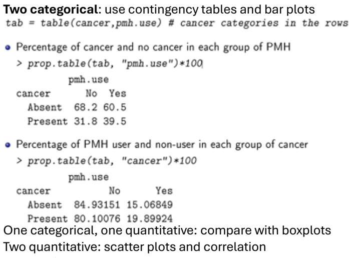
ST1131 AY24/25 S2 Nathan Loo github.com/natlv

**Categorical ordinal variables**: observations can be ordered but no specific quantitative values (e.g. ratings)
**Categorical nominal variables**: observations can be classified into categories, but categories have no specific ordering (e.g. gender, race, pregnancy status)
Note that categorical is aka random here
**Quantitative discrete variables**: usually countable, possible values form set of separate numbers (e.g. number of pets in a house, number of dengue cases in GRC)
**Quantitative continuous variables**: possible values form an interval (e.g. age, height, weight, blood pressure, IQ)
Note that we can group into ranges (age range, competition prize rank/no prize) -> treat as categorical ordinal
A **frequency table** is a listing of possible values, together with the frequency of each value.
The **proportion** of observations in a certain category is the count of observations in that category divided by the total number of observations. **Percentage** = proportion * 100. Proportions and percentages are **relative frequencies**.
**Bar plot** used to display single categorical variable with vertical bar for each category, height is category frequency. Summarize a bar plot: mention same points as for freq table. Mention if there are groups of categories with high/low proportions. If there is an ordering to the categories, mention if there is any apparent trend in proportions.
**Histogram** uses bars to display frequencies or relative frequencies of the possible outcomes for a quantitative variable. Can be created using frequency or probability (density). Look for: The overall pattern - data cluster together, or there is a gap such that one or more observations deviate from the rest. Any suspected outliers? Do the data have a single mound? This is known as a unimodal distribution. Data with two or more are known as bimodal or multimodal distribution. Is the distribution symmetric or skewed? (left skewed = left tail longer)
The mean is sensitive to **extreme observations**, whereas the median is not. The median is robust to extreme observations. **Summarize centre tendancy**: Highly skewed dataset, use median. Symmetric + bell-shaped, use mean.
**For unimodal distributions, usual relationship between mean and median**: Right-skewed: mean > median, symmetric: mean = median, left-skewed: mean < median
The **variance** of a set of values is the average of the squared deviations of the values from the mean. Variance formula:

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

The **standard deviation** s = sqrt(variance). Larger s means the values are more spread out from the mean.
**Linear transformations**
Orig sample mean = x̄ -> Linear transformation bx + a for all x -> new mean is bx̄ + a
New variance is orig variance multiplied by b^2
New s.d.is orig s.d.multiplied by absolute value of b
If a distribution is **bell-shaped**, then approximately
- 68% of the observations fall within 1 standard deviation of the mean, i.e. between the values x̄ - s and x̄ + s
- About 95% of the observations fall within 2 standard deviations of the mean. (x̄ ± 2s).
- All or nearly all the observations fall within 3 standard deviations of the mean (x̄ ± 3s).

**Quantiles** (percentiles) Let p be a value between 0 and 1. The 100p-th quantile, qp, is a value such that 100p percent of the values fall below or at that value. Sample with values: 1,2,...,100 then 90 is a q0.9, a 90th percentile. q0.25 -> 25% of observations at or below it -> first quartile, Q1
**Summarise sample**: Use variance and sd and mean if approx. bell-shaped distribution. Use IQR and median if not. Very different samples can have same mean and variance. Picture better than numerical summaries.
**Five-number summary** in R gives good indication of center and variability of a dataset. It's minimum, Q1, median, Q3, maximum. **Boxplot** is visual representation of this. Outliers are smaller than Q1 – 1.5*IQR or greater than Q3 + 1.5*IQR
The **response** / target variable is the variable on which comparisons are made.
The **explanatory** variable is any variable you believe the response depends on.
If the explanatory is a categorical variable, it defines the groups to be compared.
In some situations, we are unable to identify the role of variables as response or explanatory. We only can explore the association of two variables when treating them equally.
**Two categorical**: use contingency tables and bar plots



```
tab = table(cancer,pmh.use) # cancer categories in the rows
```

- Percentage of cancer and no cancer in each group of PMH

```
> prop.table(tab, "pmh.use")*100
         pmh.use
cancer      No  Yes
  Absent  68.2 60.5
  Present 31.8 39.5
```

- Percentage of PMH user and non-user in each group of cancer

```
> prop.table(tab, "cancer")*100
         pmh.use
cancer          No      Yes
  Absent  84.93151 15.06849
  Present 80.10076 19.89924
```

One categorical, one quantitative: compare with boxplots
Two quantitative: scatter plots and correlation
**Correlation formula**:

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

- r is always between -1 and 1. positive r is positive association, negative r is negative association.
- two variables always have the same correlation regardless which is treated as response or explanatory.
- If |correlation| is greater than 0.8, it is considered very strong. If it is between 0.5 to 0.8, it is relatively strong. If below 0.5, it is not strong.

**Lurking variable**: variable not in the dataset that influences association between variables of primary interest.
**Confounding**: two explanatory variables associated with a response variable, but are also associated with each other. Confounding variable is a variable included in the dataset. Hard to tell which of the two explanatories, if any, is causing a change in the response.
**Observational study**: the values for the response variable and explanatory variables are observed for the sampled subjects, without anything being done to them.

**Experiment**: assign subjects to experimental conditions (treatments). Observe outcome on the response variable. Randomly assign treatments controls for lurking variables. Experimental study better for determining causality.
**Sampling frame**: list of subjects in population from which sample will be taken. Ideal is all subjects in population.
**Sampling design**: method for selecting subjects from the sampling frame. Good sampling design uses randomisation
**Simple random sample** of *n* subjects from sampling frame: each possible sample of size *n* has same chance of being selected. Representative of population, can use to infer
**Cluster sampling**: Randomly select and sample groups of individuals (when easier to contact groups than individuals)
**Stratified sampling**: Split population into groups based on a characteristic e.g. income (when want representation from all subgroups and/or groups are very different)
**Sampling bias**: result of sampling design or sampling frame. When sample is not random or sampling frame does not represent the full population (under coverage)
**Non-sampling bias**. This occurs not due to sampling design. It has nonresponse bias and response bias.
- **Nonresponse bias**: some sampled subjects cannot be reached or refuse to participate.
- **Response bias**: participant answer wrongly or dishonest.

Misleading questions can result in response bias.
Large sample size does NOT guarantee an unbiased sample.
**Good experimental study** has control group, randomly assign treatment to subjects, blinding using placebo
**Sample space S**: set of all possible outcomes of a random phenomenon.
**Event**: subset of the sample space S that corresponds to a particular outcome or a group of possible outcomes.
**Probability** of an event is the proportion of times that this event occurs, in a long run of trials. It is between 0 and 1.
**Axioms**: Let A be an event within sample space S, P(A) = probability of event A. - P(A) ⩾ 0. - P(S) = 1.
- If A and B are mutually exclusive (disjoint) events, then P(A ∩ B) = 0 and P(A ∪ B) = P(A) + P(B).
- If A1, A2, . . . , Ar are pairwise mutually exclusive (no two of them can occur simultaneously), then P(A1 ∪ A2 · · · ∪ Ar ) = P(A1) + P(A2) + · · · + P(Ar ).
- (Additive Law) For any three events A, B, C: P(A∪B∪C) = P(A)+P(B)+P(C)−P(A∩B)−P(A∩C)−P(B∩C)+P(A∩B∩C)

**Other axiom implications**: P(complement of A) = 1 − P(A)

P(A ∪ B) = P(A) + P(B) − P(A ∩ B)
P(A) = P(A ∩ B) + P(A ∩ complement of B )
**true probability**: Can't alw repeat experiment many times until proportion settles down. So repeat several times then estimate true probability. -> Sample proportion estimates the true proportion. We might have reason to believe outcomes in our sample space have particular probabilities, so we assign these probabilities to the individual outcomes.
**Independent events**: P(A ∩ B) = P(A)P(B), P(A|B) = P(A)
Events from unrelated experiments always independent
Mutually exclusive -> dependent iff A > 0 and B > 0
**Conditional probability of A given B** when P(B) > 0

$$P(A|B) = \frac{P(A\ \cap\ B)}{P(B)}$$

**Partition of sample space** S created by mutually exclusive events B1 ∪ B2 ∪ B3 · · · ∪ Bn = S. Then for any event A,

$$P(A) = \sum_{i=1}^{n} P(A \cap Bi)$$

**Bayes Theorem**: A and B on same sample space S, P(A) > 0

$$P(B|A) = \frac{P(B) \times P(A|B)}{P(A)}$$

**Sensitivity** of a test is the probability the test is positive, given the person has the disease. Let D refer to disease, Sen = P(+|D). P( False positive ) = P(+| Dc ).
**Specificity** of a test is the probability that the test is negative, given that the person does not have the disease. Spec = P(−|Dc ). P( False negative ) = P(−| D).
**Prevalence** of a disease is no. of people who currently have the disease, divided no. of people in the population.
Example: A new bio-marker diagnosis assay has a sensitivity of 0.95 for a particular disease. It also has a specificity of 0.99 for the absence of the disease. The prevalence of the disease is 0.005. What is the probability that a person with a positive test result actually has the disease?

$$\begin{aligned} P(D|+) &= \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+\cap D) + P(+\cap D^c)} \\ &= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)} \\ &= \frac{0.95(0.005)}{0.95(0.005) + (1-0.99)(0.995)} \end{aligned}$$

**Random variable** is a numerical measurement of the outcome of a experiment. (we cannot know the precise value beforehand, if not it is not random). The values a random variable takes are defined on the sample space S.
**Probability distribution** of a random variable specifies its possible values and their probabilities.
**Discrete random variable** X takes on a set of separate values {0, 1, 2, 3, ...}. Its probability distribution assigns a probability px to each possible value of X. Bar plot uses a rectangle for each possible value X can take on. Width of each rectangle identical, but height is proportional to px.
**Mean of discrete random variable = Expected value = Sum of (Probabilities multiplied by Possibilities)**.
If we obtain a large number of observations from a population that follows a probability distribution, the sample mean of those observations would be close to the mean of that probability distribution.
Properties of the Mean
**1 (Linear transformation)** Let *X* be a random variable with *E(X) = μ*. Let *Y = bX + a*, where *b* and *a* are known constants. Then *E(Y) = bE(X) + a = bμ + a*.
**2** If (1) $X_1, X_2, ..., X_n$ are **n random variables** with their means are $\mu_1, \mu_2, ..., \mu_n$; (2) $a_1, a_2, ..., a_n$ are known constants; then $a_1X_1 + a_2X_2 + \cdots + a_nX_n$ is also a **random variable** with $E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$
**3** Let $X_1, X_2, ..., X_n$ denote *n* random variables identically distributed. They have same probability distribution hence have the same mean *μ*.
Let $\bar{X}$ denote the mean of all these variables. Then $\bar{X}$ is a random variable. Its mean is the same as the mean of each $X_i$. $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$ This is the special case of the second property when $a_i = 1/n$, *i = 1, ..., n*.
**Variance (risk) of discrete random variable**

$$\sigma^2 = \sum_{x}(x-\mu)^2\, p_x$$

Properties of the Variance

**1** (Linear Transformation) Let *X* be a random variable with variance $\sigma^2$. Let *Y = bX + a*, where *b* and *a* are known constants, then: $Var(Y) = b^2 Var(X) = b^2\sigma^2$.
**2** If (1) $X_1, X_2, ..., X_n$ are **n random variables** with respective variance $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$;
(2) $a_1, a_2, ..., a_n$ are known constants; then
$a_1X_1 + a_2X_2 + \cdots + a_nX_n$ is a random variable with
$Var(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$.
Let $X_1, X_2, ..., X_n$ denote *n* random variables identically distributed. They have the same mean and same variance, $\sigma^2$.
Let $\bar{X}$ denote the mean of all these variables. Then the variance of $\bar{X}$ *is* $Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$
**Continuous random variable** X has possible values that form an interval. Its probability distribution is specified by a curve that helps determine probabilities of intervals. This curve is referred to as a probability density function, or pdf. Each interval has probability 0-1. (This is area under the curve above that interval.) Total area under pdf curve = 1.
**Mean of continuous random variable** X with pdf f(x)= E(X)=

$$\mu = \int x f(x)dx$$

**Variance of continuous random variable** X with pdf f(x)=

$$Var(X) = \sigma^2 = \int (x-\mu)^2 f(x)dx$$

Mean and variance properties here same as for discrete
100p th Quantile, qp: point on x-axis such that area under the curve and to the left of qp, is equal to p.
**Binomial Distribution**: Suppose we have n trials, each of which has two possible outcomes. The outcome of interest is called a success and the other outcome is called a failure. Each trial has the same probability of success p. The n trials are independent. The total number of successes in the n trials is a binomial random variable. We write X ~ Bin(n, p). A Bin(1, p) distribution is also referred to as a Bernoulli trial or a Bernoulli distribution with success probability p.
- E.g. 20 seeds will each germinate independently with probability 0.6. Then Z, a random variable for total number of germinated, follows Bin(20, 0:6) distribution.
- E.g. P(breast cancer over lifetime) = 1/9, sample 50 women independently, follow them over their lifetime. If we set Y to be the total number of breast cancer, then Y ~ Bin(50, 1/9).

Suppose X follows Bin(n,p) distribution. Then probability of x successes in these n trials is P(X=x) = $(nCx)\, p^x\, (1-p)^{n-x}$
Mean of X, E(X) = np. Variance of X, Var(X) = np(1 – p)
**Poisson Distribution**: Random variable X follows Poisson distribution with parameter λ if P(X=k) = $\frac{e^{-\mu}\,\mu^k}{k!}$
where *e* is approx. 2.71828, λ is expected no,. of events per time unit, *μ* = λt is expected no. of events over time period t
- Example: Suppose the number of deaths from typhoid fever over a 1-year period is Poisson distributed with parameter *μ* = 4.6. Probability distribution, Y of no. of deaths over a 3-months period: For Y, because *μ* = 4.6, t = 1 year, it follows that λ = 4.6. For 3-months period, we have *μ* = 4.6 * 0.25 = 1.15. Therefore,

$$P(Y=0) = e^{-1.15} = 0.317$$

$$P(Y=1) = \frac{1.15}{1!}e^{-1.15} = 0.364$$

$$P(Y=2) = \frac{1.15^2}{2!}e^{-1.15} = 0.209$$

$$P(Y=3) = \frac{1.15^3}{3!}e^{-1.15} = 0.08$$

$$P(Y \geq 4) = 1 - (0.317 + 0.364 + 0.209 + 0.08) = 0.03$$

Poisson dist with parameter *μ*: mean = variance = *μ*
**Poisson Approximation to the Binomial Distribution**
- Binomial with large n and small p can be accurately approximated by Poisson distribution with parameter μ = np.
- Mean of this distribution is np and variance is np(1 − p), where (1 − p) is approximately equal to 1 for small p, and thus np(1 − p) ≈ np, ie mean and variance are almost equal.
- Binomial distribution cumbersome for large n.

**Normal Distribution** aka Gaussian: symmetric, bell-shaped, characterised by mean *μ* and variance $\sigma^2$
If X is random normally distributed variable: X~N(*μ*, $\sigma^2$)
Highest point of normal distribution curve: x = *μ*
Normal distribution symmetric about *μ*. This implies:
1) If *d* > 0, *P(X ≤ μ − d) = P(X ≥ μ + d)* 2) $q_{1-p} = 2\mu - q_p$
**Linear Transformation of Normal Random Variables**
Add constant to a normal variable, get new normal variable.
Sum of normal variables is a normal variable.
If $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, and *X* and *Y* are independent random variables, *a* is a constant then

$$X + a \sim N(a + \mu_x, \sigma_x^2)$$
$$X + Y \sim N(\mu_x + \mu_Y, \sigma_x^2 + \sigma_Y^2)$$

The addition could have >2 terms. If $X_1, X_2, ..., X_n$ are independently identically distributed (IID) $N(\mu, \sigma^2)$, then

$$X_1 + ... + X_n \sim N(n\mu, n\sigma^2)$$

Product of a normal variable with constant is a normal variable. For any real numbers *a* and *b*, if $X \sim N(\mu_x, \sigma_x^2)$, and $Y \sim N(\mu_Y, \sigma_Y^2)$ then

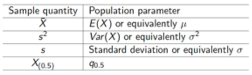
$$aX \sim N(a\mu_x, a^2\sigma_x^2)$$
$$aX + bY \sim N(a\mu_x + b\mu_Y, a^2\sigma_x^2 + b^2\sigma_Y^2)$$

When *a = 1* and *b = −1*, $X - Y \sim N(\mu_x - \mu_Y, \sigma_x^2 + \sigma_Y^2)$
**Standardisation of Normal Variable**
N(0, 1) is called standard normal distribution. If X~N(*μ*, $\sigma^2$): the Z-score of X, referred to as Z = $\frac{(X-\mu)}{\sigma} \sim N(0,1)$
Any observation of X with abs(Z-score) > 3 is an outlier.
Example: Test scores (X) follow N(*μ* = 1500; $\sigma^2$ = 90000) distribution. Find P(X ≤ 1800): pnorm(1800, mean = 1500, sd = sqrt(90000)) Find P(X ≥1630): pnorm(1630, mean = 1500, sd = sqrt(90000), lower.tail = FALSE)
Example: Measuring patients' blood pressure. The random variable X of this measurement for hypertensive patients follows a Normal distribution with *μ* = 95 and *σ* = 12. You wish to develop a screening test for hypertension with sensitivity 0.90. What cut-off pressure should you use?
Sensitivity = P(X ⩾ C) = 0.9
We need to find $q_{0\cdot 10}$ for X ~ N(95, 144). This equals 79.6.
In R, qnorm(0.1, mean = 95, sd = sqrt(144)) Ans: C = 79.62
**Normal Approximation to the Binomial Distribution**
If n is moderately large and p is not close to 0 or 1 then Bin(n, p) tends to be symmetric and is well approximated by a normal distribution N(np, np(1 − p)). The condition for the approximation to be good is np(1 − p) ⩾ 5.
**R and Binomial Distribution** (lower.tail TRUE by default)
Let X~Bin(3889, 0.531), then P(X ≤ 2000) is pbinom(2000, 3889, 0.531), P(X > 2000) is pbinom(2000, 3889, 0.531, lower.tail = FALSE), P(X < 2000) is pbinom(1999, 3889, 0.531), P(X ≥ 2000) is pbinom(1999, 3889, 0.531, lower.tail = FALSE)
Let *X ~ Bin(100, 0.5)*. The quantile $q_{0\cdot 9}$ is value such that the left area of it is 0.9, or $P(X \leq q_{0\cdot 9}) = 0.9$. qbinom(0.9, 100, 0.5)
**R and Normal Distribution**: P (X ≤ x) = P(X < x) pnorm(x, mean, sd) P (X ≥ x) = P(X > x) lower.tail=FALSE

Example: Let *X* be the height of NUS students and assume that $X \sim N(170, 10^2)$. The 90th percentile or $q_{0\cdot 9}$ is then qnorm(0.9, 170, 10) = 182.8155; That means 90% of students are as tall as or shorter than 182.8. The proportion of students whose height are from 150 to 190 is pnorm(190, 170, 10) – pnorm(150, 170, 10).
**Sampling distribution**: Distribution of a sample statistic, eg sample mean X̄ or sample proportion p̂, is sampling dist. It specifies probabilities for interval of values of a statistic in a sample of subjects. It helps determine how close to the population parameter a sample statistic is likely to fall.
**Sampling distribution of Sample Proportion p̂**
(voting outcome, Bin(1, p) ) California election: p̂ = 0.531
We know this statistic has sd equal to $\sqrt{\frac{p(1-p)}{3889}}$
We don't know p so we replace in eqn with best estimate p̂. sd of sampling distribution is estimated as 0.008. this is the **standard error** of p̂. n is large -> distribution of p̂ is well approximated by a normal distribution. Almost all the observations of p̂ are within 3 sd of the mean.
0.531 ± 3 × 0.008 = (0.507, 0.555)
**Central Limit Theorem**
Suppose we have independent observations $X_1, X_2, ..., X_n$ from a distribution with mean μ and variance $\sigma^2$.
As a general guide, suppose n ⩾ 30. Then the sample mean is approximated by a normal distribution, N(μ, $\sigma^2$/n). The approximation gets better as n larger. The approximation gets better if the Xᵢ's themselves not too skewed. Note that population distribution may be unknown & data distribution may not be normal (although near normal -> btr approx.)
**Central Limit Theorem Special Case** – regardless size of n, If $X_1, X_2, ..., X_n$ are independently from a normal distribution N(μ, $\sigma^2$). Then their sum, $X_1 + X_2 + ... + X_n$ follows N(nμ, n$\sigma^2$) and sample mean $(X_1 + X_2 + ... + X_n)$ / n follows N(μ, $\sigma^2$ / n) These values follow normal distribution exactly (not approx.)
**Sampling Distribution of Sample Mean**
- For a random sample of size n from a population with mean μ and standard deviation σ, the sampling distribution of X̄ has its center equal to the population mean μ, and its variability described by standard deviation σ / sqrt(n).
- If the population distribution is normal to begin with, then X̄ is Normal. If the population distribution is not normal, then the sampling distribution of X̄ approaches normal as n becomes larger, n ⩾ 30.

Normal Population Distribution: Histogram of X̄ has normal distribution. Variability of bell-shaped less as $n$ increase. Bell shapes are all centered at the population mean $\mu$ . Sampling distribution of X̄ depends on $\mu$, $\sigma^2$ and $n$. Num of samples drawn, N, no impact if sample size $n$ is constant.
Non-Normal Population Distribution: Histograms of X̄ are or close to bell-shaped. They are more symmetric when $n$ is larger. Variability (spread) of the bell-shape gets less as we increase $n$. Bell shapes are all centered at the population mean $\mu$ . Sampling distribution of $\bar{X}$ depends on $\mu$, $\sigma^2$ and $n$.
Consider just one sample of size n. Let us form a histogram of the observations from this single sample. This is referred to as the **data distribution**. The larger n, the closer data distribution to the population distribution.
Back to the example of BMI of people at the beginning of this topic. What is the sampling distribution of X̄? • Population distribution is unknown, but data distribution is not skewed. • From the given data/sample, X̄ = 24.9 and s = 4.77. • n = 34 is considered as large enough, data distribution is not skewed. • Hence, by CLT, the sampling distribution of X̄ is approximated by N(μ, $\sigma^2$/34), where μ is estimated by 24.9 and σ is estimated by s = 4.77. • Conclude: sampling distribution of X̄ is approximated by N(24.9, 0.669).
When the sampling distribution of X̄ is normal, we 95% sure that a random value of X̄ will fall within 2 standard deviations (2σ / sqrt(n) ) from the population mean. When n is large, the sampling distribution is approx. normal even if population distribution is not, so we can still make the same claim.
**Statistical inferences** Estimation of population parameters
- **Point estimate**: a single number that is our best guess for the population parameter. Find using appropriate statistic from a random sample. E.g. for population mean μ, can use sample mean X̄. For population proportion p, can use sample proportion p̂. Point estimates vary from sample to sample (random samples). Point estimate does not provide an idea about how close it is to the true value it estimates.

Recommended sample quantities as point estimates for their population analogs.



| Sample quantity | Population parameter |
|---|---|
| $\bar{X}$ | $E(X)$ or equivalently $\mu$ |
| $s^2$ | $Var(X)$ or equivalently $\sigma^2$ |
| $s$ | Standard deviation or equivalently $\sigma$ |
| $X_{(0.5)}$ | $q_{0.5}$ |

- **Interval estimate**: Interval of numbers within which the parameter value is believed to fall (based on data observed, numbers around point estimate). **Confidence interval**: an interval containing the most believable values for a parameter. The probability this method produces an interval that contains the parameter is the **confidence level**. e.g. when conf level 0.95, the interval is 95% confidence interval

**Confidence interval = Point estimate ± margin of error**
- margin of error measures how accurate the point estimate is likely to be in estimating a parameter. It is a multiple of the standard deviation of the sampling distribution of the point estimate. For instance, when the sampling distribution is approximately Normal, a 95% confidence interval has a margin of error equal to 1.96 standard deviations.

HOW TO CALCULATE:
1. let confidence level be CL% e.g. 97%
2. Remaining area in tails: 1 – 0.CL e.g. 1- 0.97 = 0.03
3. Each tail gets half e.g. 0.03 / 2 = 0.015
4. Get z-value from Z-table for value in step 3. e.g. 1 – 0.015 -> 2.17

If $\bar{X}$ is your sample mean, the **97% confidence interval** is:

$$\bar{X} \pm 2.17 \cdot \frac{\sigma}{\sqrt{n}} \quad \text{(or } \frac{s}{\sqrt{n}} \text{ if } \sigma \text{ is unknown)}$$

formula for the 95% confidence interval for *p* will be

$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

CI for p should be used only when n is sufficiently large that n p̂ (1 – p̂ ) ≥ 5. If not, we need more observations.
**General Procedure for Confidence Interval for Proportion**
Let x be the confidence level, e.g. x = 0.95. To find a 100x% CI, the steps are: Find p̂ from the given sample. Ensure that np̂ (1 − p̂) ⩾ 5. Otherwise, obtain more observations. Find α = 1 − x. Find quantile $q_{1-\alpha/2}$ from N(0, 1), aka $z_{1-\alpha/2}$. Return the desired confidence interval as

$$\hat{p} \pm q_{1-\alpha/2} \times \sqrt{\frac{\hat{p}\,(1-\ \hat{p})}{n}}$$

**Confidence**: long-run interpretation that describes how well the method performs over many different random samples.

If we form many 95% confidence intervals for p, then in the long run 95% of these intervals would contain the true p.

**Factors affecting the length of CI for p**

- Sample size: Larger sample size n -> narrower interval
- Conf level: Higher conf level (smaller α) -> wider interval
- The true value of p in the underlying population (which we cannot change).

**Confidence Interval Length/Width**

For a $(1-\alpha)100\%$ CI, the formula is

$$\hat{p} \pm q_{1-\alpha/2} \times \sqrt{\frac{p(1-p)}{n}}$$

The length/width of this interval is

$$2 \times q_{1-\alpha/2} \times \sqrt{\frac{p(1-p)}{n}}$$

If we want a (1 – α)*100% CI where the length is ≤ D, solving for n gives us

$$n \geq (\frac{2 \times q_{1-\frac{\alpha}{2}}}{D})^2\, p(1-p)$$

If p is unknown, we use p = 0.5, as it corresponds to the largest possible value of margin of error for a given α. It will give us smallest possible n we should have for the sample.

- California Election: suppose before starting the survey, we did not know how many voters to take, but we knew that we wanted to make a 95% CI, and it should have length of at most 0.1. Then before collecting sample, we could compute $n \geq (\frac{2\times1.96}{0.1})^2\, 0.5(0.5) = 384.16$ -> min sample 385

**Confidence Interval for Mean**

variance of $\bar{X}$ is $\sigma^2 / n$ where $\sigma^2$, the population variance, is unknown. From $X \sim N(\mu, \sigma^2 / n)$, if we estimate $\sigma^2$ by $s^2$, then we'll have $\frac{(\bar{X}-\mu)}{s/sqrt(n)} \sim t_{n-1}$

$t_{n-1}$ -> t-distribution with n-1 degrees of freedom

**Properties of t-distribution**

The t-distribution is symmetric about 0, just like a N(0,1) distribution. The probabilities under the t-distribution depend on the degrees of freedom, df. The t-distribution has thicker tails and more variability than that of N(0,1). The larger the value of df, the closer the t-distribution gets to the N(0,1). When df ⩾30, t_df are nearly identical to N(0,1).

Since $(\bar{X}-\mu)/(s/\sqrt{n})$ follows $t_{n-1}$, we have

$$P(-t_{n-1,0.975} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq t_{n-1,0.975}) = 0.95,$$

where $t_{n-1,0.975}$ is quantile of probability 0.975 from a $t_{n-1}$.

Hence,

$$P\left(\bar{X} - t_{n-1,0.975}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,0.975}\frac{s}{\sqrt{n}}\right) = 0.95$$

The 95% CI for $\mu$ is then

$$\bar{X} \pm t_{n-1,0.975}\frac{s}{\sqrt{n}}.$$

df = 6 -> find in R using qt(0.975, 6)

Suppose that we have a sample of size $n$

1. This sample must be obtained by randomization, either by a random sample or a randomized experiment.
2. The distribution of the data should be approximately normal or symmetric.

Then a 95% confidence interval for a population mean $\mu$ is

$$\bar{X} \pm t_{n-1,0.975} \times \frac{s}{\sqrt{n}}$$

where $t_{n-1,0.975}$ corresponds to the 0.975-quantile of a $t$-distribution with $(n-1)$ degrees of freedom.

Example

From the output, we can see that sample mean $\bar{X} = 3.208$ kg, sample standard deviation $s = 0.506$, $n = 47$.

From the $t$-distribution with $df = 46$, we have $t_{46,0.975} = 2.01$.

Hence the 95% confidence interval is

$$3.208 \pm 2.01 \times \frac{0.506}{\sqrt{47}} = (3.06,\ 3.36)$$

We are 95% confident that the interval (3.06, 3.36) contains $\mu$.

Let $x$ be the confidence level. To find a 100x% CI for $\mu$, the steps are:

1. Find $\bar{X}$ from the given sample.
2. Check the assumptions.
3. Find $\alpha = 1 - x$ and derive $t_{n-1,1-\alpha/2}$.
4. Return the desired confidence interval as $\bar{X} \pm t_{n-1,1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$

**Note**: If $n$ is large enough and $\sigma$ is known, then the margin of error is $z_{\alpha/2}(\sigma/\sqrt{n})$ instead.

**Robustness**: A statistical method is said to be **robust** with respect to a particular assumption if it performs adequately even when that assumption is modestly violated.

E.g. sample obtained by randomization, data distribution normal / symmetric

**Factors affecting length of CI for μ**

- Sample size, confidence level similar to factors for p
- The variance $\sigma^2$ of the population distribution (which we cannot change).

**Approximating Sample Size**

- Unknown $t_{n-1,1-\alpha/2}$: We replace this by $q_{1-\alpha/2}$ from a $N(0,1)$ distribution. We should reduce the impact of this approximation by making sure $n$ will be at least 30.
- Unknown $s$: We can estimate $s$ by looking for $s$ from a similar study, or by conducting a pilot study to get an initial estimate.

Thus the formula to use in practice is

$$n \geq \left(\frac{2q_{1-\alpha/2}s}{D}\right)^2$$

Example: new study to assess mean weight of babies born to first time mothers. we want to obtain sample of size n, such that length of 95% CI at most 2kg. Assume variability of new study like previous observations e.g. estimate s by = 0.5064. We use q0.975 = 1.96. Then we collect n observations for the new study (ans: 99)

$$n \geq \left(\frac{2q_{1-\alpha/2}s}{D}\right)^2 = \left(\frac{2(1.96)0.5064}{0.2}\right)^2 = 98.5$$

Use $\hat{p}$ for categorical data (e.g. yes/no, success/fail) -> count proportion of successes, and $\bar{X}$ for quantitative (e.g. height/weight) -> find average of measured values

**Hypothesis Testing** (check if data supports a statement about a population, these statements are hypotheses) – a hypothesis usually claims a parameter takes a particular numerical value or falls in a certain range of values

**GENERAL STEPS FOR HYPOTHESIS TESTING**

Step 1: Assumptions. most important assumption is that we must have data that come from randomization. Other assumption may be about the sample size (e.g. that it must be large enough); or assumption about the shape of the population distribution (e.g. that it is symmetric).

Step 2: Stating hypotheses. **Null hypothesis H0**: states that the parameter takes a particular value. (represent no effect) **Alternative hypothesis H1**: states that the parameter falls in some alternative range of values.

From alternative hypothesis, can determine side of test.

- If the statement in H1 is: the parameter is not equal to the value under H0, then we have a two-sided test.
- If the statement in H1 is: the parameter is larger than the value under H0, then we have a right-sided test.
- If the statement in H1 is: the parameter is smaller than the value under H0, then we have a left-sided test.

Step 3: **Test Statistic**. describes how far point estimate falls from H0 parameter value. Usually measured by no. of std errors between point estimate and H0 parameter value.

- In order to compute the value of the test statistic (*) we will need: the value of point estimate from the sample, its sampling distribution, and the parameter value specified under H0.
- Test statistic is a variable. The value calculated in (*) is just an observation of this variable from a given sample.
- The distribution of a test statistic under H0 is called null distribution.

Step 4: **p-value**. "If the null hypothesis H0 were true, how likely would we be to observe a test statistic as extreme (or more extreme) than what we got?" Very small p-value -> either assumption that H0 is true is not correct, or sample not representative of population. p-value small (close to 0) provides strong evidence against H0.

Step 5: Conclusion. If a significance level α was pre-specified (usually 0.05 or 0.01), we need to make a decision on the validity of H0. -> Compare p-value to α. If p-value ≤ α, then we reject H0. Otherwise we retain (don't reject) H0. When we reject H0, we say the test is statistically significant

**ONE SAMPLE DATA**

**HYPOTHESIS TESTING FOR PROPORTIONS**

Assumptions: categorical variable, data obtained using randomisation, sample size n sufficient large such that the sampling distribution of the sample proportion $\hat{p}$ is approximately normal when the null is true.

n * p0 * (1 – p0) >= 5, where p0 is the value specified in h0

Hypothesis: select null h0 and alternative h1

Test Statistic

Test statistic measures how far the sample proportion $\hat{p}$ falls from the hypothesis value $p_0$.

When $H_0$ is true, $p = p_0$, the sampling distribution of $\hat{p}$ is then

$$\hat{p} \sim N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$$

The test statistic is the distance of $\hat{p}$ from $p_0$ in term of| its standard deviation:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \quad \text{this } Z \sim N(0,1).$$

p-value summarises evidence against h0 and supporting h1.

- If H1 is a two-sided test, then p-value is the two areas (left and right tail probabilities) of the test statistics Z.
- If H1 is a right-sided test, then p-value is the area on the right side of test statistics Z.
- If H1 is a left-sided test, then p-value is the area on the left of test statistics Z.

Assume that, at the Step 2, we chose to perform a **two-sided test** with

$$H_0 : p = 0.5 \quad \text{vs} \quad H_1 : p \neq 0.5$$

What would the $p$-value for this two-sided test be?

```
> # area in the right of test statistic
> pnorm(3.866, lower.tail = FALSE)
[1] 5.531747e-05
> # area in the left tail (left of -3.866)
> pnorm(-3.866)
[1] 5.531747e-05
> # p-value
> 2*pnorm(3.866, lower.tail = FALSE)
[1] 0.0001106349
```

**HYPOTHESIS TESTING FOR MEANS** (1 sample t-test)

Assumptions: quantitative variable, data obtained using randomisation, population distribution is approximately Normal (This assumption is crucial when n is small)

Hypothesis: The null hypothesis of a test about the mean has the form H0 : μ = μ0 (μ0 is the hypothesized mean of the population). Two-sided alt hypothesis would be H1: μ ≠ μ0. One-sided alternative is either H1: μ < μ0 or H1 : μ > μ0.

Test Statistic: $T = \frac{(\bar{X}-\mu0)}{s/sqrt(n)}$ where $\bar{X}$ is point estimate of population mean. If H0 is true, then T follows a t distribution with (n – 1) degrees of freedom. recall that n is sample size. Observed sample mean $\bar{X}$ is approximately T standard errors away from the null value μ0.

The null distribution of the test statistic is t n-1.

| Alternative hypothesis | $p$-Value |
|---|---|
| $H_1 : \mu \neq \mu_0$ | Two tail probability from $t_{n-1}$ |
| $H_1 : \mu > \mu_0$ | Right area of $T$ from $t_{n-1}$ |
| $H_1 : \mu < \mu_0$ | Left area of $T$ from $t_{n-1}$ |

If a significance level α is given, we can make a decision on reject or do not reject H0 by comparing p-value with α.

Example: 95% Confidence Interval for Mean Bill Length

- From the sample, $X$ = 47.5, n = 123 and s / √ n = 0.278.
- From the $t_{122}$ distribution, we have $t_{122,0.975}$ = 1.98
- Hence, the 95% CI is 47.5 ± 1.98 × 0.278 = (46.95, 48).
- The interval (46.95, 48) does not contain the null value 38.8.
- This is consistent with the decision from the test (i.e. reject 38.8).

The consistency between a CI and a significance test happens when

- the CI has confidence level of 100x% and the test is concluded at significance level α = (1 − x);
- the test must be of two-sided test to match with the CI which has lower bound and upper bound;
- both the CI and the test use the same standard error.

**Type I Error**: reject H0 when H0 is true. Probability = α

**Type II Error**: not reject H0 when H0 is false. Probability = β

Power of a test is 1 – β. Power of a test is the probability of correctly rejecting H0, when it is in fact false.

- cannot reduce both types of errors simultaneously

**TWO SAMPLE DATA**

**Independent Samples, Equal Variances (pooled, Ftest)**

Assumptions: A quantitative response variable for both groups, Two samples are independent, The population distribution of each group is approximately Normal (this assumption is important when n is small),

The variances of the two populations are the same. We shall check this by a test in R. If the test is significant (has p-value smaller than 0.05), then this assumption fails and we have to resort to the unequal variance test.

H0: the two samples are from two populations with the same variance. In R, we use equal var test "var.test(x,y)"

Hypothesis: The null hypothesis of a test for comparing two means: H0 : μX − μY= 0 where $\mu_x$ and $\mu_Y$ is population mean of 2 groups. A two-sided test has alternative hypothesis be:

H1 : μX − μY ≠ 0 One-sided alternative is either:

H1 : μX − μY < 0 or $H_1$ : μX − μY > 0

Test Statistic: The pooled estimate of common variance, $\sigma^2$, is estimated by $s_p^2 = \frac{(n_1-1)s_X^2 + (n_2-1)s_Y^2}{n_1+n_2-2}$.

The test statistic is then

$$T = \frac{(\bar{X}-\bar{Y})-0}{se} \quad \text{where} \quad se = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Under H0, T follows t-distribution with (n1+n2 –2) df

| Alternative hypothesis | $p$-Value |
|---|---|
| $H_1 : \mu_X - \mu_Y \neq 0$ | Two tail probability from $t_{n_1+n_2-2}$ |
| $H_1 : \mu_X - \mu_Y > 0$ | Right area of $T$ from $t_{n_1+n_2-2}$ |
| $H_1 : \mu_X - \mu_Y < 0$ | Left area of $T$ from $t_{n_1+n_2-2}$ |

**Independent Samples, Unequal Variances (Welch's)**

Use when test of equal variance is significant, i.e. p-value is small, so we reject the assumption that variances are equal.

Hypothesis: H0 : μX − μY = 0

H1 : μX − μY ≠ 0 or H1 : μX − μY > 0 or H1 : μX − μY < 0

Test statistic for this version is

$$T = \frac{(\bar{X}-\bar{Y})-0}{se} \quad \text{where} \quad se = \sqrt{\frac{s_X^2}{n_1}+\frac{s_Y^2}{n_2}}$$

If H0 true, T flw t-distribution. df may be complex/not integer

Rely on R to calculate df and p-value

**Dependent Samples**

e.g. the two groups/samples comprise same set of subjects/individuals, e.g. before n after treatment.

Two samples dependent means each observation in one sample has a matched observation in the other sample.

Instead of comparing two means, we can take the difference of matched observations and compare the mean of differences with 0. Set of n differences can be treated as one-sample data. Let μ be mean of differences of matched subjects in population, H0 : μ = 0. Similar to one-sample.

In R, t.test(diff, mu = 0, alternative = "greater") equivalent to t.test(Yes,No,alternative = "greater",paired = TRUE)

**Normality Assumption**

One assumption in test for population mean μ is that the population distribution is approximately Normal.

Need to check if the population distribution is approximately Normal. However, we do not have the population distribution. Hence, we need to check this assumption using the sample distribution instead.

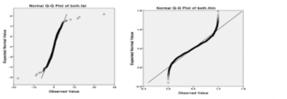
If the sample distribution is approximately Normal, then there's a high probability the population follows it as well.

**QQ Plots**

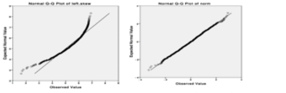
- Given a sample: 14, 13, 1, 20, 7, 14, 2, 19, 8, 10. How check if sample follows a Normal distribution? First standardize by taking each value, deduct the mean, then divide by the standard deviation, then sort. Standardized sample: -1.52, -1.37, -0.59, -0.43, -0.12, 0.34, 0.50, 0.50, 1.27, 1.43. If normal, the standardized sample should have values matching the appropriate quantiles from a standard Normal distribution: $q_{0.05}, q_{0.15}, \cdots, q_{0.95}$ or solved as -1.64, -1.04, -0.67, -0.39, -0.13, 0.13, 0.39, 0.67, 1.04, 1.64
- If observed close to expected, we assume sample follows Normal distribution. We can plot the expected quantiles against the observed quantiles to see if they match. This is what we call a quantile-quantile plot, or QQ plot.

QQ plot plots the standardized sample quantiles against the theoretical quantiles of a N(0; 1) distribution. If they fall on a straight line, then we would say that there is evidence that the data came from a normal distribution.

From the points on the plot, we can usually tell whether our sample data has longer or shorter tail than the normal, and on which side of the mean this occurs.



Left: both tail longer than normal. Right: both tail shorter



Left: Left tail longer than normal, right tail shorter than normal. Right: both tails are normal.

- Right tail is below the straight line: longer than normal.
- Right tail is above the straight line: shorter than normal.
- Left tail is below the straight line: shorter than normal.
- Left tail is above the straight line: longer than normal.
- longer=heavier=thicker=fatter, shorter=lighter=thinner.

**Shapiro-Wilk Test** (quantitative, gd for small samples only)

- H0: sample is from a Normal distribution.
- H1: sample is not from a Normal distribution.
- A small p-value would reject H0.
- We want large p-value for Normality assumption to hold.

**Linear Regression**

- response=dependent=target=output variable
- explanatory=independent=predictor=input variable=regressor=covariate

A regression of the response variable Y on the regressor X is a mathematical relationship between the mean of Y and different values of X. Linear regression means that this relationship is linear, of the form: Y = β0 + β1X + ε.

ε is a random variable. It has variance $\sigma^2$.

β0 is the Y-intercept, and β1 is the slope of the line, known as coefficients or parameters of the model.

The word "linear" refers to the linearity in the parameters.

The following are still linear regression models: Y = β0+ β1 sin(X) + ε , Y = β0 + β1 log(X) + ε , Y = $\beta_0$ + $\beta_1$ $e^x$ + ε

The following are NOT linear regression models:

Y = β0 sin(β1X) + ε , Y = β0 e^(β1X) + ε

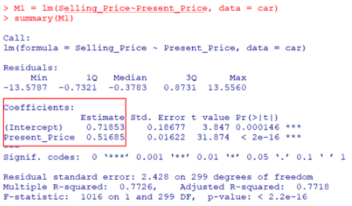
The word "simple" refers to only one regressor in the model.

Linear model has >1 regressor: multiple linear regression

Model Assumptions: Data were obtained by randomization. Relationship between X and Y is linear. The error term ε~N(0, $\sigma^2$) where $\sigma$ is a constant. (Note: we do not check these assumptions bfr building a model, but after fitting model)

Implications of Assumptions: • For any particular X value, the response is a variable that has a normal distribution: Y ~ N(β0 + β1X, $\sigma^2$). • For any particular X value, the mean of variable Y is ($\beta_0$ + $\beta_1$X). • For any values of X, the variance of Y is always the same: $\sigma^2$. • $\beta_0$, $\beta_1$ and $\sigma^2$ are the parameters to be estimated.

Ordinary Least Squares Estimation: Lowest sum of squared residuals used to find line of best-fit



```
> M1 = lm(Selling_Price~Present_Price, data = car)
> summary(M1)

Call:
lm(formula = Selling_Price ~ Present_Price, data = car)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5787  -0.7321  -0.3783   0.8731  13.5560 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.71853    0.18677   3.847 0.000146 ***
Present_Price  0.51685    0.01622  31.874  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.428 on 299 degrees of freedom
Multiple R-squared:  0.7726,	Adjusted R-squared:  0.7718 
F-statistic:  1016 on 1 and 299 DF,  p-value: < 2.2e-16
```

The OLS estimates of the slope ($\beta_1$) and the intercept ($\beta_0$) are rounded as 0.52 and 0.72 respectively. They are the point estimates of β1 and β0. We write $\hat{\beta_0}$ = 0.72, $\hat{\beta_1}$ = 0.52

We say that our fitted or estimated regression line is

$\hat{Y}$ = 0.72 + 0.52X where Y is Selling_Price, X is Present_Price.

Given a value of X, $\hat{Y}$ is point estimate of mean selling price.

In the summary, point estimate of σ is $\hat{\sigma}$ = 2.428.

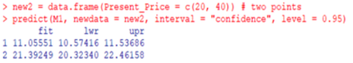
Point estimate of $\sigma^2$ computed using **raw residuals** i.e. ei = Yi – $\hat{Y}$i. Raw residuals are our best estimates for εi.

In R, M1$res lists all raw residuals of model M1

**Interpolation**: estimating mean response for an X value that had not been observed, but is within the range of observed values. **Extrapolation**: estimating mean response for an X value that is outside the range of observed values.

**Interval Estimates**: for β1 and β0: confint(M1, level = 0.95)

for mean response: e.g. estimate at X=20 and X=40



```
> new2 = data.frame(Present_Price = c(20, 40)) # two points
> predict(M1, newdata = new2, interval = "confidence", level = 0.95)
       fit      lwr      upr
1 11.05551 10.57416 11.53686
2 21.39249 20.32340 22.46158
```

fit: mean estimate. lwr, upr: lower and upper CI

**Hypothesis testing**

- In simple linear regression, there is only one F-test, and it is equivalent to the t-test. (H0: β1 = 0, H1: β1 != 0)
- t-test for β1: Test statistic is a t-statistic, t = $\hat{\beta_1}$ / SE($\hat{\beta_1}$), which can be found from the R output. (t-value of variable) Beside on the right is the p-value (Pr). small p-value -> data provide strong evidence that variable is significant
- F-test tests if whole model is significant. H0: all coefficients except intercept are zero. H1: at least one of the coefficients except intercept are nonzero. Larger F-statistic and smaller corresponding p-value means more significant (model explains more variability)
- if H0 of F-test not rejected, means regressor(s) used in model not significant, should use a new model (intercept model), Y = β0 + ε or $\hat{Y}$ = $\hat{\beta_0}$. In R, lm(dep_var~1, data = data1)

**Checking if Built Model Satisfies Assumptions**

Randomization: From the steps of data collection

Linearity: can check this assumption using a scatter plot between response Y and regressor X and the residuals plot.

Normality: is checked using the residuals of the built model.

Constant variance: checked using residuals of built model.

**After Scatterplot of Y vs X**: possible cases

1. Ideal case: Proceed with model analysis
2. Linearity assumption violated (e.g. look like quadratic): Possible fix: add higher order terms in X, e.g. $X^2$ to model.
3. Variance not constant: One possible fix: transform the response by taking ln(Y), square root ( √ Y), or the reciprocal (1/Y), to be the response of model. Transformation will change the interpretation of the coefficient $\beta_1$.

**Residual Plots**: Used to check normality assumption, check for need to add higher order terms in X, check for non-constant variance and need to transform Y

**Standardized residual (SR)** = (Y – $\hat{Y}$) / std. err. of (Y – $\hat{Y}$)

where Y is actual value, $\hat{Y}$ is predicted value from model

In R, rstandard(M1) lists standardized residuals of model M1

**What plots to make?** ri's (SR) on y-axis against $\hat{Y}$i on x-axis, SR on y-axis against X on x-axis, SR histogram, SR QQ-plot

**What to expect from plots if linear regression model is a good fit?** Plots of SR against $\hat{Y}$, X: points scatter randomly about 0, within interval (-3, 3). Histogram and QQ plot: normally distributed. Note SR from fitted model not exactly indep, but if n large, can expect show randomness.

**What issues to look out for and how to fix?**

- Funnel shape in plots of SR against $\hat{Y}$, X: constant variance assumption violated. Transform response or modify model.
- Curved band when plotting Y against X: linearity assumption violated. Modify model. - Non-normality in QQ plot: normality assumption violated. Modify model.

**Outliers**: point has SR greater than 3 or less than -3

**Influential point**: point that greatly affects parameter estimates. Outlier may or may nor be influential. Influence measured using Cook's distance, which measures effect of deleting a given observation. Poins with large Cook's distance could be influential points (can use 1 as threshold)

Using R, which(cooks.distance(M1) > 1) gives index of point

**Coefficient of determination of linear model, R-squared**: Helps check goodness of fit of the model. It is the proportion of total variation of response (about sample mean $\bar{Y}$) explained by model, is between 0 and 1, can never be 1 if there repeated X values with different Y values.

Value of R-squared = 0.7725 -> 77.25% of variation in the response variable is explained by the fitted regression

**Multiple R-squared**: Use: only one predictor or initial try

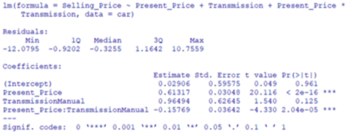
**Adjusted R-squared**: Use: when multiple predictors or comparing the fit of two models

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$ where k is num(regressor)

**R-squared and Cor(X, Y)**: In a simple model, R = √ $R^2$ = |Cor(X, Y)|. If $\hat{\beta_1}$ < 0 (resp. > 0), then Cor(X, Y) = −R (resp. Cor(X, Y) = R).

Possible to have small R-squared but significant regression (F-test very small p-value). Use **Multiple Linear Regression**

MLR vs SLR: Similarities: Regression function is linear in β's (parameters), check assumptions using residuals, t-tests for indiv coefficients of a regressor, F-test for overall regression. Differences: test for significance of a categorical variable with >2 categories, Using adjusted R^2 to compare models. Generally, Possible to have interaction between 2 variables e.g. part of eqn is β3X1 * I(X2 = Automatic)



```
lm(formula = Selling_Price ~ Present_Price + Transmission + Present_Price * 
    Transmission, data = car)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0795  -0.9202  -0.3255   1.1642  10.7559 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.02906    0.59575   0.049    0.961    
Present_Price                    0.61317    0.03048  20.116  < 2e-16 ***
TransmissionManual               0.96494    0.62645   1.540    0.125    
Present_Price:TransmissionManual -0.15769    0.03642  -4.330 2.04e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.341 on 297 degrees of freedom
Multiple R-squared:   0.79,	Adjusted R-squared:  0.7879 
F-statistic: 372.5 on 3 and 297 DF,  p-value: < 2.2e-16
```

R chooses "Manual" for the indicator variable of X2.

Fitted regression line is $\hat{Y}$ = 0.029 + 0.613X1 + 0.965 · I(X1 = Manual) − 0.158X1 · I(X2 = Manual).

Fitted regression line for Manual is $\hat{Y}$ = 0.994 + 0.455(Present_Price), and Fitted regression line for Automatic is $\hat{Y}$ = 0.029 + 0.613(Present_Price).