Displaying Data

2 Numerical Variables

- For 2 numerical variables, we use a scatterplot or a line plot
- A **scatter plot** displays 2 numerical variables in which each observation is represented as a point on a graph with 2 axes
- Position along the horizontal axis (the *x*-axis) indicates the measurement of the explanatory variable
- Position along the vertical axis (the *y*-axis) indicates the measurement of the response variable
- The trend indicates whether an association between the 2 variables is positive (lower left to the upper right), negative (left to the lower right), or absent (no discernible pattern)

1 Numerical & 1 Categorical Variable

- For 1 numerical and 1 categorical, we use a stripchart, box plot, multiple histograms or a line plot
  eg. stripchart is the left, box plot is the right
- The **strip chart (dot plot)** is a graphical display of a numerical variable and a categorical variable in which each observation is represented as a dot
- Explanatory variable is categorical, response variable is numerical
- Numerical measurement is on one axis and the category it belongs to is on the other
- Need to spread/jitter, the points along the horizontal axis to reduce overlap of points, so that they can be more easily seen
- A **box plot** is a graph that uses lines and a rectangular box to display the median, quartiles, range, and extreme measurements of the data (represented by a dot)
- A box plot indicates where the bulk of the observations lie
- **Multiple histograms** can be used, one for each category
- Histograms must be stacked above one another with the same scale as shown on the right so that the position and spread of the data are most easily compared

Line Graph

- A **line graph** uses dots connected by line segments to display trends in a measurement over time
- One *y*-measurement is displayed for each point in time or space, which is displayed along the *x*- axis
- Adjacent points along the *x*-axis are connected by a line segment
- When the baseline for the y-axis is 0, the area under the curve between 2 time points is proportional to the total number of new cases in that period

Maps

- A **map** is the spatial equivalent of the line graph, using a colour gradient to display a numerical response variable at multiple locations on a surface
- The explanatory variable is location in space
- One measurement is displayed for each point or interval of the surface

P-value (uncertainty)

- The **P-value** is the probability of obtaining a test statistic as extreme as we did, if the null hypothesis were true
- Data is not always a perfect match to the expectation of the null hypothesis due to chance
- To describe the mismatch between data and a null hypothesis, we calculate the chance of getting that data, or data that is even more different from that expected, while assuming the null hypothesis
- If the probability is small, then the null hypothesis is inconsistent with the data and we reject it in favour of the alternative hypothesis
- If the probability is not small, then we do not have enough evidence to doubt the null hypothesis, and we would not reject it
- The smaller the *P*-value, the stronger the evidence is, against the null hypothesis
- The *P*-value is not the probability that the null hypothesis is true
- The P-value is the probability of obtaining a result as extreme or more extreme than that observed
- The *P*-value is calculated from the null distribution for the test statistic
- Pr[14] is the probability of obtaining 14 out of 18, right-handed toads
- To include the equally extreme results at the left tail of the null distribution multiply by 2

  eg. $P=2\times(Pr[14]+Pr[15]+Pr[16]+Pr[17]+Pr[18]) = 2 \times 0.0155 = 0.031$
- If *P* is less than or equal to 0.05, then it is **significant** and we reject the null hypothesis
- If *P* is greater than 0.05, it is **non-significant** and we do not reject it
- P = 0.031, is less than 0.05, so we reject the null hypothesis that left-handed and right-handed toads are equally frequent in the toad population
- The **significance level *(α)*** is a probability used as a criteria for rejecting the null hypothesis. If the *P*-value is less than or equal to α, then the null hypothesis is rejected. If the *P*-value is greater than α, then the null hypothesis is not rejected
- When a statistical test returns a *P*-value larger than the significance level, the result is non-significant