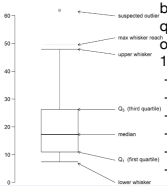
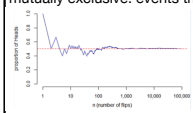
categorical: each observation belongs to one of a set of categories
quantitative: observations of it take on numerical values that represent different magnitudes of the variable
discrete quantitative: the values form a set of separate numbers
continuous quantitative: possible values from an interval
ordinal categorical: observations can be ordered
nominal categorical: classified into categories but have no specific ordering
frequency table: list all possible values, together with the frequency of each
  - mention the modal category and the proportion of it
proportion: count of observations in the category divided by total number
percentage: proportion multiplied by 100
bar plots: used to display a single categorical variable and the height is proportional to the frequency of the category.
  - mention the groups with high/low proportions and if there is ordering in the categories, mention if there is any apparent trend in proportions.
histogram: bars portray the frequencies of the possible outcomes for a quantitative variable
  - mention the overall pattern. data cluster together, or there is a gap such that one or more observations deviate from the rest. any outliers?
  - single mound(unimodal)? two mounds(bimodal)? or more(multimodal)?
  - is the distribution symmetric or skewed?
skew left: left tail is longer than the right tail. (median > mean)
skew right: right tail is longer than the left tail. (mean > median)
sample mean: sum of all the observations divided by the sample size
  - sensitive to extreme values
median: the middle value of the ordered observations
  - robust to extreme observations
range: difference between the largest and smallest observations
  - easy to compute but sensitive to extreme oobservations
variance: average of the squared deviations of the values from the mean
standard deviation: square root of the variance
68% of values are between mean - sd and mean + sd
95% of values are between mean - 2*sd and mean + 2*sd
all or nearly all of values are between mean - 3*sd and mean + 3*sd
100p-th quantile: 100p percent of the values fall below or at that value
Q1 = q0.25 = lower quartile: 25% observations are at or below it
Q2 = q0.5 = median = 50% ; Q3 = q0.75 = upper quartile = 75%
IQR: distance between upper and lower quartile, tells us how spread out the "middle" of the sample is.
dataset is symmetric and bell-shaped: report mean and sd and range
distribution is not bell-shaped: report median and IQR and range
the larger the intervals are, the more variability there is in the data and the histogram will be less peaked.



boxplots: visual of min, max, lower and upper quartile, outliers
outlier: < Q1 - 1.5*IQR (min whisker) or > Q3 + 1.5*IQR (max whisker)
  - boxplot does not portray mounds like histogram
  - can indicate skewness if distribution is unimodal
  - mention the median
  - if there are outliers, how many? which side?
  - more than one boxplot, compare median and IQR
  - used to compare one categorical and one quantitative
association: particular value for one variable is more likely to occur with certain values of the other variable.
response: variable on which comparisons are made
explanatory: variable you believe the response variable depends on
contingency table: display for two categorical variables
  - compare the percentages of the different categories of the response
barplots: visualisation of two categorical variables
scatterplot: visualisation of two quantitative variables.
  - mention the relationship/association. is it positive/negative/no association.
  - can the trend be approximated by a straight line, if so how do the points vary about the line?
  - are some observations unusual, departing from the overall trend?
correlation: indicates the strength and +ve/-ve of the association, value is between -1 and 1.
lurking variable: usually unobserved and not measured, that influences the association between the variables of primary interest and has potential to be a confounding variable.
confounding: when two explanatory variables are associated with a response variable and are also associated with each other, it is in the dataset.
observational study: values for the response variable and explanatory variables are observed for the sampled subjects, without anything being done by them.
sample surveys: study that asks questions/take measurements of the subjects in a sample drawn from the population randomly.
  - identify the population -> compile a list of subjects in the population from which the sample will be taken ->specify a method for selecting subjects from the sampling frame -> collect data
  - if we use chance rather than convenience, we can get a better representation of population
  - good sampling designs employ randomisation
simple random sample: n subjects from a sampling frame is one in which each possible sample of size n has the same chance of being selected.
  - representative sample allows us to make inferences about the population.
  - number the subjects -> generate a set of n random numbers -> subjects with the same number in the set of n numbers are selected
sampling designs : cluster random sampling & stratified random sampling
collecting data: personal face-to-face interview (easier to get response but costly), telephone interview(cheaper but subjects can refuse to participate), self-administered questionnaire (cheaper and less labour-intensive but more subjects may fail to participate).
sampling bias: result of sampling design or sampling frame
  - when the sample is not random and sampling frame does not represent the full population
non-response bias: some sampled subjects cannot be reached or refuse to participate
response bias: occur when the participant is not honest when answering or answering wrongly
experimental study: conducted by assigning subjects to certain experimental conditions and them observing the outcome on the response variable.
  - can control for lurking variables by randomly assigning the treatment
  - more confident to determine the causality between the explanatory and response variables.
  - it could be not ethical sometimes to out subjects under the treatment
  - is to investigate the association between treatment and response
 - have a control comparison group, randomly assign treatments to subjects and blind the study
randomization to eliminate bias that may appear if we assign subjects by hand, to balance the groups on lurking variables that we know it affects the response and that may be unknown to us.
mutually exclusive: events that can never occur simultaneously and the intersection is null



when n is small, P(A) is a little erratic, when it increases the value of proportion settles down.
P(A) = number of A outcomes / total number of possible outcomes
sample proportion estimates the true proportion.
independent: P(A intersect B) = P(A) * P(B)
mutually exclusive implies dependence.
P(A|B) = P(A intersect B) / P(B) ; P(A|B) = P(A) if A and B are independent
bayes theorem: P(B|A) = P(B) * P(A|B) / P(A)
sensitivity: probability that the test is positive, given that the person has the disease
specificity: probability that the test is negative, given that the person does not have the disease
prevalence: number of people who currently have the diesel, divided by total population
random variable: numerical measurement of the outcome of a experiment.
probability distribution: specifies its possible valued and their probabilities
discrete random variable: takes a set of seperate values and its probability distribution assigns a probability px to each possible value of X.
  - px is between 0 and 1 and the sum of the probabilities for all the possible values is 1.
 - use bar plot the visualise, the width of each rectangle is identical, height is proportional to p.
  - mean = x* px (sum of probabilities multiplied by possibilities) also known as E(X)
  - E(X_bar) = E(X)
  - Var(X) = (x- mean)^2 * px and larger variance = more variability
continuous random variables: random variable X has values that form an interval.
  - mean = E(X) = integration of x * f(x)
  - variance = Var(X) = integration of (x - mean)^2 * f(x)
  - P(X <= qp) = p, qp is the 100pth quantile
binomial distribution: X ~ Bin(n, p)
  - n trials, each of which has 2 possible outcomes. The outcome of interest is called success and the other outcome is called a failure.
  - each trial has the same probability of success, p
  - the n trials are independent.
  - Bin(1, p) is also referred as Bernoulli(p)
  - E(X) = np ; Var(X) = np(1 - p)
Poisson distribution: X ~ Poisson(lambda)
  - E(X) = Var(X) = lambda
  - Binomial(n , p) = Poisson(np) ; (1 - p) is approximately 1 for small p so np(1 - p) ~ np, where the mean and variance are almost equal.
normal distribution: X ~ N(mu, sigma^2)
  - it is symmetric, bell-shaped and
  - highest point of the curve is at x = mu
  - the curve is symmetric about mu
  - the larger the sigma^2, the larger the spread of values.
  - adding a constant to a normal variable, sum of normal variables and product of a normal variable with a constant all give new normal variables.
  - Z = (x - mean) / sigma ~ N(0, 1) also known as Z - score of X
  - outliers have Z-score greater than 3 or less than -3
  - Bin(n , p) = N(np, np(1 - p)) if np(1 - p) >= 5
sampling distribution: distribution of X_bar or p_hat
  - specifies probabilities for the interval of values of a statistic in a sample of subjects
 - help us determine how close to the population parameter a sample statistic is likely to fall
sampling distribution of p_hat: approximately N(p, p*(1 -p) / n)
  - np(1 - p) >= 5; p_hat = (X1 + … + Xn) / n
  - many samples, many p_hat, histogram of all the p_hat gives us the idea of the sampling distribution of p_hat
  - standard deviation of p_hat is the standard error of p_hat
sample mean: X_bar
  - varies from sample to sample and it is a random variable
  - if X1, …, Xn are independent from a normal distribution,
    X1 + … + Xn ~ N(n*mean, n*sigma^2)
  - X1 + … + Xn / n ~ N(mean, sigma^2/n)
when population distribution is Normal, Xi's are also Normal, X_bar follows Normal
  - histogram of X_bar has normal distribution
  - variability of the bell-shaped decreases as n increases
  - bell shapes are all centred at the population mean
  - it depends on mean, sigma^2 and n
  - larger sample size, sampling distribution is more peaked, X_bar is closer to true mean
when the population distribution is not normal, random sample of Xi is not normal
  - apply CLT when n >= 30, X_bar is approximately normal ~N(mean, sigma^2/n)
  - histograms of X_bar is still or close to bell-shaped and more symmetric when n is larger
  - variability of the bell-shape lowers as n increases.
  - bell shapes are all centred at the population mean
  - sampling distribution of X_bar depends on mean, sigma^2 and n.
central limit theorem: suppose than n >= 30, the sample mean can be approximated by a normal distribution by a normal distribution N(mean, sigma^2/n)
  - approximation gets better as n gets larger and if Xi's distribution is not too skewed.
  - the population distribution might be unknown
  - the data distribution might not be normal
data distribution: histogram visualises the observations from a single sample
  - larger n, the closer data distribution to the population distribution

point estimate: single number that is our best guess for the population parameter
- it varies from sample to sample as we are not taking the mean of the population and we are taking a random sample each time.
- it does not provide an idea about how close it is to the true value.
- p_hat is a point estimate for p
- X_bar is the point estimate for mu (mean)
- S^2 is the point estimate for Var(X)
- S is the point estimate for standard deviation
- X(0.5) is the point estimate for q0.5(50th percentile)
interval estimate: an interval of numbers which the parameter value is believed to fall
- indicates precision by giving an interval of numbers around the point estimate
- made up of numbers that are the most believable values for the unknown parameter, based on the data observed.
confidence interval: interval containing the most believable values for a parameter.
confidence level: the probability that this method produces an interval that contains the parameter.
- point estimate +- margin of error
margin of error: measures how accurately the point estimate is likely to be in estimating a parameter. it is a multiple of the standard deviation.
- for sample proportion: sqrt(p*(1-p)/n)
for population proportion, p_hat ~ N(p, p*(1-p)/n)
- n*p_hat*(1-p_hat) must >= 5
- 95%: p_hat +- 1.96 * sqrt(p_hat*(1-p_hat)/n)
- 99%: p_hat +- 2.58 * sqrt(p_hat*(1-p_hat)/n)
- 80%: p_hat +- 1.28 * sqrt(p_hat*(1-p_hat)/n)
- 100%: [0, 1] can have this without any sample, a lower CI can significantly narrow down the range of values for the parameter.
- __% that the proportion of ____ in the population falls in (__, __)
confidence: refers to a long-run interpretation, describes how well the method performs over many different random samples
- if we form many 95% CI for p, then in the long run, 95% of these intervals will contain the true p
error probability: 1 - CI probability that it does not contain the true value.
- we can increase the CI but the interval will become wider.
factors affecting the length of CI
- sample size: larger n, smaller interval
- CI: higher the CI, smaller alpha, wider interval
- the true value of p in the underlying population (cannot change)
- variance (sigma^2) of the population distribution (cannot change)
(1 - alpha)100% CI with length <= D, n >= (2 * q1-alpha/2 / D)^2 * p(1-p)
- if p is unknown, use p = 0.5 and it gives largest possible margin of error
- will give us the smallest possible n we need for the sample.
for population mean, X_bar ~ N(mean, sigma^2/n)
- sample must be obtained by randomisation
- distribution of data should be approximately normal or symmetric
- (X_bar - mean) / (s / sqrt(n)) ~ t(n - 1) t-distribution with df of n - 1
- 95%: X_bar +- t(n-1, 0.975)* s/sqrt(n)
t-distribution is also symmetric about 0, just like N(0, 1)
- probabilities under this depends on the degrees of freedom
- has thicker tails and more variability than N(0,1)
- the larger the value of df, the closer it is to N(0,1)
- when df is 30 or more, it is nearly identical to N(0, 1)
robust: when respect to a particular assumption if it performs adequately even when that assumption is modestly violated.
- randomisation assumption: not robust
- data distribution assumption: is robust
- sample size is large enough, we only need to ensure no extreme outlier
(1 - alpha)100% CI with length <= D, n >= (2 * t(n-1, 1-alpha/2) * s / D)^2
Hypothesis testing: using probability to quantify how plausible a value for a parameter is
hypothesis: statement about a population, claiming that a parameter takes a particular value or falls in a certain range of data.
step 1: assumptions
- data must come from randomisation
- sample size must be large enough or shape of the population distribution is symmetric.
step 2: stating hypotheses
null hypothesis: statement that the parameter takes a particular value
alternative hypothesis: statement that the parameter falls in some alternative range of values.
- H0 usually represents no effect and H1 represents effect of some type.
- right sided test: H1 parameter is larger than the value in H0
- left sided test: H1 parameter is smaller than the value in H0
- two sided test: H1 parameter is not equal to the value in H0
step 3: test statistic (describes how far the point estimate falls from the value in H0)
- we need the value of point estimate from the sample and its sampling distribution
- parameter value specified under H0.
step 4: p-value (probability that the test statistic receives the value as observed for more extreme, given that H0 is true)
- when it is small, either H0 is rejected or sample is not representative of the population
- small p-value provides strong evidence against H0.
- two sided test: p-value is left and right tail probabilities
- right sided test: p-value is the area on the right side of the test statistic
- left sided test: p-value is the area on the left of the test statistic
step 5: conclusion: if p-value <= alpha, we reject H0 as we have strong evidence against H0.
for proportions, variable measured is categorical
- distribution assumption fulfilled when n*p0*(1 - p0) >= 5, p0 is value in H0.
- H1: p > p0 OR p < p0 OR p != p0
- Z = (p_hat - p0) / sqrt(p0 * (1 - p0) / n) ~ N(0, 1)
for means, variable measures is quantitative
- normal assumption is crucial when n is small
- H1: mu > mu0 OR mu < mu0 OR mu != mu0
- T = (X_bar - mu0) / (s / sqrt(n)), follows t-distribution with (n-1) degrees of freedom.
type I error: occurs when we reject H0 when it is in fact true. denoted by alpha
type II error: occurs when we do not reject H0 when it is in fact false. denoted by beta
power of the test: probability of correctly rejecting H0 when it is in fact false. denoted by 1-beta
two independent samples
- in an experimental study, study units are assigned randomly to different treatments
- in an observational study, we draw a random sample from the population, and then observe variables. OR we draw a random sample from a population and then a sample from another population to observe the variable of interest.
two dependent samples: when we have two samples comprises the same set of subjects
- arise when the measurements are taken from the same set of subjects before and after a treatment to see how effective the treatment is
equal variance test: var.test(x, y) and it is not equal if the p-value is small
independent sample with equal variances
- quantitative response variable for both groups and the samples are independent.
- population distribution of each group is approximately normal and it is crucial when n is small
- variances are equal
- S^2p = ((n1 - 1)*S^2x + (n2 - 1)*S^2y) / (n1 + n2 - 2)
- T = (X_bar - Y_bar) / Sp*sqrt(1/n1 + 1/n2) ~ t(n1 + n2 -2)
independent samples with unequal variances
- variances are unequal
- T = (X_bar - Y_bar) / sqrt(S^2x/n1 + S^2y/n2)
dependent samples
- get the set of n differences, D1, …, Dn and treat it as one-sample data, H0: mu = 0
Normality assumption to check if the population distribution is approximately Normal
- checked using the sample distribution using histogram qq plots or shapiro-wilk test
qq plot
- might tail below straight line or left tail above straight line: longer than normal
- right tail above the straight line or left tail below the straight line: shorter than normal
shapiro-wilk test: when p-value is small, it does not follow a normal distribution
response variable OR dependent variable OR target variable OR output variable
explanatory OR regressor OR independent OR predictor OR input OR covariate
regression: mathematical relationship between the mean of Y and the different values of X
linear regression: the relationship is linear, Y = B0 + B1*X + E
assumption of a simple linear regression model, Y ~ X
- data obtained by randomization
- relationship between X and Y is linear
- error term, E, ~N(0, sigma^2) and sigma^2 is a constant
these assumptions implies that for any X value, Y ~ N(B0 + B1*X, sigma^2), hence variance is always the same and mean of Y is (B0 + B1*X)
the fitted regression line gives Y_hat and it is the point estimate for the mean of Y
interpolation: estimating the mean response for an X value that had not been observed, but it is within the range of observed values
extrapolation: estimating the mean response for an X value that is outside the range.
sigma^2 gives us the idea of the variability of the response values around the fitted line
smaller variance, closer to the fitted line
in simple linear model, the F-test is equivalent to the t-test
t-test tests for the significance of the coefficients of the regressor terms.
- H0: regressor is not significant, H1: regressor is significant
- t = B1_hat / SE(B1_hat)
- for a simple linear model, the null distribution of t is t(n-2)
F-test: tests for the significance of the whole model
- H0: model is not significant, OR H0: all coefficients except intercept are zero,
- H1: model is significant OR at least one coefficient except intercept is non-zero
- if we do not reject the H0, Y = B0 + E, Y_hat = B0_hat, Y does not depend on any
checking for assumptions
- normality and constant variance: checked using residuals of built model
- linearity: checked using scatter plot between response and regressor
- randomisation: checked in the steps of data collection
if linearity assumption is violated, can add higher order terms in X like X^2 to the model
if variance is not constant, can transform the response like ln(Y), sqrt(Y) and 1/Y
standardised residuals: Y - Y_hat / SE of (Y - Y_hat)
plot of SR against Y_hat or X: points scatter randomly about 0, within (-3, 3)
- funnel shape means the constant variance assumption is violated.
histogram and QQ plot of SR: normally distributed
- non-normal means normality assumption violated
when sample size n is large, expect SR to show randomness
linear scatter plot of Y against X
- curved band means the linearity assumption is violated
outliers have SR greater than 3 or less than -3
influential points: one that affect the parameter greatly and can test with Cook's distance
- if cooks distance > 1, which is the effect of deleting the point, it is an influential point
- not all outliers are influential points
R^2: proportion of total variation of the response that is explained by the model
- takes on values between 0 and 1
- if there are repeated values of X with different Y values, it can never be 1
- in a simple model, one regressor, R = |Cor(x, y)|
- used to comment on the goodness of fit of the model
adjusted R^2: measure of fit
- use this to compare multiple linear regression models
- 1 - (1 - R^2)*(n - 1) / (n - k - 1), k is the number of regressors
indicator variables: categorical variables in the model
- I(Xn = 1) include into the model only if the Xn variable is 1
Interaction term: interaction between the two variables
- add Xi * Xj into the model