect for scientific studies are rarely ready for analysis; they often require **preprocessing**.
- This typically involve
  - handling missing information
  - identifying outliers and possibly removing them (ONLY WHEN THEY ARE DEEMED TO BE DATA ENTRY MISTAKES)
  - data transformation
  - creating new variables based on the existing ones

## Coefficient of variation

- Suppose that we want to compare the dispersion of bwt to that of lwt using their standard deviations based on the birthwt data.



```
> numSummary(birthwt[,c("bwt", "lwt")], statistics=c("mean", "sd"),
+ quantiles=c(0,.25,.5,.75,1))
      mean      sd      n
bwt 2944.5873 729.21430 189
lwt  129.8148  30.57938 189
```

- It seems that bwt is more dispersed than lwt since it has higher standard deviation compared to lwt.
- However, the two variables are not comparable; they have different units.

## Coefficient of variation

- In many situations, we can avoid these issues by using another measure of variation called the **coefficient of variation** instead of standard deviation.
- To quantify dispersion independently from units, we use the coefficient of variation, which is the standard deviation divided by the sample mean (assuming that the mean is a positive number):

$$CV = \frac{s}{\bar{x}}$$

- The coefficient of variation for bwt (birth weight in grams) is  $729.2/2944.6 = 0.25$  and for bwt.1b (birth weight in pounds) is  $1.6/6.5 = 0.25$ .

## Scaling and shifting variables

- Why the coefficient of variation ( $CV = s/\bar{x}$ ) is independent of measurement units in the above example?
- In general, when we multiply the observed values of a variable by a constant  $a$ , its mean, standard deviation, and variance are multiplied by  $a$ ,  $|a|$ , and  $a^2$ , respectively.
- That is, if  $y = ax$ , then

$$\begin{aligned}\bar{y} &= a\bar{x}, \\ s_y &= |a|s_x, \\ s_y^2 &= a^2s_x^2,\end{aligned}$$

- Therefore,

$$CV_y = \frac{s_y}{\bar{y}} = \frac{as_x}{a\bar{x}} = \frac{s_x}{\bar{x}} = CV_x.$$

## Scaling and shifting variables

- If instead of scaling the observed value, we shift them by a constant  $b$ , the sample mean shifts by  $b$  units.
- However, since the difference between observed values and the mean do not change, the standard deviation and variance remain unchanged.
- In general, if we shift the observed values by  $b$ , i.e.,  $y = x + b$ , then

$$\bar{y} = \bar{x} + b,$$

$$s_y = s_x,$$

$$s_y^2 = s_x^2.$$

## Variable standardization

- **Variable standardization** is a common *linear* transformation, where we subtract the sample mean  $\bar{x}$  from the observed values and divide the result by the sample standard deviation  $s$ :

$$y_i = \frac{x_i - \bar{x}}{s}.$$

- Subtracting  $\bar{x}$  from the observations shifts the sample mean to zero.
- This, however, does not change the standard deviation.
- Dividing by  $s$ , on the other hand, changes the sample standard deviation to 1. by  $s$ .
- Therefore, variable standardization creates a new variable with mean 0 and standard deviation 1.