Statistical Inference and estimation

STAT 4101L

Surya Lamichhane

Learning Objectives

After of completion of this lesson you will be able to understand and apply:

- Point estimation
- Confidence intervals
- Hypothesis testing
- ANOVA
- Test of independence

Motivation

About 697,000 people die of heart disease in the United States every year–that's 1 in every 5 deaths. Coronary heart disease (CHD) is the most common type of heart disease, killing approximately 382,820 people annually. Every year about 805,000 Americans have a heart attack.

Data and data exploration

In [1]:
```
library(dplyr)
library(ggplot2)
health_data = read.csv("~/Desktop/STAT4101L_all_files/Stat4101L-Rfiles/DataS
knitr::kable(head(health_data[,1:5]),  "simple")
dim(health_data)
```

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke
No	16.60	Yes	No	No
No	20.34	No	No	Yes
No	26.58	Yes	No	No
No	24.21	No	No	No
No	23.71	No	No	No
Yes	28.87	Yes	No	No

319795 · 18

In [ ]:
```
unique(health_data$Race)
```

Application examples

1. Estimate the proportion of US adult population who have heart-disease. (about 6% of US adults reported having heart disease [https://www.cdc.gov/nchs/hus/topics/heart-disease-prevalence.htm]).
2. Examine the risk of heart disease among the ethnic group/gender (Black men have a 70% higher risk of heart failure compared with white men [https://my.clevelandclinic.org/health/articles/23051-ethnicity-and-heart-disease])
3. Estimate the average BMI of US adults
4. Check if there is any significant difference between the BMI of US adults by heart disease status yes/no.
5. Check if all the ethnic group have same effect on BMI of US adults.
6. Check if there is any association between heart disease and ethnicity.

Probability distributions in R

1. Binomial distribution

```
x = number of successes
n = total number of trails
p = probablity of success
pmf : p(X = x), dbinom(x, size = n, prob = p)
CDF : p(X < x), pbinom(q, size = n, prob = p), q = quantile
inv_CDF : qbinom(CDF, size = n, prob = p) computes x
random variable: rbinom(r, size = n, prob = p), r = how many
random number you want to draw
```

- Bernoulli distribution is specific case of Binomial distribution when size = 1

1. Uniform distribution

```
a = lower limit, b = upper limit
pdf : f(x; a, b) = 1/(b-a), dunif(x, a, b)
CDF : P(X < x), punif(q, a, b)
inv_CDF : qunif(CDF, a , b) computes x
random variable: runif(r, a, b), r = number of random
varaibles that we want to draw
```

---

1. Normal distribution

```
mu = mean, sd = standard deviation
pdf : f(x; mu, sd), dnorm(x, mu, sd)
CDF : P(X < x), pnorm(q, mu, sd)
inv_CDF : qnorm(p, mu, sd) computes x
random variable: rnorm(r, mu, sd), r = number of random
varaibles that we want to draw
```

In [ ]:
```
dnorm(10, mean = 74, sd = 5)
```

- mean $\mu = 0$, and $sd = 1$ yields standard normal

1. t-distribution

```
df degree of freedom
dt(x, df, ncp, log = FALSE) computes density
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE) computes CDF
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE) computes
quantile
rt(n, df, ncp) generates random number
```

---

1. chi-square distribution

```
dchisq(x, df, ncp = 0, log = FALSE)
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df, ncp = 0)
```

Inference for proportions (Using normal approximation)

- A sampling distribution summarizes the distribution of a sample statistic (e.g., mean or proportion)
- calculated from multiple samples drawn from the same population.

Example: if we want to know the average height of students in a school, we might take several random samples of students and compute the mean height for each. The distribution of these sample means constitutes the sampling distribution of the mean.

- The Central Limit Theorem (CLT) is a crucial statistical concept asserting that the sampling distribution of the sample mean approaches a normal distribution as sample size increases, irrespective of the population distribution's shape, provided the sample size is sufficiently large.
- This normal approximation is valuable as it enables making probabilistic statements about sample means even when the population distribution is unknown. It underpins various statistical inference techniques like confidence intervals and hypothesis testing.

Conditions for the Central Limit Theorem include:

- Independence: Samples must be independent, drawn randomly without influence from each other.
- Sample Size: Sample size should be "sufficiently large," often considered at least 30, though larger samples may be necessary for skewed or heavy-tailed populations.
- Population Distribution: The shape of the population distribution is irrelevant as long as the population standard deviation is finite. Extreme non-normality might require larger samples for the approximation to hold.

In practical applications, such as hypothesis testing or constructing confidence intervals for population parameters, the normal approximation provided by the Central Limit Theorem is frequently relied upon.

1. Inference for a single proportion (p) using normal approximation

Assumptions

The assumptions for inference for proportions using normal approximation typically include:

- Random Sample: The data should be collected from a random sample or a well-defined randomization process to ensure that it represents the population of interest without bias.
- Large Sample Size: The sample size should be sufficiently large. A common rule of thumb is that both np and n(1-p) should be greater than 5, where n is the sample size and p is the proportion of interest.
- Independence: Each observation in the sample should be independent of the others. The occurrence of one event should not affect the occurrence of another event in the sample.
- Approximately Normal Distribution: While the original distribution of the population may not necessarily be normal, the distribution of the sample proportion (p-hat) tends to be approximately normal for large sample sizes due to the Central Limit Theorem.

Estimation of population proportion

a. population proportion

$$p = \frac{\text{number of sucesses}}{\text{number of possible cases}}.$$

b. Point estimate: Sample proportion

$$\hat{p} = \frac{\text{number of sucesses in a sample of size n}}{\text{sample size n}}.$$

c. Sampling distribution of sample proportion $\hat{p}$ is approximately normal with mean = p and $sd = \sqrt{\frac{p\,(1-p)}{n}}$, called standard error (SE).

d. Large sample $(1-\alpha)100\%$ confidence interval for p:

- Confidence interval estimation depends on the normal approximation, so we require independent trials and a large enough sample
- Confidence interval, $\hat{p} - E < p < \hat{p} + E$ is usually written as $\hat{p} \pm E$, where E is the margin of error.
- The margin of error $E = z_{\frac{\alpha}{2}} * SE$.
- α is called the level of significance, and $z_{\frac{\alpha}{2}}$ called the critical value

e. Most widely used confidence level is 95%, but you might see 90% and 99% sometimes.

f. Interpretation of 95% confidence interval:

- We usually say we are 95% confident that the estimated confidence interval contains the true population proportion.
- The actual meaning is that when we construct the corresponding confidence intervals from each of all possible samples of same size, 95% of the several constructed confidence interval estimates would actually contain the true value of the underlying population proportion.
- Does not mean the probability that the estimated confidence interval contains the true value is 95%.

Example 1 (Point estimate and confidence interval estimate)

Estimate the proportion of US adult population who have heart-disease using the health_data.

a. Point estimate (sample proportion)

- Compute Manually

In [ ]:
```
p_hat = mean(health_data$HeartDisease == "Yes")
print(p_hat)
```

b. Confidence interval estimate

In [ ]:
```
n = nrow(health_data) # sample size
SE = sqrt(p_hat *(1-p_hat)/n)
alpha_half = (1 -.95)/2
Z_alpha_half = qnorm(1 -alpha_half)
E = Z_alpha_half * SE
L = p_hat - E; U = p_hat + E
```

In [ ]:
```
cat("95% confidence interval:", "\n")
cat("Lower limit = ", L,
    ",Upperr limit = ", U)
```

- Compute using prop.test function

Syntax

```
prop.test(x, n, p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

?prop.test( ), get help for the full documentation.

In [ ]:
```
#`@ prop.test(x,n,correct=FALSE)
#`@ x = number of successes in a given sample
#`@ n = sample size
x =  sum(health_data$HeartDisease == "Yes")
prop.test(x,n, conf.level = 0.95)
```

Interpretation

We are 95% confident that the true proportion of US adult population who have heart-disease lies in between 0.0846 and 0.0866.

Hypothesis testing (testing claim about a population proportion)

a. claim: about 8.4% of US adults reported having heart disease

b. claim: less than 8.6% of US adults reported having heart disease

c. claim: more than 8.4% of US adults reported having heart disease

1. Type of test

    - if the original claim is about equality, it goes to null, and alternative is opposite of it.
    - if the original claim is about less or greater, it goes to alternative, and null is opposite of it (many texts define null with equality only).

    a. Two tailed test:

    null H0: $p = p_0$
    alternative H1: $p \neq p_0$

    b. Left-tailed

    null H0: $p \geq p_0$
    alternative H1: $p < p_0$

    c. Right-tailed

    null H0: $p \leq p_0$
    alternative H1: $p > p_0$

1. Test statistic

    - It is a z-score calculated assuming null is true (or true value is at null)
    - It is an observed value from the data when null is true.

$$z = \frac{\hat{p} - p_0}{SE},\ SE = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

1. Critical region or rejection region:

    - This region is defined based on a prespecified cutoff called the significance level (α)
    - This region determines whether the null to be rejected or not.
    - If the test is two sided, $-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$ are the critical values that separate the rejection region of normal density.
    - If the test is left-tailed, $-z_\alpha$ is a critical value that separates the rejection region of normal density.
    - If the test is right-tailed, z_α is a critical value that separates the rejection region of normal density.

---

1. p-value:

The P-value is the probability of observing the test statistic (i.e., summary of the data) that is as extreme or more extreme than currently observed test statistic when the null hypothesis is true. This can be expressed as Pr(data|H0)

- It is a probability value, that tells you how likely your data (or test statistic) could have occurred when null hypothesis is true.
- This tells you whether the observed value (or test statistic) is an extreme value (unacceptable by the data) or not when null is true
- Two tailed test: p-value = 2*p(z > |T|)
- Left tailed test: p-value = p(z < T)
- Right tailed test: p-value = p(z > T)

---

1. Test conclusion
    - If $p - value > \alpha$, we fail to reject the null (we cannot accept the alternative but we don't say accept the null)
    - If $p - value < \alpha$, we reject the null (we can say accept the alternative)
    - If original claim is null then we either fail to reject the claim or reject the claim
    - If original claim is alternative then eitherwe do not support the claim or we suppoort the claim.

Computation of p-value In R

```
Two tailed test: p-value = 2*(1- pnorm(|z|))
Left tailed test:  p-value = pnorm(z)
Right tailed test:  p-value = 1- pnorm(z)
```

Recall:

1. estimation for population proportion (p)
    - Point estimate, Sample proportion ($\hat{p}$)
    - Confidence interval, $\hat{p} \pm E$

1. Hypothesis testing of proportion
    - Two tailed test
    - Left tailed test
    - Right tailed test

Example 2 (Hypothesis tesing for one proportion)

a. claim: about 8.4% of US adults reported having heart disease, use $\alpha = 0.05$

```
null            H0: p = 0.084
alternative     H1: p ≠ 0.084
```

In [ ]:
```
prop.test(x, n, p = 0.084,
          alternative = "two.sided")
```

Test conclusion

- p-value = 0.001156 is less than significance level 0.05
- We reject the null
- Hence, we have sufficient evidence to reject the claim that about 8.4% of US adults reported having heart disease.

b. claim: less than 8.6% of US adults have heart disease, use $\alpha = 0.05$

```
Null, H0: p >= 0.086, Alt H1: p < 0.086
```

In [ ]:
```
prop.test(x, n, p = 0.086,
          alternative = "less")
```

- p-value = 0.2082 is greater than significance level 0.05
- We fail to reject the null (we cannot support the alternative)
- Hence, we do not have sufficient evidence to support the claim that less than 8.6% of US adults have heart disease.

---

c. claim: more than 8.4% of US adults reported having heart disease, use $\alpha = 0.05$

In [ ]:
```
prop.test(x, n, p = 0.084,
          alternative = "greater")
```

- p-value = 0.0005782 is less than significance level 0.05
- We reject the null (we support the alternative)
- Hence, we have sufficient evidence to support the claim that more than 8.4% of US adults have heart disease.

Comparing two proportions (proportion of two independent groups)

Example: Examine the risk of heart disease among the ethnic groups: Black and White.

a. Confidence interval for the difference of two proportions

- Confidence interval:

$$(\hat{p_1} - \hat{p_2}) - E < p_1 - p_2 < \hat{p_1} - \hat{p_2}) + E$$

- Margin of error:

$$E = z_{\frac{\alpha}{2}} * SE, \;\; SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

b. Hypothesis test for difference of two proportions

- Test statistic

$$T = \frac{\hat{p_1} - \hat{p_2}}{SE}.$$

In [ ]:
```
p_hat = health_data %>%
  filter(Race %in% c("Black", "White"))  %>%
  group_by(Race) %>%
  summarise(sample_size = n(), successes = sum(HeartDisease == "Yes"), p_hat

knitr::kable(p_hat)
```

```
Two sided test

H0: pB -pW = 0
H1: pB -pW ≠ 0
```

In [ ]:
```
prop_diff <- prop.test(x = c(1729, 22507),
                       n = c(22939, 245212),
                       alternative = "two.sided")
# x vector of successes
# n vector of sample sizes
# Printing the results
prop_diff
```

Test conclusion

- p-value < 2.2e-16, the null is rejected
- There is sufficient evidence to reject the claim that proportion of heart disease of Black population and that of white population are aqual.

Left-tailed test

H0: pB -pW = 0
H1: pB -pW < 0

In [ ]:
```
prop_diff <- prop.test(x = c(1729, 22507), n = c(22939, 245212), alternative
# Printing the results
prop_diff
```

Test conclusion

- p-value < 2.2e-16, the null is rejected
- There is sufficient evidence to support the claim that proportion of heart disease of Black population is less than that of the white population.

Right-tailed test

H0: pB -pW = 0
H1: pB -pW < 0

In [ ]:
```
prop_diff <- prop.test(x = c(1729, 22507), n = c(22939, 245212), alternative

# Printing the results
prop_diff
```

Test conclusion

- p-value < 2.2e-16, the null is rejected
- There is no sufficient evidence to support the claim that proportion of heart disease of Black population is greater than that of the white population.

Inference for Means (Using normal approximation)

1. Estimation of population mean from sample data

- The best point estimate for population mean μ is sample mean $\bar{X}$.
- The confidence interval estimate for mean μ is $\bar{X} - E < \mu < \bar{X} + E$, where
  - For known variance, $E = z_{\frac{\alpha}{2}} * SE$, and $SE = \frac{\sigma}{\sqrt{n}}$ (from normal distribution).
  - For unknown variance, $E = t_{\frac{\alpha}{2}, df=n-1} * SE$, and $SE = \frac{s}{\sqrt{n}}$ (from t-distribution). (**Note: This is the usual case.**)
- The confidence interval can be written as $\bar{X} \pm E$.

- Example:

a. Estimate the average BMI of US adults

In [2]:
```
avg_BMI = mean(health_data$BMI)
print(avg_BMI)
```

```
[1] 28.3254
```

b. Construct a 95% confidence interval for the mean BMI of US adults.

- Manual computation

In [ ]:
```
X_bar = mean(health_data$BMI)
s = sd(health_data$BMI)
n = nrow(health_data)
alpha = 0.05
alpha_half = alpha/2
t = 1 - qt(alpha_half, df = n-1)
SE = s/sqrt(n)
E = t * SE
L = X_bar - E; U = X_bar + E
cat('95% confidence interval is', '\n')
cat('Lower limit: ', L, ',', 'Upper limit: ', U)
```

Interpretation

We are 95% confident that the population BMI of US adults lies in between 28.29213 and 28.35867. Here 95% confidence means, among all possible samples of same size n, 95% of such constructed estimated intervals actually contain the true mean and our interval is one of those possible interval. Therefore, we say we are 95% confidence.

- Using t-test function

In [3]:
```
x = health_data$BMI
t.test(x, conf.level = 0.95)
```

```
        One Sample t-test

data:  x
t = 2520.1, df = 319794, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 28.30337 28.34743
sample estimates:
mean of x
  28.3254
```

2. Hypothesis testing (testing claim about a population mean)

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

According to the National Health and Nutrition (NHN) Survey, the average adult population of the United States has a BMI of around 28.29.

a. Test whether the BMI of health data supports the above claim, use significance level equals 0.1.

$$\text{null } H0 : mu = 28.29,$$
$$\text{alternative } H1 : mu \neq 28.29$$

In [4]:
```
x = health_data$BMI
t.test(x, alternative = "two.sided", mu = 28.29)
```

```
        One Sample t-test

data:  x
t = 3.1494, df = 319794, p-value = 0.001636
alternative hypothesis: true mean is not equal to 28.29
95 percent confidence interval:
 28.30337 28.34743
sample estimates:
mean of x
  28.3254
```

Test conclusion

- p-value = 0.001636 and given significance level is 0.1
- Since p-value is less than significance level we reject the null
- Hence, we have sufficient evidence to reject the claim that average BMI of US adults is 28.29.

b. Test the claim that BMI of Black US adults is greater than 30.1, use significance level equals 0.05.

```
null            H0: mu = 30.1
alternative     H1: mu > 30.1
```

In [5]:
```
x = health_data %>% filter( Race == "Black") %>% select(BMI)
x = unlist(x)
t.test(x, alternative = "greater", mu = 30.1)
```

```
        One Sample t-test

data:  x
t = 1.4728, df = 22938, p-value = 0.07041
alternative hypothesis: true mean is greater than 30.1
95 percent confidence interval:
 30.09182      Inf
sample estimates:
mean of x
 30.16999
```

Test conclusion

- p-value = 0.07041 is greater than the sig level = 0.05.
- We fail to reject the null.
- We do not have sufficient evidence to support the claim that BMI of Black US adults is greater than 30.1

c. Test the claim that BMI of Asian US adults is less than 25.3. use significance level equals 0.05.

```
null            H0: mu = 25.3
alternative     H1: mu < 25.3
```

In [ ]:
```
x = health_data %>% filter( Race == "Asian") %>% select(BMI)
x = unlist(x)
t.test(x, alternative = "less", mu = 25.3)
```

Test conclusion

- p-value = 0.0662 is greater than the sig level = 0.05.
- We fail to reject the null.
- We do not have sufficient evidence to support the claim that BMI of Asian US adults is less than 25.3.

Comparing the means of two independent groups (two independent samples t-test)

- Independent samples:
  - Cases of each group are unrelated (or are not naturally paired/matched) to one another.
  - Not necessarily of same sizes
- Dependent samples:
  - Cases of each group are somehow related (or are naturally paired/matched) to one another.
  - Always of same sizes

- The two-sample t-test (aka **independent samples t-test**) is a method used to test whether the unknown population means of two independent groups are equal or not.

- **Requirements(or assumptions):**

  - data values are independent, are randomly sampled from two normal populations
  - two independent groups have equal variances (Homogenity of varaince). Violations of this assumption can affect the accuracy of the test, especially when sample sizes are unequal (Welch's t-test is applied in this case).

1. Confidence interval for difference between the means of two groups.

```
  a. Provide the point estimate of the difference between the
mean BMI of male and female population.

  b. Construct $95\%$ confidence interval estimate of the
difference between the mean BMI of male and female
population.
```

In [6]:
```
gender_mean_BMI = health_data %>%
                    group_by( Sex) %>%
                    summarise(sample_mean = mean(BMI))

knitr::kable(gender_mean_BMI)
```

Sex	sample_mean
Female	28.16244
Male	28.50532

In [7]:
```
# condidence interval for difference of means (independent sample t-test) fo
male_BMI = health_data %>%
       filter(Sex == "Male") %>%
        select(BMI)

female_BMI = health_data %>%
         filter(Sex == "Female") %>%
        select(BMI)

t.test(x = male_BMI, y = female_BMI, conf.level = 0.95)
```

```
        Welch Two Sample t-test

data:  male_BMI and female_BMI
t = 15.368, df = 318168, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2991518 0.3866106
sample estimates:
mean of x mean of y
 28.50532  28.16244
```

2. Hypothesis testing for the mean differences of two groups.

a. Test the claim that there is no significant difference between the BMI of Asian and Hispanic population.

null	H0: mu_A = mu_H
alternative	H1: mu_A ≠ mu_H

In [8]:
```
A_BMI = health_data %>% filter( Race == "Asian") %>% select(BMI)
H_BMI = health_data %>% filter( Race == "Hispanic") %>% select(BMI)
t.test(x = A_BMI, y = H_BMI, alternative =  "two.sided")
```

```
        Welch Two Sample t-test

data:  A_BMI and H_BMI
t = -53.612, df = 17438, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.731395 -3.468170
sample estimates:
mean of x mean of y
 25.21830  28.81809
```

Conclusion: We reject the claim that there is no significant difference between the BMI of Asian and Hispanic population.

b. Test the claim that White population has lower BMI than the that of Black population in average.

null	H0: mu_W = mu_B
alternative	H1: mu_W < mu_B

In [9]:
```
B_BMI = health_data %>% filter( Race == "Black") %>% select(BMI)
W_BMI = health_data %>% filter( Race == "White") %>% select(BMI)
t.test(x = W_BMI, y = B_BMI, alternative =  "less")
```

```
        Welch Two Sample t-test

data:  W_BMI and B_BMI
t = -41.036, df = 26210, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.935738
sample estimates:
mean of x mean of y
 28.15342  30.16999
```

Test Conclusion

- p-value < 2.2e-16 is less than 0.05.
- We support the claim that White population has lower BMI than that of Black population in average.

c. Test the claim that Asian population has lower BMI than the that of white population in average.

null H0: mu_A = mu_W
alternative H1: mu_A > mu_W

In [ ]:

3. Comparing results from two dependent groups (pairwise t-test)

- The pairwise t-test is a method used to compare the means of two measurements taken from the same individual, object, or naturally paired units.

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = TRUE)
```

Example 1:

We want to compare the height of the married couples. You selected a sample of 10 couples and want to compare the height of couples to determine if there is a statistically significant difference between them.

H0: mu_d = 0
H1: mu_d is not equal 0

In [10]:
```
Woman_height = c(5.3, 4.9, 5.5, 5.2, 5.7, 5.4, 5.9, 5.6, 5.0, 5.5)
Husband_height = c(5.6, 5.9, 5.4, 5.7, 5.6, 5.9, 6.39, 5.9, 5.3, 5.2)
d = Husband_height - Woman_height
cat("difference of the height of married couples")
round(d,2)
```

```
difference of the height of married couples
```
0.3 · 1 · -0.1 · 0.5 · -0.1 · 0.5 · 0.49 · 0.3 · 0.3 · -0.3

In [11]:
```
t.test(x = Woman_height, y = Husband_height,
       alternative = "two.sided",
      paired = TRUE)
```

```
        Paired t-test

data:  Woman_height and Husband_height
t = -2.4187, df = 9, p-value = 0.03869
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.55929281 -0.01870719
sample estimates:
mean difference
         -0.289
```

Example 2

We want test whether the test score is significantly improved in test two than test one. A comparison is made based on test scores of 35 randomly selected students from a Statistics class.

Example 3

A pharmaceutical company is conducting a clinical trial to compare the effectiveness of two different drugs (Drug A, and Drug B) in treating hypertension (high blood pressure). The company wants to determine if there are any statistically significant differences in blood pressure reduction between the two drugs.

Analysis of Variance (ANOVA)

When a dependent (quantitative) variable is further categorized into multiple treatments (or factors), and we want to compare the effect of those treatment groups on the given dependent variable we use ANOVA method.

Example 1 (One way ANOVA)

We want to investigate how the BMI differs between the various races of the US population.

Example 2 (Two way ANOVA)

We want to examine the effect of irrigation and sunlight exposure on crop yield

- t-test examines whether means of two groups are different or not,
- while ANOVA can be used for more than two groups.

One way ANOVA

When the dependent variable is categorized (or describe) according to one way (by one treatment/factor) into more than two independent groups, and we want to examines the effects (or means) of those independent groups on the dependent variable, we use one way ANOVA.

Example 1 (One way ANOVA)

We have a sample data that describes BMI of randomly selected 500 US adults which is divided into multiple groups by race, and first 25 samples is given in 5 rows below:

BMI by race

In [12]:
```
# simulated sample data
df = data.frame(White = runif(5, 20, 28),
                Black = runif(5, 21, 30),
                Asian = runif(5, 18, 25),
                Hispanic = runif(5, 21, 29),
                Others = runif(5, 23, 30)
)
df = round(df, 1)
knitr::kable(df)
```

White	Black	Asian	Hispanic	Others
22.7	27.5	19.2	21.4	26.8
25.6	23.9	22.1	23.2	25.4
21.9	23.3	22.2	22.2	28.1
26.7	23.9	23.8	24.6	29.2
20.3	26.1	24.9	21.4	29.8

a. Here the response variable is BMI, which is divided by one factor/treatment race (with five categories/groups).

b. If we want to compare mean BMIs of two races, we use two-independent sample t-test,

c. If we want to compare means of more than two races, we use one way ANOVA.

- Why don't we use pair-wise t-test?

  a. Computational complexity: comparing 5 groups, we require $C(5,2) = 10$ pair-wise t-test.

  b. Problem with type I error: when number of pair increases, type I error increases exponentially, so we have a problem of rejecting the null in most of the cases.

One way ANOVA testing framework

For k-many groups,

$$\text{Null} \quad H_0 : \mu_1 = \mu_2 = \ldots, \mu_k, \ (\text{all the means are same})$$
$$\text{Alternative} \quad H_1 : \ \text{at least two means are unequal/ at least one mean is different fro}$$

- Why do we define alternative in this way instead of saying all the means are different?
  - Alternative is a statement that contradicts null, so we consider a weak statement that is enough to contradict the null. (note: if we say all the means are different, it is vary strong, we do not use it as an alternative).
  - It means that some can be equal but not all.
  - The alternative hypothesis in one-way ANOVA is designed to capture the general idea that there are differences among the group means, without restricting it to a specific pattern of differences.

- Test conclusion based on p-value and α
    - If p-value is greater than α, we fail to reject the null, (so it is reasonable to say all means are equal)
    - If p-value is less than α, we reject the null, so we have sufficient evidence to say that at least two means are different.

- Does ANOVA suggest us which means are different and how they are different?
    - If the null of one way ANOVA rejected, we can say at least two means are different, but ANOVA does not suggest which pairs are different and how they differ.
    - To find such pairs we need to use pair-wise comparision, for this course we use **modified pairwise comparision t-test**.

- Does ANOVA requires equal sample size for each group?
    - No, ANOVA does not require equal sample sizes.

Requirements:

```
- The population have distribution that are approximately
normal
- The population for each group has same varaince (or same
standard deviation)
- The samples are SRS and independent among the groups
- Data is categorized in only one way (Treatment/factor)
```

Test Statistics

- F-statistic $F = \frac{MS(Treatment)}{MS(Error)} \approx \frac{\text{variance between sample means}}{\text{variance within samples}}$
    - F-statistic is the ratio of two varainces: variance between the groups and variance within groups.
    - If all the means are approximately equal, then numerator becomes small and so is F-statistic, and vice versa,
    - If means are different, the numerator becomes large and so is F-statistic, and vice versa
    - Hence, smaller F-score supports Null, while larger value supports alternative.

- Why do we call this method Analysis of variance even though we are comparing the means?
    - Even though our goal of study is to comapre the means between the groups, our test conclusion is based on comparing the varaince between the gorups and within the groups.
    - Hence, we refer this method as Analysis of varaince.

Test conclusion of ONE WAY ANOVA

1. $p - value < \alpha$
    - Initial conclusion: if p-value is less than α we reject the null (i.e. accept alternative)
    - Formal conclusion: We have sufficient evidence to say that at least one mean is different from others.
    - In this case we can say at least two groups have different mean (different effect).
2. $p - value \geq \alpha$
    - Initial conclusion: if p-value is greater than α we fail to reject the null
    - Formal conclusion: There is insufficient evidence to reject the claim that all the means are same.
    - In this case we might think of all the groups have same mean (same effect).

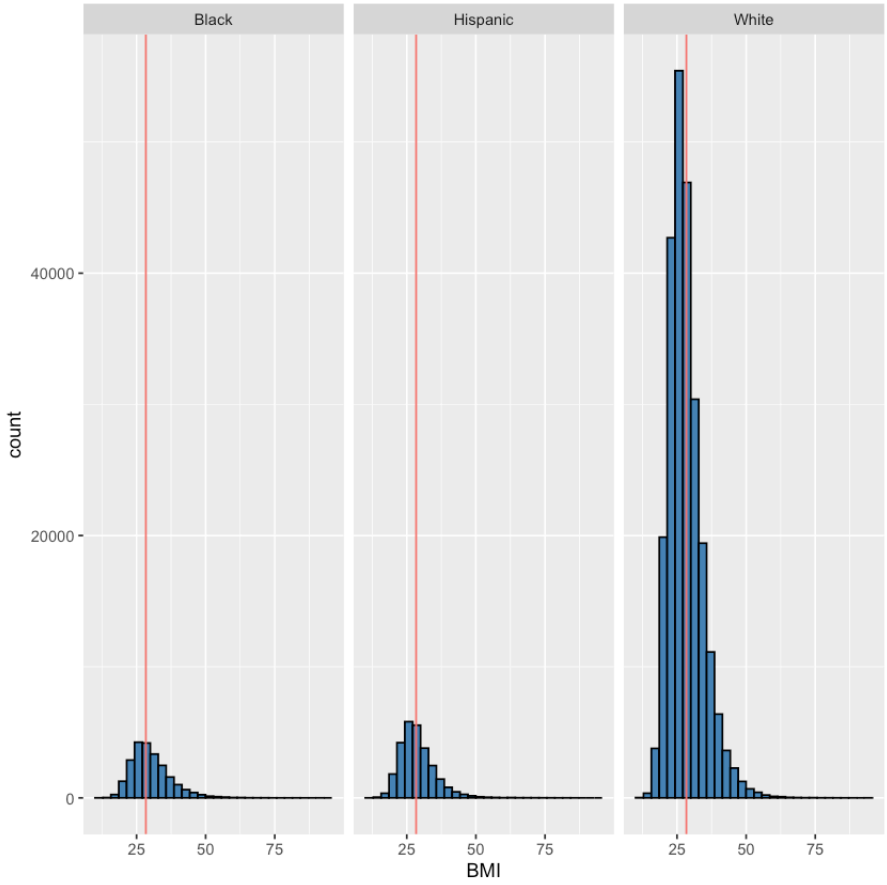
One-Way Anova in R

- Test the effect of Race on BMI's, i.e. examine whether the BMI differs by race among Black, White and Hispanic.

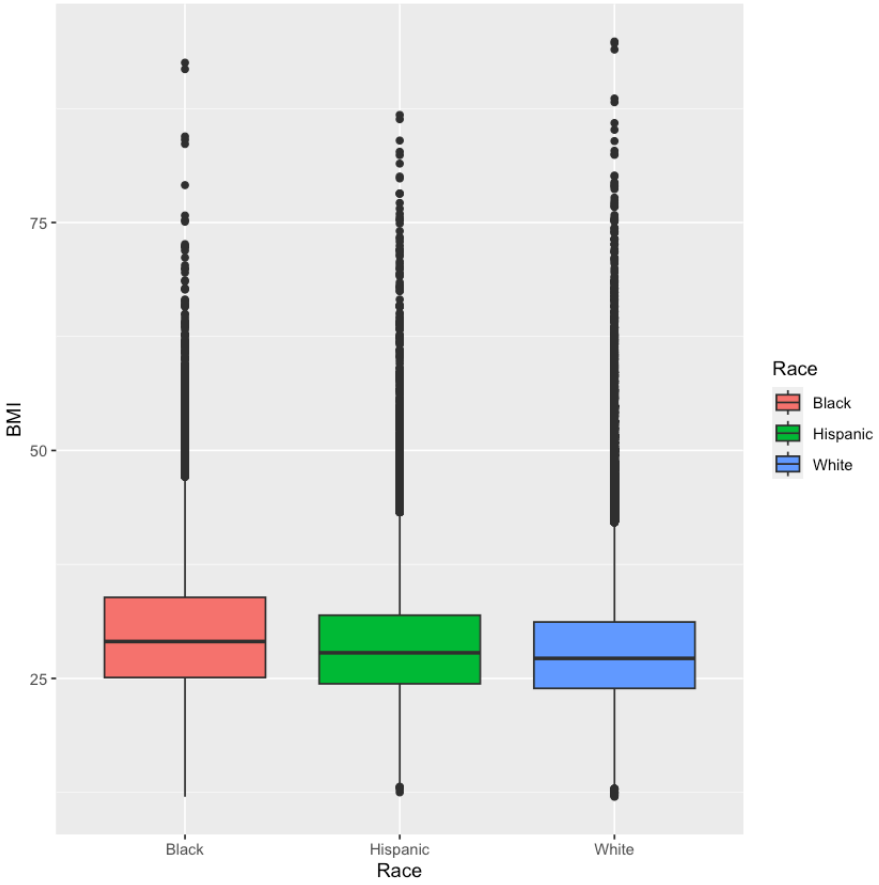
In [13]:
```
library(ggplot2)
df <- health_data %>% filter(Race %in% c('Black', 'White', 'Hispanic'))
stat_summary <- df %>%
  group_by(Race) %>%
  summarise(mean_BMI = mean(BMI))

ggp <- ggplot(df, aes(x=BMI))+
  geom_histogram(color="black", fill="steelblue")+
  geom_vline(aes(xintercept=mean(BMI), col = 'red'), show.legend = FALSE) +
  facet_wrap(Race ~ .)
ggp
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In [14]:
```
# Change box plot colors by groups
p<- ggplot(df, aes(y=BMI, x=Race,  fill= Race)) +
   geom_boxplot()
p
```

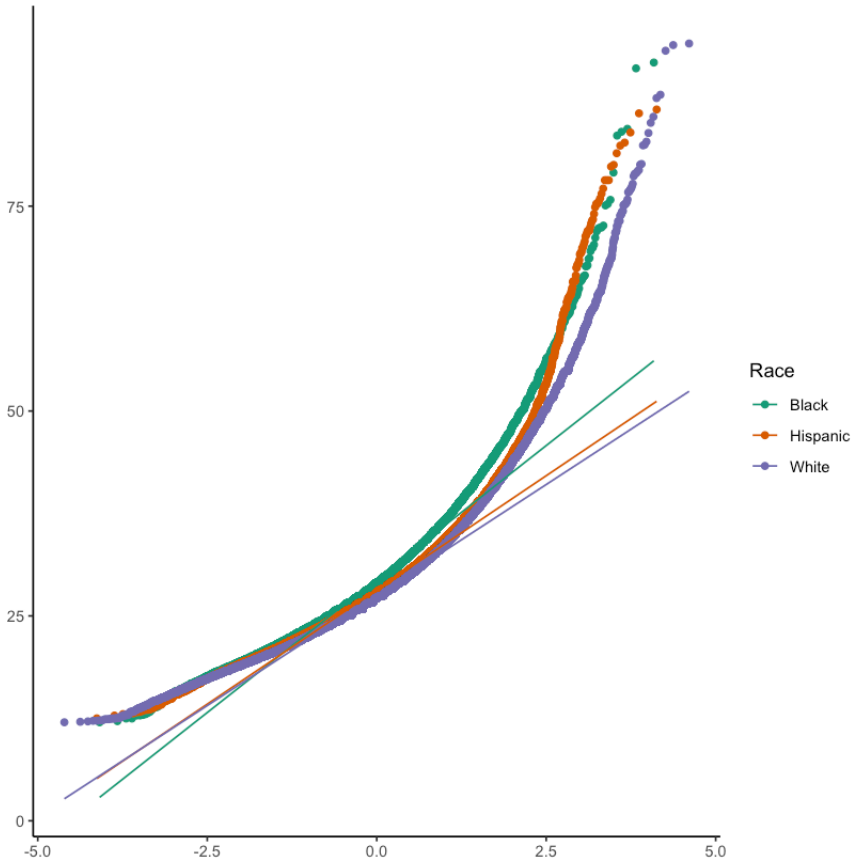


Normality using quantile-quantile plot

If the data is normally distributed, the points in a Q-Q plot will lie on a straight diagonal line.

In [15]:
```
# Change qq plot colors by groups
ggp<-qplot(sample = BMI, data = df, color=Race)
ggp +
  stat_qq_line() +
  scale_color_brewer(palette="Dark2") +
  theme_classic()
```

```
Warning message:
“`qplot()` was deprecated in ggplot2 3.4.0.”
```



In [16]:
```
Anova1 <- aov(BMI ~ Race, data = df)
summary(Anova1)
```

```
                Df    Sum Sq Mean Sq F value Pr(>F)
Race             2     91334   45667    1148 <2e-16 ***
Residuals   295594 11760469      40
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of the result of one-way ANOVA tests

As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the groups means of BMI.

Multiple pairwise-comparison between the means of groups

In one-way ANOVA test, rejecting null indicates that some of the group means are different, but we don't know which pairs of groups are different.

It's possible to perform multiple pairwise-comparison, to determine if the mean difference between specific pairs of group are statistically significant.

Tukey multiple pairwise-comparisons

We can compute Tukey HSD (Tukey Honest Significant Differences, R function: TukeyHSD()) for performing multiple pairwise-comparison between the means of groups.

- The function TukeyHSD( ) takes the fitted ANOVA as an argument.

In [17]:
```
TukeyHSD(Anova1)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = BMI ~ Race, data = df)

$Race
                     diff        lwr        upr p adj
Hispanic-Black -1.3519078 -1.4841562 -1.2196593     0
White-Black    -2.0165711 -2.1186411 -1.9145011     0
White-Hispanic -0.6646634 -0.7587581 -0.5705686     0
```

The above result shows that all the means are different.

Two-Way ANOVA

Recall: ANOVA stands for analysis of variance and it tests the effects of various treatment group on a quantitative dependent variable.

- When a dependent variable is further categorized by two different treatments (or factors) and we want to examine the effect of two treatments (or factors) on the dependent variable, we use two way ANOVA.
- A two-way ANOVA tests the effect of two independent variables on a dependent variable.
- It also examines whether the dependent variable is affected by the interaction of two factors (or treatment) or not.

Example 1 (Two way ANOVA)

We have a sample data that describes BMI of randomly selected 500 US adults which is divided into multiple groups by race, and gender as follows:

```
- Here, response BMI
- factor (or treatment ), race
- factor (or treatment ), gender
```

BMI by Race and Gender

In [18]:
```
# construct an example of sample data
set.seed(0)
df1 = cbind(White = runif(45, 20, 28),
                Black = runif(45, 21, 30),
                Asian = runif(45, 18, 27),
                Hispanic = runif(45, 21, 29)
)
rownames(df1) = rep("Male", 45)

df2 = cbind(
                White = runif(45, 20, 28)-1,
                Black = runif(45, 21, 30)-1,
                Asian = runif(45, 18, 25)-1,
                Hispanic = runif(45, 21, 29)-1
)

rownames(df2) = rep("Female", 45)

df = rbind(df1, df2)
df = round(df, 1)
knitr::kable(as.table(head(df,5)))
```

	White	Black	Asian	Hispanic
Male	27.2	25.8	19.3	28.4
Male	22.1	28.1	20.2	25.8
Male	23.0	21.2	18.5	25.5
Male	24.6	25.3	23.8	25.2
Male	27.3	27.6	25.9	28.9

Mean BMI by race

In [19]:
```
tt = colMeans(df)
knitr::kable(tt)
```

	x
White	23.75000
Black	24.95222
Asian	20.96222
Hispanic	24.77667

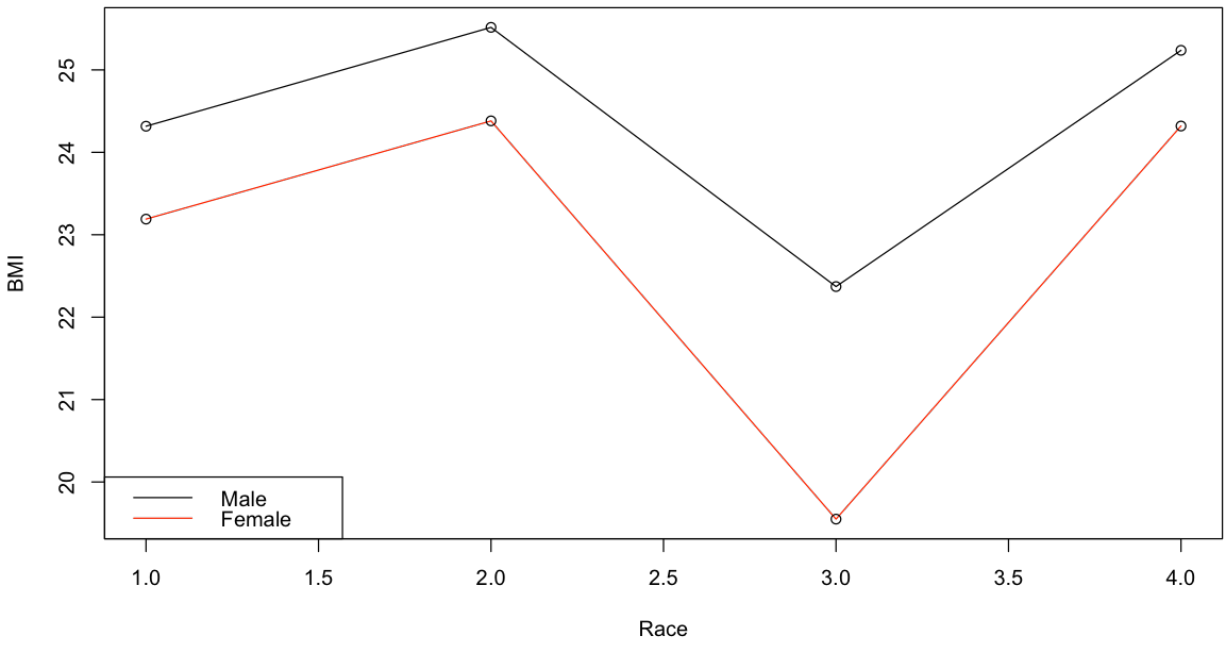
Mean BMI by gender

In [20]:
```
tt = c(mean(df1), mean(df2))
names(tt) = c("Male", "Female")
knitr::kable(tt)
```

	x
Male	24.36048
Female	22.86005

Check Interaction

In [21]:
```
t1 = colMeans(df1)
t2 = colMeans(df2)
options(repr.plot.width = 10, repr.plot.height = 6)
plot(1:4, t1, 'l', ylim = range(t1, t2), xlab = "Race", ylab = "BMI")
points(1:4, t1)
lines(1:4, t2, 'l', col = 'red')
points(1:4, t2)
legend(x = "bottomleft",
       #inset = c(-.2, 0), # You will need to fine-tune the first
                           # value depending on the windows size
       legend = c("Male", "Female"),  # Legend texts
       #lty = c(1, 2),
       col = c('black', 'red'), # Line colors
       lwd = 1,
       xpd = TRUE) # You need to specify this graphical parameter to
                   # put the legend outside the plot
```



Example 2 (Two way ANOVA)

We want to examine the effect of irrigation and sunlight exposure on crop yield

Paddy yield per hectare(in tons) by Sunlight exposure (in column) and Irrigation (in row)

In [22]:
```
set.seed(123)
df1 = cbind(
                Low = runif(45, 2, 4),
                Midium = runif(45, 2.5, 5),
                High = runif(45, 2.1, 4.5)
)
rownames(df1) = rep("Low", 45)

df2 = cbind(
                Low = runif(45, 2, 3),
                Midium = runif(45, 3, 5),
                High = runif(45, 3.5, 5.5)
)
rownames(df2) = rep("High", 45)

df = rbind(df1, df2)
df = round(df, 1)
knitr::kable(head(df[sample(90, 20),],10))
```

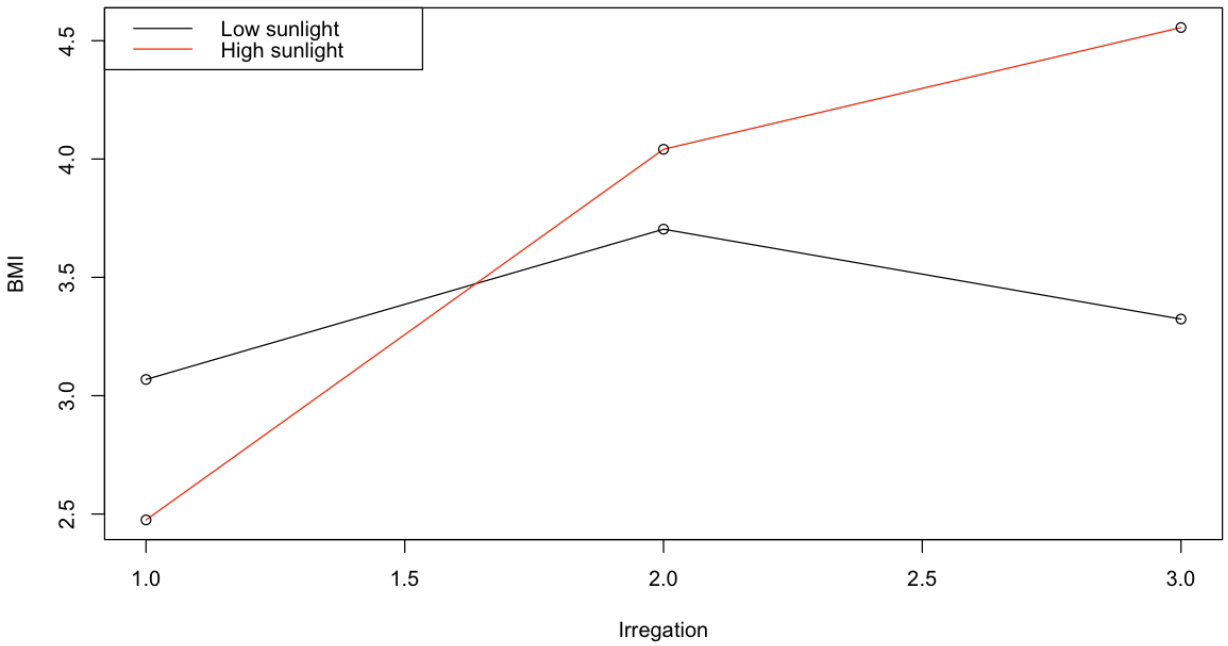
	Low	Midium	High
Low	2.6	2.5	3.5
Low	2.9	3.9	3.3
High	2.0	4.1	4.6
High	2.2	4.8	3.6
High	2.7	3.5	3.8
Low	3.9	3.0	3.5
Low	3.3	3.6	3.8
High	2.4	3.5	4.4
Low	3.4	4.4	2.4
Low	3.1	3.6	4.0

In [23]:
```
# Let convert  row-factor and column-factor in to two columns
d1 = c(runif(45, 2, 4), runif(45, 2.5, 5),runif(45, 2.1, 4.5))
d2 = c(runif(45, 2, 3), runif(45, 3, 5),runif(45, 3.5, 5.5))
irreg_lev = rep(c('low', 'medium', 'high'), each = 45)
sun_exp = rep(c('low', 'high'), each = 135)
df_yield = data.frame(sun_exp, irreg_lev, production = c(d1, d2))
subset_df = head(df_yield[sample(nrow(df_yield), 10), ], 10)
rownames(subset_df) = NULL
knitr::kable(subset_df)
```

sun_exp	irreg_lev	production
high	low	2.687743
high	high	5.006220
low	low	2.069750
high	low	2.093917
low	medium	2.841350
high	medium	3.173880
low	low	2.829471
high	low	2.000465
low	medium	3.547290
high	medium	3.929428

- Here, response is rice production (in metric tons)
- factor1 (or treatment1)is Sunlight exposure
- factor2 (or treatment2) is Irrigation
- Interaction term is Sunlight exosure * Irrigation

In [24]:
```
set.seed(123)
t1 = colMeans(df1)
t2 = colMeans(df2)
plot(1:3, t1, 'l', ylim = range(t1, t2), xlab = "Irregation", ylab = "BMI")
points(1:3, t1)
lines(1:3, t2, 'l', col = 'red')
points(1:3, t2)
legend(x = "topleft",
       #inset = c(-.2, 0), # You will need to fine-tune the first
                           # value depending on the windows size
       legend = c("Low sunlight", "High sunlight"),  # Legend texts
       #lty = c(1, 2),
       col = c('black', 'red'), # Line colors
       lwd = 1,
       xpd = TRUE) # You need to specify this graphical parameter to
                   # put the legend outside the plot
```



Interaction effect

- Interaction effects represent the combined effects of factors on the dependent measure.
- When an interaction effect is present, the impact of one factor depends on the level of the other factor.
- ANOVA has the ability to test interaction effects.

Therefore, in two way ANOVA we examine the following:

a. Interaction effect

- If we found any significance of interaction effect, then we considering an interaction term in regression model is significant ($y = \beta_0 + \beta_1 X_{treat_1} + \beta_2 X_{treat_2} + \gamma X_{treat_1} * X_{treat_2}$.)
- If there is no evidence of interaction effect, then we can only consider an additive model ($y = \beta_0 + \beta_1 X_{treat_1} + \beta_2 X_{treat_2}$.)

In our examples: example 1, shows that there is no interaction effect between row factor and column factor, while in example 2, we see the interaction effect. It is quite reasonable that if the sunlight exposure is low, high irrigation can have adverse effect in rice production, while for the high sunlight exposure, high irrigation can have positive

effect in production. Hence, impact of irrigation depends on the sunlight exposure. So, we can say there is an interaction effect.

b. Factor1 (Row) effect

- We test whether the variability of groups in Factor1 is significant or just occurred by chance.
- having Factor1 effect means at least one group has different effect (or mean) than others

c. Factor2 (Column) effect

- We test whether the variability of groups in Factor2 is significant or just occurred by chance.
- having Factor2 effect means at least one column group has different effect (or mean) than others

- **Note that effect of factor1 and factor 2, i.e. $\beta_1 X_{treat_1} + \beta_2 X_{treat_2}$ is jointly called main effect**

Two Way ANOVA Test

Requirements:

- Population distribution of each group is approximately normal
- Population have the same variance (homogeneity of variance)
- The samples ar SRS
- The samples are from populations that are categorized in two ways

Hypothesis tesing

Step 1: Test for Main effect (i.e row effect and column effect)

a. Test for Row effect

$$\text{Null } H_0: \text{ there is no row effect}$$
$$\text{Alternative } H_1: \text{There is a row effect}$$

- Finds the F-statistic and p-value of the row effect
- If p-value < alpha, we have sufficient evidence to say that there is a row effect
- Otherwise, we have insufficient evidence to reject the claim that all the groups have same effect (there is no row effect).

b. Test for column effect

$$\text{Null } H_0: \text{ there is no column effect}$$
$$\text{Alternative } H_1: \text{There is a column effect}$$

- Finds the F-statistic and p-value of the column effect
- If p-value < alpha, we have sufficient evidence to say that there is a column effect
- Otherwise, we have insufficient evidence to reject the cliam that all the groups have same effect (there is no column effect).

Step 2: Test for interaction effect

$$\text{Null } H_0: \text{ there is no interaction effect}$$
$$\text{Alternative } H_1: \text{There is an interaction effect}$$

- Finds the F-statistic and p-value of the interaction effect
- If p-value < alpha, we have sufficient evidence to say that there is an interaction effect, then interaction model is significant .
- If p-value > < alpha, we have insufficient evidence to say that there is an interaction effect, then additive model is well enough to model the response.

Two-way ANOVA in R

In [25]:
```
anova2 = aov(production ~ sun_exp + irreg_lev + sun_exp * irreg_lev, data =
summary(anova2)
```

```
                  Df Sum Sq Mean Sq F value   Pr(>F)
sun_exp            1   8.12    8.12   23.02 2.68e-06 ***
irreg_lev          2  82.54   41.27  117.04  < 2e-16 ***
sun_exp:irreg_lev  2  41.53   20.76   58.89  < 2e-16 ***
Residuals        264  93.09    0.35
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Test of Independence

Learning Objectives

After of completion of this lesson you will be able to understand and apply:

- Test of independence of two categorical variables

Chi-Square test of Independence

- The primary use of the chi-square test is to examine whether two categorical variables are independent or not
- The t-test for independent means compares means of two quantitative (i.e. numerical) variables,
- Chi-Square test of Independence compares the categorical variables
- Expected cell frequency = $\frac{(\text{row total})*(\text{column total})}{\text{grand total}}$
- $\chi^2 = \sum \frac{(\text{Observe freq - Expected freq})^2}{\text{Expected freq}}$

H0: In the population, the two categorical variables are independent.
H1: In the population, the two categorical variables are dependent.

In [26]:
```
health_data %>%
    select(HeartDisease,Smoking) %>%
    table()
```

```
             Smoking
HeartDisease     No    Yes
         No  176551 115871
         Yes  11336  16037
```

Example 1.

Test if there is any association between HeartDisease and Smoking.

1. Create a contingency table

In [28]:
```
ConTable = health_data %>%
    select(HeartDisease,Smoking) %>%
    table
ConTable
```

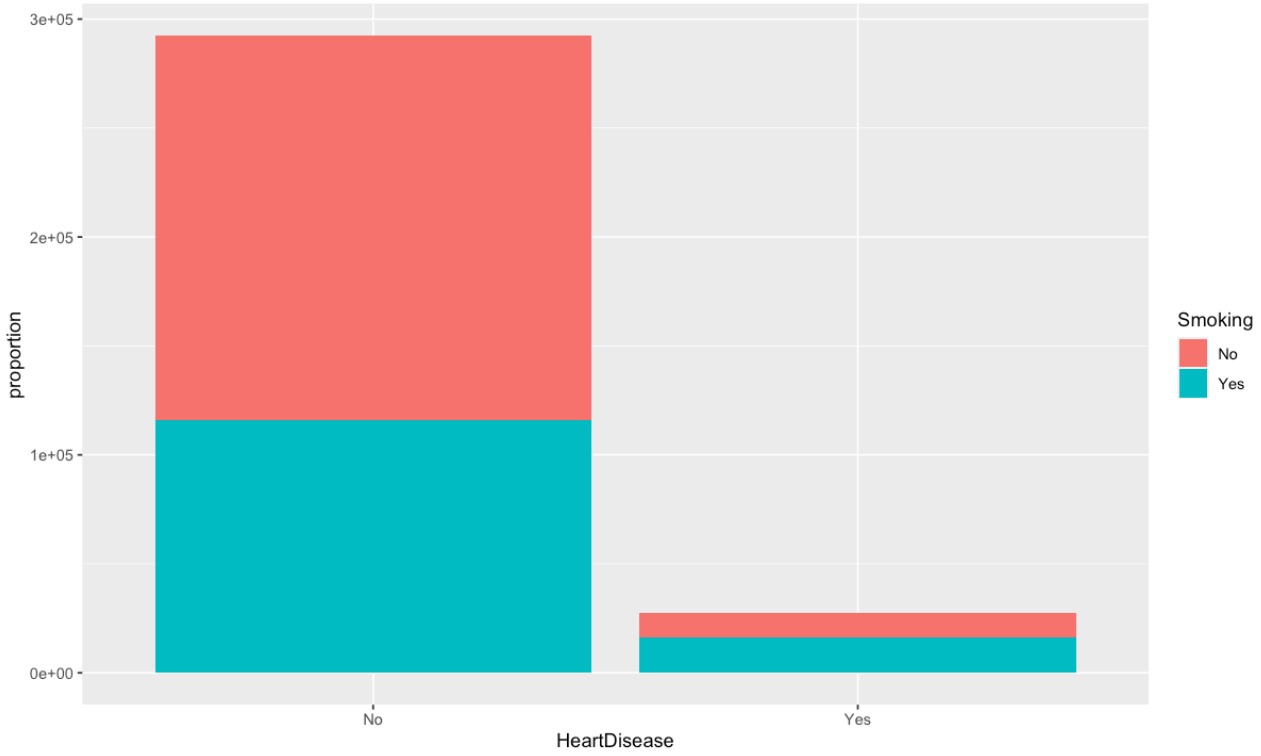
```
             Smoking
HeartDisease     No    Yes
         No  176551 115871
         Yes  11336  16037
```

In [29]:
```
propTable = round(prop.table(ConTable),2)
propTable
```

```
             Smoking
HeartDisease   No  Yes
         No  0.55 0.36
         Yes 0.04 0.05
```

Exploratory visualization: stacked bar plot

In [30]:
```
ggplot(health_data, aes(x = HeartDisease, fill = Smoking)) +
geom_bar() +
xlab("HeartDisease") +
ylab("proportion")
```



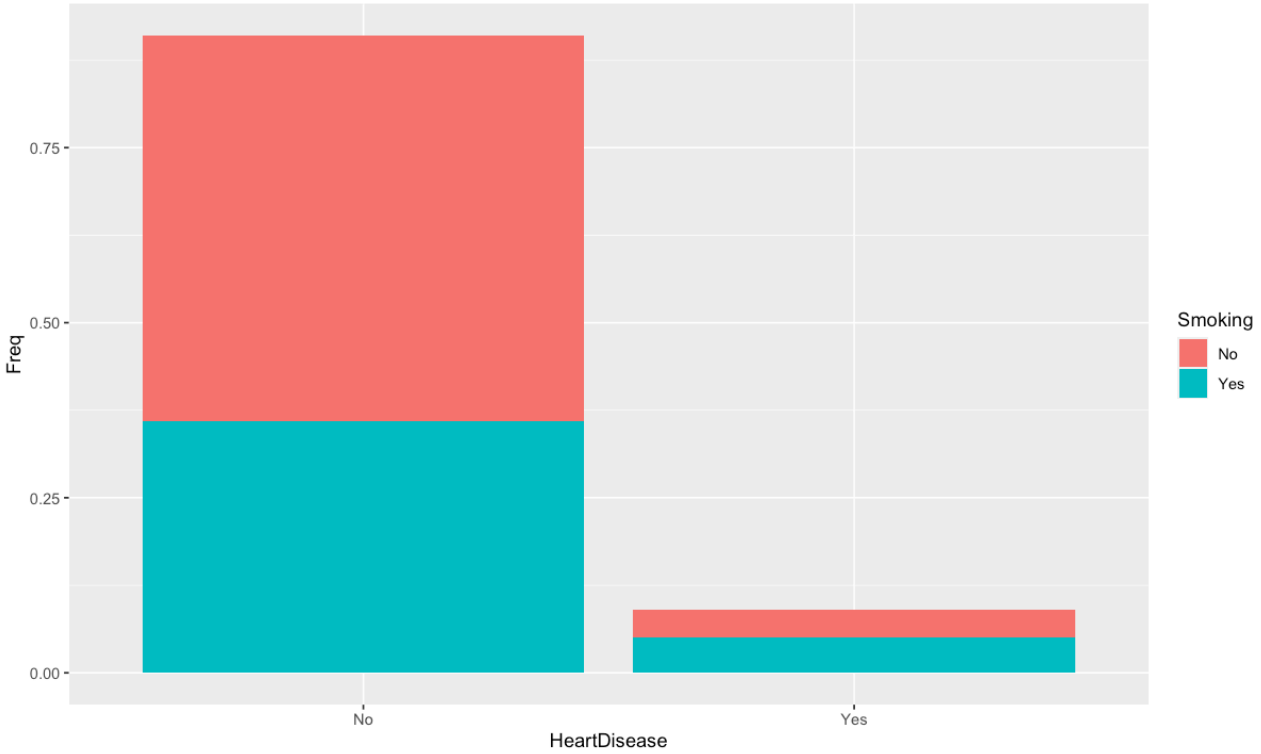
Exploratory visualization: stacked bar plot of proportion table

In [31]:
```
tab = as.data.frame(propTable)
tab
```

A data.frame: 4 × 3

HeartDisease	Smoking	Freq
<fct>	<fct>	<dbl>
No	No	0.55
Yes	No	0.04
No	Yes	0.36
Yes	Yes	0.05

In [32]:
```
ggplot(tab, aes(x = HeartDisease, y = Freq, fill = Smoking)) +
 geom_col()
```



Test if there is any association between HeartDisease and Smoking.

```
Null        H0: HeartDisease and Smoking are independent.
Alternative H1: HeartDisease and Smoking are dependent.
```

In [ ]:
```
#ConTable = table(HeartDisease = health_data$HeartDisease, Smoking = health_
test = chisq.test(ConTable)
test
```

- Observation: $p - value < 2.2e - 16$ which is much less than alpha of 0.05. Hence, we reject the null hypothesis
- Conclusion: HeartDisease and smooking are related.

Expected Frequencies

In [ ]:
```
test$expected
```

Observed Frequencies

In [ ]:
```
test$observed
```

Good Luck!

In [ ]: