# Describing Distributions with Numbers

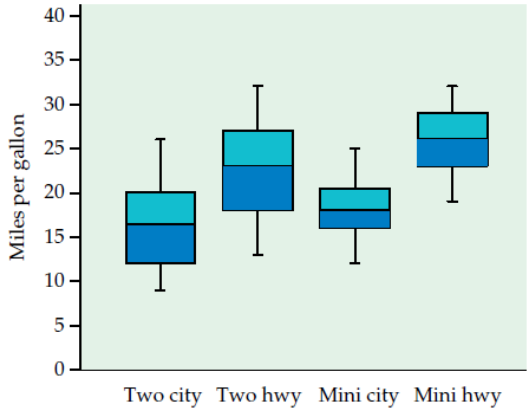
- A brief description of a distribution should include its shape and numbers describing its center and spread. We describe the shape of a distribution based on inspection of a histogram or a stemplot.

- Mean = the average value (*sensitive to the influence of a few extreme observations, not resistant)*
- Median = the middle value

  6 9 8 5 6 7 2 3 7 9 - order them - 2 3 5 6 6 7 7 8 9 9 35 – 7

- The simplest useful numerical description of a distribution consists of both a **measure of center** and a **measure of spread**.
- We can describe the spread or variability of a distribution by giving several percentiles.

The five-number summary leads to another visual representation of a distribution, the *boxplot*. Figure 1.19 shows boxplots for both city and highway gas mileages for our two groups of cars.




- The **interquartile range IQR** is the distance between the first and third quartiles.
- Call an observation a suspected **outlier** if it falls more than 20+ 1.5 × IQR = outlier above the third quartile or below the first quartile. 12 - 1.5 × IQR = outlier
- **The standard deviation** measures spread by looking at how far the observations are from their mean. How to calculate the standard deviation?
  1. First, compute the **variance $s^2$** of a set of observations, which is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations x1, x2, . . . , xn is:

     $$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

     The number n − 1 is called the **degrees of freedom** of the variance or standard deviation.
  2. The **standard deviation *s*** is the square root of the variance $s^2$

- s2 and s will be large if the observations are widely spread about their mean, and small if the observations are all close to the mean.

*Why do we square the deviations?*

- Squared deviations point to the mean as center in a way that distances do not.

*Why do we emphasize the standard deviation rather than the variance?*

- The standard deviation is the natural measure of spread for Normal distributions.
- The standard deviation s measures spread about the mean in the original scale.

*Properties of the standard deviation*

- **s** measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- **s** = 0 only when there is no spread. This happens only when all observations have the same value. Otherwise, s > 0. As the observations become more spread out about their mean, s gets larger.
- **s**, like the mean x, is not resistant. A few outliers can make **s** very large.

*How to choose the measure of spread?*

- The **five-number summary** is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. In R, fivenum function.
- Use *mean* and *standard deviation* only for reasonably symmetric distributions that are free of outliers.

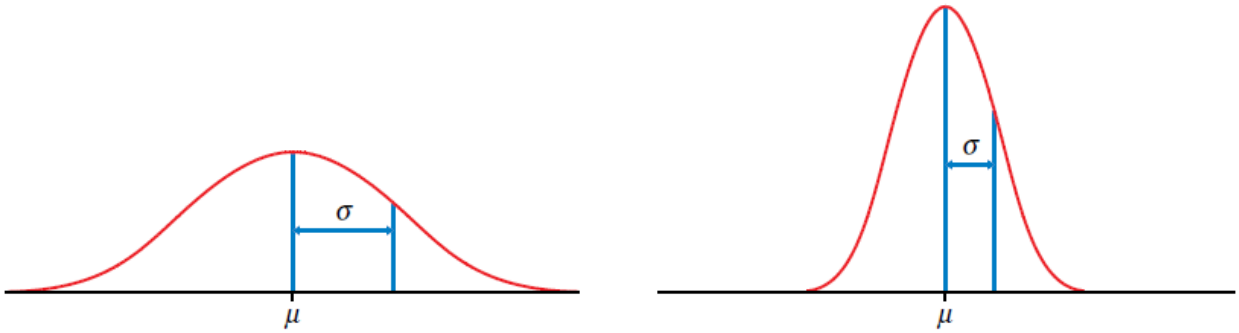
### LINEAR TRANSFORMATIONS

- **linear transformation** changes the original variable x into the new variable xnew given by an equation of the form: Xnew = a + bx.
  Adding the constant **a** shifts all values of **x** upward or downward by the same amount. In particular, such a shift changes the origin (zero point) of the variable. Multiplying by the positive constant **b** changes the size of the unit of measurement.
- Linear transformations do **not** change the shape of a distribution.

## 1.3 Density Curves and Normal Distributions

- **A density curve** is as a smooth approximation to the irregular bars of a histogram.
- A density curve is always on or above the horizontal axis and has area exactly **1** underneath it.
- A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range.

- **A mode** of a distribution described by a density curve is a peak point of the curve, the location where the curve is highest.
- $\bar{\mathbf{x}}$ and **s** are the mean and standard deviation of the actual data (sample).
- **μ** and **σ** are the idealized mean and standard deviation (population).



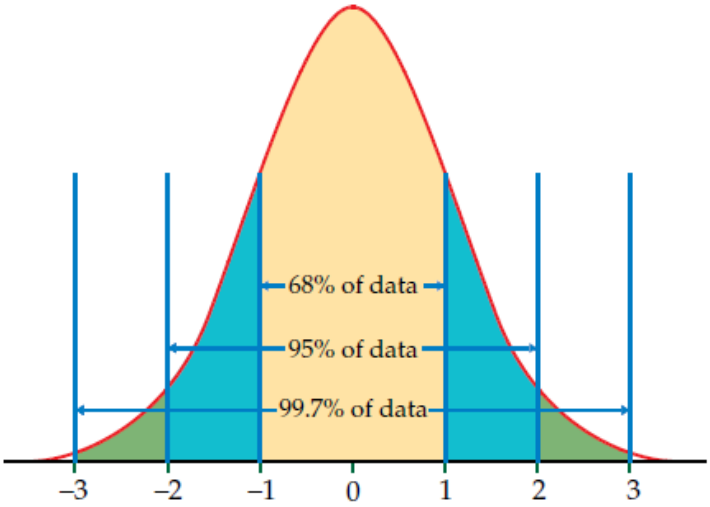

- **Normal curves** – the **μ** and **σ** alone specify the shape of the **normal distributions - N(μ,σ)**, and the shape of density curves in general reveals σ. These are special properties of Normal distributions.

*Why are the Normal distributions important in statistics?*

1. Normal distributions are good descriptions for some distributions of real data.
2. Normal distributions are good approximations to the results of many kinds of chance outcomes, such as tossing a coin many times.
3. Many statistical inference procedures based on Normal distributions work well for other roughly symmetric distributions.

## THE 68–95–99.7 RULE

- In the Normal distribution with mean **μ** and standard deviation **σ**:
    1. Approximately 68% of the observations fall within **σ** of the mean **μ**.
    2. Approximately 95% of the observations fall within **2σ** of **μ**.
    3. Approximately 99.7% of the observations fall within **3σ** of **μ**.




## STANDARDIZING AND *z*-SCORES

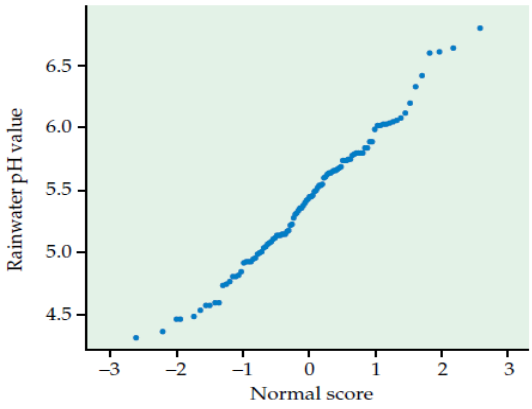
- All Normal distributions are the same if we measure in units of size σ about the mean μ as center. Changing to these units is called standardizing. To standardize a value, subtract the mean of the distribution and then divide by the standard deviation: $z = (x - \mu) / \sigma$ (*z-score*).
- **A z-score** tells us how many standard deviations the original observation falls away from the mean, and in which direction. Observations larger than the mean are positive when standardized, and observations smaller than the mean are negative.
- *Standardizing* is a linear transformation that transforms the data into the standard scale of z-scores. We know that a linear transformation does not change the shape of a distribution, and that the mean and standard deviation change in a simple manner. In particular, the standardized values for any distribution always have mean 0 and standard deviation 1.
- If the variable we standardize has a Normal distribution, standardizing does more than give a common scale. It makes all Normal distributions into a single distribution, and this distribution is still Normal. Standardizing a variable that has any Normal distribution produces a new variable that has the standard Normal distribution.

## THE STANDARD NORMAL DISTRIBUTION

- The standard Normal distribution is the Normal distribution N(0, 1) with mean 0 and standard deviation 1. If a variable X has any Normal distribution N(μ, σ) with mean μ and standard deviation σ, then the standardized variable $Z = (X - \mu) / \sigma$ has the standard Normal distribution.

## NORMAL QUANTILE PLOTS

- It is risky to assume that a distribution is Normal without actually inspecting the data. The most useful tool for assessing Normality is another graph, the **Normal quantile plot.** If the data distribution is close to any Normal distribution, the plotted points will lie close to a straight line.


Rainwater pH value
6.5
6.0
5.5
5.0
4.5
−3
−2
−1
0
1
2
3
Normal score

## CORRELATION & REGRESSION

*How to display a relationship between two quantitative variables?*

- With a **scatterplot -** it displays he form, direction, and strength of the relationship between two quantitative variables. (plot())
- To display a relationship between a categorical explanatory variable and a quantitative response variable, make a side-by-side comparison of the distributions of the response for each category.

*How to quantify the relationship between the two variables?*

- The **correlation** measures the direction and strength of the *linear* relationship between two quantitative variables. Correlation is usually written as ***r***. cor(x, y) – in R

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

1. First, we **standardize the observations** both for x and y.
2. Then, we calculate ***r*** by averaging the products.

*What are the properties of r?*

1. Correlation makes no use of the distinction between explanatory and response variables. It makes no difference which variable you call **x** and which you call **y** in calculating the correlation.
2. Correlation requires that both variables be quantitative, so that it makes sense
3. Because r uses the standardized values of the observations, r does not change when we change the units of measurement of x, y, or both.
4. Positive r indicates positive association between the variables, and negative r indicates negative association.
5. The correlation r is always a number between −1 and 1. Values of r near 0 indicate a very weak linear relationship. The strength of the relationship increases as r moves away from 0 toward either −1 or 1. Values of r close to −1 or 1 indicate that the points lie close to a straight line.
6. Correlation measures the strength of only the **linear** relationship between two variables. Correlation does not describe curved relationships between variables, no matter how strong they are.
7. Like the mean and standard deviation, the correlation is **not resistant**: r is strongly affected by a few outlying observations.

8. Correlation is not a complete description of two variable data, even when the relationship between the variables is linear. You should give the **means and standard deviations** of both x and y along with the correlation.

# Least-Squares Regression

- A **regression line** is a straight line that describes how a response variable **y** (weight)changes as an explanatory variable **x** (calories) changes. We often use a regression line to predict the value of y for a given value of x. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

## Fitting a line to data

- https://www.youtube.com/watch?v=ZkjP5RJLQF4&t=723s
- https://www.youtube.com/watch?v=JvS2triCgOY&t=105s
- **Fitting a line to data** means drawing a line that comes as close as possible to the points.
- Equation for fitting the line is: y = b0 + b1x
- In this equation, b1 is the **slope**, the amount by which ***y*** changes when ***x*** increases by one unit. The number b0 is the **intercept**, the value of **y** when **x = 0**.
- HW: fit the line of these values and **predict** the rate of change for the 2019

| Total No. of passengers (millions) | Year |
|---|---|
| 130 | 2014 |
| 133 | 2015 |
| 140 | 2016 |
| 150 | 2017 |
| 159 | 2018 |

### LEAST-SQUARES REGRESSION LINE

- The least-squares regression line of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
- **Equation:**
    - We have data on an explanatory variable x and a response variable y for n individuals. The means and standard deviations of the sample data are ***x*** and ***sx*** for x and ***y*** and ***sy*** for

y, and the correlation between x and y is r. The equation of the least-squares regression line of y on x is

$$y = b0 + b1x$$

with slope

$$b1 = r^* (sy/sx)$$

and intercept

$$b0 = y - b1x$$

HW: Fit the least-squares regression line for:

NEA increase (cal) −94 −57 −29 135 143 151 245 355

Fat gain (kg) 4.2 3.0 3.7 2.7 3.2 3.6 2.4 1.3

*What are the properties of the least-squares regression line?*

- The slope and intercept of the least-squares line depend on the units of measurement—you can't conclude anything from their size.
- The expression b1 = r* (sy/sx) for the slope says that, along the regression line, a **change of one standard deviation in x corresponds to a change of r standard deviations in y**. When the variables are perfectly correlated (r = 1 or r = −1), the change in the predicted response y(hat) is the same (in standard deviation units) as the change in x. Otherwise, when −1 < r < 1, the change in y(hat) is less than the change in x. As the correlation grows less strong, the prediction y(hat) moves less in response to changes in x.
- The least-squares regression line **always passes through the point (x-bar, y-bar**)

*What connects correlation and regression?*

- **The square of the correlation, $r^2$**, is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x.
- When reporting a regression, give **r2** as a measure of how successfully the regression explains the response.
- **$r^2$** = variance of predicted values **y(hat)** / variance of observed values **y**
- The squared correlation gives the variance the responses would have if there were no scatter about the least-squares line as a fraction of the variance of the actual responses. This is the exact meaning of **"fraction of variation explained"** as an interpretation of r2.

*What should I be careful about with respect to correlation?*

## Residuals

- A regression line describes the overall pattern of a linear relationship between an explanatory variable and a response variable. Deviations from the overall pattern are also important. In the regression setting, we see deviations by looking at the scatter of the data points about the regression line. The vertical distances from the points to the least-squares regression line are as

small as possible in the sense that they have the smallest possible sum of squares. Because they represent “left-over” variation in the response after fitting the regression line, these distances are called **residuals**.

- A residual is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

  residual = observed y − predicted y

- The residuals from the least-squares line have a special property: **the mean of the least-squares residuals is always zero.**
- An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in **the y direction** of a scatterplot have large regression residuals, but other outliers need not have large residuals**.**
- An observation is influential for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the leastsquares regression line.
- **Correlation does not imply causation.**

## RANDOMNESS AND PROBABILITY

- A **random phenomenon** has outcomes that we cannot predict but that nonetheless have a regular distribution in very many repetitions.
- Probability describes only what happens in the long run. Most people expect chance outcomes to show more short-term regularity than is actually true.

- We call a phenomenon **random** if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions.
- The **probability** of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions.

### Probability Model

- A description of a random phenomenon in the language of mathematics is called **a probability model**.
- The **sample space S** of a random phenomenon is the set of all possible outcomes. The name "sample space" is natural in random sampling, where each possible outcome is a sample and the sample space contains all possible samples.
  E.g. S = {heads, tails}
  S = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}
  **Sample space for tossing a coin four times**
  Toss a coin four times and record the results. Counting shows that there are 16 possible outcomes. The sample space S is the set of all 16 strings of four H's and T's.

- How can we describe probability mathematically? We need to assign probabilities not only to *single* outcomes but also to sets of outcomes.
- An **event** is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.

### Probability Properties & Rules

1. Any probability is a number between 0 and 1. Any proportion is a number between 0 and 1, so any probability is also a number between 0 and 1.
   *Rule 1*. The probability P(A) of any event A satisfies $0 \le P(A) \le 1$.

2. All possible outcomes together must have probability 1. Because every trial will produce an outcome, the sum of the probabilities for all possible outcomes must be exactly 1.
   *Rule 2.* If S is the sample space in a probability model, then P(S) = 1.

3. If two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual probabilities. If one event occurs in 40% of all trials, a different event

occurs in 25% of all trials, and the two can never occur together, then one or the other occurs on 65% of all trials because 40% + 25% = 65%.
*Rule 3.* Two events A and B are **disjoint** if they have no outcomes in common and so can never occur together. If A and B are disjoint, P(A or B) = P(A) + P(B) -- This is the **addition rule** for disjoint events.

4. The probability that an event does not occur is 1 minus the probability that the event does occur. If an event occurs in (say) 70% of all trials, it fails to occur in the other 30%.
*Rule 4.* The complement of any event A is the event that A does not occur, written as $A^c$. The complement rule states that $P(A^c) = 1 - P(A)$

## EQUALLY LIKELY OUTCOMES

- If a random phenomenon has k possible outcomes, all equally likely, then each individual outcome has probability 1/k. The probability of any event A is

  P( ) = count of outcomes in A / count of outcomes in S

## THE MULTIPLICATION RULE FOR INDEPENDENT EVENTS

- *Rule 5.* Two events A and B are independent if knowing that one occurs does not change the probability that the other occurs. If A and B are independent,

  P(A and B) = P(A)P(B)

  This is **the multiplication rule** for independent events

### Dependent Events

Here is another example of a situation where events are dependent.

*Example* -- Taking a test twice. If you take an IQ test or other mental test twice in succession, the two test scores are not independent. The learning that occurs on the first attempt influences your second attempt. If you learn a lot, then your second test score might be a lot higher than your first test score. This phenomenon is called a carry-over effect.

### Conditional probabilty

The new notation P(A | B) is a conditional probability. That is, it gives the probability of one event (the next card dealt is an ace) under the condition that we know another event (exactly 1 of the 4 visible cards is an ace). You can read the bar | as “given the information that.”

Multiplication rule for the he probability that both of two events A and B happen together can be found by: P(A and B) = P(A)P(B | A) -- Here P(B | A) is the conditional probability that B occurs, given the information that A occurs.

## Bayes's Rule

Suppose that A1, A2, . . . , Ak are disjoint events whose probabilities are not 0 and add to exactly 1. That is, any outcome is in exactly one of these events. Then if C is any other event whose probability is not 0 or 1,

$$P(A_i \mid C) = \frac{P(C \mid A_i)P(A_i)}{P(C \mid A_1)P(A_1) + P(C \mid A_2)P(A_2) + \cdots + P(A_k)P(C \mid A_k)}$$

## LOGISTIC REGRESSION

- ***Logistic regression is multiple regression but with an outcome variable that is a categorical variable and predictor variables that are continuous or categorical.***
- ***E.g. male / female – categorical; weight, height – continuous;***
- When we are trying to predict membership of only two categorical outcomes the analysis is known as **binary** logistic regression, but when we want to predict membership of more than two categories we use **multinomial** (or polychotomous) logistic regression.
- Instead of predicting the value of a variable Y from a predictor variable $X_1$ or several predictor variables (Xs), we predict the *probability* of Y occurring given known values of $X_1$ (or Xs).

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots b_n X_{ni})}}$$

- Despite the similarities between linear regression and logistic regression, there is a good reason why we cannot apply linear regression directly to a situation in which the outcome variable is categorical. The reason is that one of the assumptions of linear regression is that the relationship between variables is **linear**.
- One way around this problem is to transform the data using the *logarithmic* transformation. This transformation is a way of expressing a non-linear relationship in a linear way. The logistic regression equation described above is based on this principle: it expresses the multiple linear regression equation in logarithmic terms (called the *logit*) and thus overcomes the problem of violating the assumption of linearity.

## ASSESSING THE MODEL

- These parameters are estimated by fitting models, based on the available predictors, to the observed data. The chosen model will be the one that, when values of the predictor variables are placed in it, results in values of Y closest to the observed values. Specifically, the values of the parameters are estimated using **maximum-likelihood estimation - MLE**, which selects coefficients that make the observed values most likely to have occurred.
- In multiple regression, if we want to assess whether a model fits the data we can compare the observed and predicted values of the outcome (if you remember, we use $R^2$, which is the Pearson correlation between observed values of the outcome and the values predicted by the regression model). Likewise, in logistic regression, we can use the observed and predicted values to assess the fit of the model. The measure we use is the **log-likelihood**. Large values of the log-likelihood statistic indicate poorly fitting statistical models, because the larger the value of the log-likelihood, the more unexplained observations there are.
- **The deviance** is very closely related to the log-likelihood: it's given by deviance = −2 × log-likelihood/ The deviance is often referred to as **−2LL** because of the way it is calculated. It's actually rather convenient to (almost) always use the deviance rather than the log-likelihood because it has a chi-square distribution which makes it easy to calculate the significance of the value.

- We can use the **Akaike information criterion (AIC)** and the **Bayes information criterion (BIC)** to judge model fit. The AIC is the simpler of the two; it is given by: AIC = −2LL + 2*k* in which −2LL is the deviance (described above) and *k* is the number of predictors in the model. The BIC is the same as the AIC but adjusts the penalty included in the AIC (i.e., 2k) by the number of cases: BIC = −2LL + 2*k* × log(n) in which n is the number of cases in the model.
- More crucial to the interpretation of logistic regression is the value of **the odds ratio**, which is the exponential of B (i.e., $e^B$ or exp(B)) and is an indicator of the change in odds resulting from a unit change in the predictor. The odds of an event occurring are defined as the probability of an event occurring divided by the probability of that event not occurring and should not be confused with the more colloquial usage of the word to refer to probability.

$$\text{odds} = \frac{P(\text{event})}{P(\text{no event})}$$

$$P(\text{event } Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i})}}$$

$$P(\text{no event } Y) = 1 - P(\text{event } Y)$$

## ASSUMPTIONS

- **Independence of errors**: Cases of data should not be related; for example, you cannot measure the same people at different points in time (well, you can actually, but then you have to use a multilevel model).
- **Multicollinearity:** Although not really an assumption as such, multicollinearity is a problem as it was for ordinary regression (see section 7.7.2.1). In essence, predictors should not be too highly correlated. As with ordinary regression, this assumption can be checked with tolerance and VIF statistics, the eigenvalues of the scaled, uncentred cross-products matrix, the condition indices and the variance proportions.
- **In RStudio:** Model1 <-glm(Accuracy ~ Language, data = Data, family = binomial())
  Model1 <-glm(Accuracy ~ Language*Gender, data = Data, family = binomial())