

STAT 3011 Study Guide

October 7, 2022

Chapter 1

1. *Population and sample:*
 - (a) Population: A collection of all units in interest.
 - (b) Subject: An individual unit of a population.
 - (c) Sample: An observed subset of the units of a population.
2. *Parameter and statistic:*
 - (a) Parameter: A number that describes a population, which is usually unknown.
 - (b) Statistic: A number that describes a sample. It must be computable from the sample, and therefore is known.
3. *Statistical inference* is the procedure of using a sample to learn about the population.

Chapter 2

4. Types of variables:
 - (a) Categorical
 - (b) Quantitative
 - i. Discrete
 - ii. Continuous
5. Graphical summaries for categorical variables:
 - (a) Pie chart
 - (b) Bar plot
6. Graphical summaries for quantitative variables:
 - (a) Stem-and-leaf plot
 - (b) Histogram
 - (c) Box plot
7. Finding the center and spread of a quantitative data from its graphical summaries:
 - (a) Stem-and-leaf plot and histogram:
 - i. The center is located at the longest stem(s) in the stem-and-leaf plot or the tallest rectangular bar(s) in the histogram.

- ii. The range begins at the lowest valued stem in the stem-and-leaf plot with nonzero length or the leftmost rectangular bar in the histogram with nonzero height; ends at the highest valued stem in the stem-and-leaf plot with nonzero length or the rightmost rectangular bar in the histogram with nonzero height.
- (b) Box plot:
 - i. The center is located at the thick line in the box.
 - ii. The range begin at the bottom of the lower whisker and ends at the top of the upper whisker.
 - iii. The inter-quartile range (IQR) is the height of the box.
- 8. The *outliers* refer to the observations that fall far from the bulk of the data.
- 9. Shape of a distribution:
 - (a) Number of modes:
 - i. If the histogram of a distribution contains one peak, then the distribution is *unimodal*.
 - ii. If the histogram contains two peaks, then the distribution is *bimodal*.
 - iii. If the histogram does not contain any obvious peaks, then the distribution is *uniform*.
 - (b) Skewness:
 - i. If the histogram of a distribution is symmetric about the center, then the distribution is *symmetric*.
 - ii. If the distance from the center is farther to the maximum than to the minimum, i.e. the upper tail is longer, then the distribution is *right-skewed*.
 - iii. If the distance from the center is farther to the minimum than to the maximum, i.e. the lower tail is longer, then the distribution is *left-skewed*.
- 10. Computing and interpreting the numerical summaries of quantitative data:
 - (a) Measures of center:
 - i. Mean: $\bar{x} = \frac{\sum x_i}{n}$.
 - ii. Median: M = the middle value (when the number of observations is odd); the average of the middle two values (when the number of observations is even.)

Note: For a right-skewed distribution, the mean is usually greater than the median; for a left-skewed distribution, the mean is usually less than the median.
 - (b) Measures of spread:
 - i. Standard deviation: $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$.
 - ii. Range: maximum – minimum.
 - iii. Inter-quartile range (IQR): 3rd quartile – 1st quartile.
- 11. A numerical summary of the data is *resistant* if extreme observations have little influence on its value.
Examples: The median and IQR are resistant, but the mean, range, and standard deviation are not.
- 12. *3-standard deviations rule*: For approximately bell-shaped distributions, most observations fall within 3 standard deviations of the mean (see also the 68-95-99.7 rule in Chapter 6.)

Chapter 4

13. *Experiments and observational studies:*
 - (a) In an experiment, the researcher assigns the subjects to various experiment conditions (i.e. *treatments*), and then observe the response variable.
 - (b) In an observational study, the researcher observes both the explanatory variable(s) and the response variable; there is no assigning of treatments.
14. Causal relationships can only be established through experiments.
15. Sampling frame is the list of subjects in the population from which we draw the sample.
16. Simple random sampling is a method of sampling in which every possible sample of size n from the population is equally likely to be selected. It is the best way to obtain a sample that is representative of the population.
17. Types of biases:
 - (a) Sampling bias: Using a sample that is not randomly drawn from the population.
 - (b) Nonresponse bias: Subjects in the sample are unavailable or (intentionally or unintentionally) do not respond.
 - (c) Response bias: Subjects in the sample give inaccurate responses either because they are untruthful or the questions are misleading.
18. *Convenience sample* is the type of sample which is easy to obtain. For example, *volunteer sample* is a common form of convenience sample. Convenience sample is usually unrepresentative.

Chapter 5

19. A *random phenomenon* is an event in which the individual outcomes are uncertain but the long-run behavior is regular.
20. The *probability* of an event can be interpreted as its long-run proportion of occurring.
21. A *probability model* contains a list of all possible outcomes and the probability of each outcome.
22. The *sample space* of a random phenomenon is the set of all possible outcomes.
23. The basic probability rules:
 - (a) $0 \leq P(A) \leq 1$.
 - (b) Law of total probability: $P(S) = 1$ where S is the sample space.
 - (c) Complement rule: $P(A^c) = 1 - P(A)$ where A^c is the complement of A .
 - (d) General addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
 - (e) Partition of probability: $P(A) = P(A \cap B) + P(A \cap B^c)$.
24. *Disjoint events:*
 - (a) Events A and B are disjoint if $A \cap B$ is empty.
 - (b) When A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$, and $P(A \cap B) = 0$.
25. *Independent events:* Events A and B are independent if $P(A \cap B) = P(A)P(B)$, that is, knowing that one occurs does not affect the probability that the other occurs.

26. *Conditional probability* of event A given event B is the probability that event A occurs given that event B has occurred and is denoted as $P(A | B)$. Provided that $P(B) > 0$, we have $P(A | B) = \frac{P(A \cap B)}{P(B)}$.
27. General multiplication rule: $P(A \cap B) = P(A | B)P(B) = P(B | A)P(A)$.
28. Two events A and B are independent if any of the following equivalent conditions holds:
- (a) $P(A \cap B) = P(A)P(B)$.
 - (b) $P(A | B) = P(A)$.
 - (c) $P(B | A) = P(B)$.

Chapter 6

29. *Random variable* is the numerical outcome of a random phenomenon.
30. Types of random variable:
- (a) Discrete: the possible values can be listed.
 - (b) Continuous: the possible values are from an interval.
31. Properties of discrete random variables:
- (a) Each possible outcome has a non-negative probability.
 - (b) The probabilities add up to 1.
32. Center and spread of a discrete random variable:
- (a) Mean: $\mu = \sum xP(x)$, which can be interpreted as the long-run average outcome, i.e. the expected average value of the outcomes if a very large sample is drawn.
 - (b) Standard deviation: $\sigma = \sqrt{\sum (x - \mu)^2 P(x)}$.
33. The probability distribution of continuous random variables are described by density curves, which have the following properties:
- (a) Non-negative everywhere.
 - (b) For every numbers a and b such that $a \leq b$, the probability $P(a < X < b)$ equals the area under the density curve between $X = a$ and $X = b$.
 - (c) The total area under density curve is 1.
34. *The normal distribution*:
- (a) The normal distributions are a family of distributions controlled by the parameters μ and σ , denoted as $N(\mu, \sigma)$.
 - (b) The density curves of all normal distributions have the same overall shape, which is bell-shaped, symmetric, and centered at μ .
 - (c) The $N(\mu, \sigma)$ distribution has the mean and median equal to μ , and the standard deviation equal to σ .
 - (d) By changing μ (while fixing σ), the normal density curve shifts its location horizontally but does not stretch; by changing σ (while fixing μ), the normal density curve stretches but the center remains unchanged.
 - (e) The $N(0, 1)$ distribution is known as the standard normal distribution and is denoted as Z .

35. The 68-95-99.7 rule: For any normal distribution,
- (a) approximately 68% of the distribution falls within 1σ of μ .
 - (b) approximately 95% of the distribution falls within 2σ of μ .
 - (c) approximately 99.7% of the distribution falls within 3σ of μ .
36. Using the normal probability table:
- (a) The normal probability table allows us to find the probability $P(Z < a)$ for any number a .
 - (b) To compute the probability $P(Z > a)$, rewrite it as $1 - P(Z < a)$ and then use the normal probability table to find $P(Z < a)$.
 - (c) To compute the probability $P(a < Z < b)$, rewrite it as $P(Z < b) - P(Z < a)$ and then use the normal probability table to find $P(Z < b)$ and $P(Z < a)$.
 - (d) To compute the p -th percentile of Z , search for the number a on the normal probability table such that $P(Z < a)$ is as close to $\frac{p}{100}$ as possible.
37. In R `pnorm()` and `qnorm()` for standard normal distribution
- (a) To find the left tail area of z , (cumulative probability, $P(Z < z)$), we use `pnorm(z)`.
 - (b) To find the right tail area of z , $P(Z > z) = 1 - P(Z < z)$, we use `1-pnorm(z)` or `pnorm(z, lower.tail=F)`.
 - (c) To find the area between two z values, $P(z_a < Z < z_b) = P(Z < z_b) - P(Z < z_a)$, we use `pnorm(z_b) - pnorm(z_a)`.
 - (d) To find a p -th percentile of z -score, z^* where $P(Z < z^*) = p$, we use `qnorm(p)`. For instance, `qnorm(0.9)` returns `z=1.28` and 90% of the standard normal distribution are less than or equal to this value.
38. *Standardizing and unstandardizing:*
- (a) Any normal distribution $X \sim N(\mu, \sigma)$ can be transformed into the standard normal distribution through the transformation $Z = \frac{X - \mu}{\sigma}$. Under this transformation, an observation x drawn from the $N(\mu, \sigma)$ distribution corresponds to the value $z = \frac{x - \mu}{\sigma}$ in the standard normal distribution. The value z is known as the z -score of x .
 - (b) Let $X \sim N(\mu, \sigma)$. The probabilities involving X can be computed by first standardizing X into the standard normal distribution. For example, the probability $P(X < a)$ can be computed by rewriting it as $P(Z < \frac{a - \mu}{\sigma})$ and then using the normal probability table to find $P(Z < \frac{a - \mu}{\sigma})$.
 - (c) Let $X \sim N(\mu, \sigma)$. To find the p -th percentile x^* of X , first find the p -th percentile z^* of Z from `qnorm(z*)` then unstandardize z^* into x^* using the formula $x^* = \mu + z^* \sigma$.
39. We can use `pnorm(x, mean= $\mu$ , sd= $\sigma$ )` and `qnorm(p, mean= $\mu$ , sd= $\sigma$ )` to probability and quantile with mean μ and standard deviation σ . For example to find $P(X < 100)$ where $X \sim N(80, 15)$, we use `pnorm(100, 80, 15)` which returns 0.908.
40. Normality assessment:
- (a) Histogram and/or boxplot:
Check for the unimodality, skewness, and outliers.
 - (b) Normal Q-Q plot:
A *Q-Q plot* is constructed by plotting the quantiles of a distribution against that of another distribution. In a *normal Q-Q plot*, the quantiles of the data are plotted against the quantiles of the standard normal distribution. The normal Q-Q plot allows us to check for the heavy-tailedness of the data. If the data points on the normal Q-Q plot approximately form a straight diagonal line, then it is reasonable that the data comes from a normal distribution; otherwise, the data is likely to come from some other distribution.

41. Conditions for Binomial distribution

- (a) The number of trials (n) is fixed.
- (b) Outcome from each trial is independent.
- (c) Each outcome can be categorized as success/failure.
- (d) Probability of success p remains the same.

42. Binomial distribution

- (a) The number X of successes in n independent trials is called a *binomial random variable*.
- (b) The probability distribution of the binomial random variable X , the number of successes in n independent trials with the probability of success p , is denoted as $X \sim \text{binom}(n, p)$.
- (c) For a particular number of successes ' x ',

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad \text{for } x = 0, 1, 2, \dots, n.$$

where $\binom{n}{x} = {}_n C_x = \frac{n!}{(n-x)!x!}$ represents the number of possible outcomes with x successes and $n - x$ failures.

- (d) In R, we use `dbinom(x, n, p)` to obtain the value of probability $P(X = x)$ and `pbinom(x, n, p)` to obtain $P(X \leq x)$.
- (e) Mean of binomial distribution is $\mu = np$
- (f) Variance of binomial distribution is $\sigma^2 = np(1 - p)$
- (g) Standard deviation of binomial distribution is $\sigma = \sqrt{np(1 - p)}$

Chapter 7

43. *Sampling distribution* is the probability distribution of the values of a sample statistic (e.g. sample mean, sample proportion) among all possible samples of the same size from the population.

44. *Standard error* refers to the standard deviation of a sample statistic, with all involved unknown parameters replaced by their estimates.

(Note: the exact definition of standard error depends on the context. In Chapter 7 we assume that we knew the population parameters, so the standard error is just the standard deviation of a sample statistic; in Chapters 8 and 9, we shall see that the standard error for the sample proportion is different for confidence intervals and hypothesis tests.)

Section 7.1 – Sampling distribution of a sample mean

45. Law of large numbers:

Let $x_1, x_2, \dots, x_n$ be independent observations drawn from a population with mean μ . As n increases towards infinity, $\bar{x} \rightarrow \mu$.

46. Mean and standard deviation of the sampling distribution of $\bar{x}$:

Let $x_1, x_2, \dots, x_n$ be a random sample drawn from a (not necessarily normal) population with mean μ and standard deviation σ , and let $\bar{x}$ be the sample mean.

- (a) The mean of $\bar{x}$ is $\mu_{\bar{x}} = \mu$ (same as the population mean.)

- (b) The standard deviation of $\bar{x}$ is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ (the variation of $\bar{x}$ becomes smaller as σ decreases, or as n increases.)
47. Distribution of $\bar{x}$ when the population is normal:
 If the population follows a *normal* distribution $N(\mu, \sigma)$, then $\bar{x}$ follows the normal distribution $N(\mu, \frac{\sigma}{\sqrt{n}})$ for any sample size n .
48. Central Limit Theorem for $\bar{x}$:
 If the population follows *any* distribution with mean μ and standard deviation σ , then $\bar{x}$ *approximately* follows the normal distribution $N(\mu, \frac{\sigma}{\sqrt{n}})$ when n is large (rule of thumb: $n \geq 30$.)

Section 7.2 – Sampling distribution of a sample proportion

49. Law of large numbers:
 Let $\hat{p}$ be the sample proportion of success in a random sample of size n drawn from a population with proportion of success p . As n increases towards infinity, $\hat{p} \rightarrow p$.
50. Mean and standard deviation of the sampling distribution of $\hat{p}$:
 Let $\hat{p}$ be the sample proportion of success in a random sample of size n drawn from a population with proportion of success p .
- (a) The mean of $\hat{p}$ is $\mu_{\hat{p}} = p$ (same as the population proportion of success.)
- (b) The standard deviation of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ (the variation of $\hat{p}$ becomes smaller as p moves away from 0.5, or as n increases.)
51. Central Limit Theorem for $\hat{p}$:
 The sample proportion $\hat{p}$ approximately follows the normal distribution $N(p, \sqrt{\frac{p(1-p)}{n}})$ when n is sufficiently large (rule of thumb: $np \geq 15$ and $n(1-p) \geq 15$.)

Chapter 8

52. Point estimate:

- (a) *Point estimation* is the process of using a sample statistic to estimate an unknown parameter (e.g. using $\bar{x}$ to estimate μ .)
- (b) The *bias* of an estimator is the difference between the mean of the estimator and the true value of the parameter. An estimator with a bias of 0 is called an *unbiased* estimator.
- (c) A good estimator should have both a small bias and a small standard error (see item 44.)

53. Interval estimate:

- (a) An *interval estimate* of a parameter is a range of numbers that we believe the true parameter value is likely to fall within.
- (b) The *margin of error* of a point estimate is a non-negative number that describes the point estimate θ uncertainty in estimating a population parameter.
- (c) A *confidence interval* is an interval containing the most believable values of a parameter.
- (d) The general form of a confidence interval is

$$(\text{point estimate}) \pm (\text{margin of error})$$

where

$$\text{margin of error} = (\text{critical value}) \cdot (\text{standard error of point estimate}).$$

Section 8.2.1 – Confidence interval for a population proportion p

54. Confidence interval for a population proportion:

- (a) Assumptions: Random sample with at least 15 successes and 15 failures.
- (b) A $100(1 - \alpha)\%$ confidence interval for the population proportion p is given by $\hat{p} \pm z_{\frac{\alpha}{2}}^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where the critical z -value $z_{\frac{\alpha}{2}}^*$ is the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution (see item 55 below for common values.)
- (c) The quantity $100(1 - \alpha)\%$ is the *confidence level* of the confidence interval, and α is the *error probability*.
- (d) Interpretation of the confidence interval: “We are $100(1 - \alpha)\%$ confident that the true proportion of ... is between $\hat{p} - z_{\frac{\alpha}{2}}^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and $\hat{p} + z_{\frac{\alpha}{2}}^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.”

55. Commonly used critical z -values :

Confidence level	α	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}^*$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.575

56. Properties of confidence intervals for population proportions:

- (a) The margin of error increases as the confidence level increases (due to the increase of $z_{\frac{\alpha}{2}}^*$, see item 55 above.)
- (b) For a fixed $\hat{p}$, the margin of error decreases as the sample size n increases.

57. Sample size calculations:

Given the desired margin of error m , the minimum required sample size is

$$n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\frac{\alpha}{2}}^*}{m} \right)^2,$$

where $\hat{p}$ is an educated guess of p from previous studies, or use $\hat{p} = 0.5$ if no information about p is available. The result should be rounded up.

Section 8.2.2 – Confidence interval for a population mean μ

58. Confidence interval for a population mean:

- (a) Assumptions: Random sample; the population is approximately normal (violation of the normality assumption can be tolerated when n is large and there is no outlier)
- (b) A $100(1 - \alpha)\%$ confidence interval for the population mean μ is given by $\bar{x} \pm t_{n-1, \frac{\alpha}{2}}^* \frac{s}{\sqrt{n}}$, where the critical t -value $t_{n-1, \frac{\alpha}{2}}^*$ is the upper $\frac{\alpha}{2}$ quantile of the t -distribution with $(n - 1)$ degrees of freedom (see item 59 below.)
- (c) A t -table is required for finding the critical t -value. Round down the degrees of freedom if it is unavailable on the table, except that $z_{\frac{\alpha}{2}}^*$ might be used when $n > 1,000$.
- (d) Interpretation of the confidence interval: “We are $100(1 - \alpha)\%$ confident that the true mean of ... is between $\bar{x} - t_{n-1, \frac{\alpha}{2}}^* \frac{s}{\sqrt{n}}$ and $\bar{x} + t_{n-1, \frac{\alpha}{2}}^* \frac{s}{\sqrt{n}}$.”

59. Properties of the t -distribution:

The t -distribution

- (a) is a family of distributions; its degrees of freedom determines the exact shape of the distribution.
- (b) is bell-shaped and symmetric around 0.
- (c) has heavier tails than $N(0, 1)$.
- (d) has a shape close to the shape of $N(0, 1)$ when the degrees of freedom is large.

60. Properties of confidence intervals for population proportions:

- (a) The margin of error increases as the sample standard deviation s or the confidence level increases.
- (b) The margin of error decreases as the sample size n increases.

61. Sample size calculations:

Given the desired margin of error m , the *approximate* minimum required sample size is

$$n = \left(\frac{z_{\frac{\alpha}{2}}^* s^*}{m} \right)^2,$$

where s^* is an educated guess of the population standard deviation σ . The result should be rounded up.

Note: A more accurate method to find the minimum sample size is to solve for the smallest n such that the inequality $n \geq \left(\frac{t_{n-1, \frac{\alpha}{2}}^* s^*}{m} \right)^2$ holds, but this requires trial and error.

Chapter 9

62. The goal of a *hypothesis test* is to formally test whether a certain number, θ_0 , is a feasible value of an unknown population parameter θ , based on the sample data.

63. General form of a hypothesis test:

- Assumptions: The way that a hypothesis test works is based on certain assumptions about the population distribution and how the sample is drawn from the population.
- Hypotheses: The null hypothesis H_0 has the form $\theta = \theta_0$; the alternative hypothesis H_a can either be $\theta < \theta_0$, $\theta > \theta_0$, or $\theta \neq \theta_0$, depending on the context.
- Test statistic: $\frac{\hat{\theta} - \theta_0}{\text{standard error of } \hat{\theta} \text{ under } H_0}$, where $\hat{\theta}$ is the point estimate of θ based on the sample.
- p -value: The probability of obtaining a test statistic at least as extreme as the observed one under H_0 .
- Conclusion: Reject H_0 if the p -value is less than the significance level α ; failed to reject H_0 otherwise.

64. Hypothesis test for population proportion:

Assumptions	- Random sample - At least 15 expected successes and 15 expected failures in the sample ($np_0 \geq 15$ and $n(1 - p_0) \geq 15$) under H_0		
Hypotheses	$H_0 : p = p_0$ $H_a : p < p_0$	$H_0 : p = p_0$ $H_a : p > p_0$	$H_0 : p = p_0$ $H_a : p \neq p_0$
Test statistic	$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ When H_0 is true, the test statistic approximately follows the z-distribution.		
p -value	$\Pr(Z < z^*)$	$\Pr(Z > z^*)$	$2\Pr(Z > z^*)$
Conclusion and interpretation	- If $p\text{-value} < \alpha$, reject H_0. “We have strong evidence to conclude that the true proportion of . . . is (less than / greater than / different from) p_0.” - If $p\text{-value} \geq \alpha$, failed to reject H_0. “We do not have enough evidence to conclude that the true proportion of . . . is (less than / greater than / different from) p_0.”		

65. Hypothesis test for population mean:

Assumptions	• Random sample • The population is approximately normal or the sample size is large ($n \geq 30$)		
Hypotheses	$H_0 : \mu = \mu_0$ $H_a : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_a : \mu > \mu_0$	$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$
Test statistic	$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ When H_0 is true, the test statistic follows the t -distribution with $n - 1$ degrees of freedom.		
p -value	$\Pr(T_{n-1} < t^*)$	$\Pr(T_{n-1} > t^*)$	$2\Pr(T_{n-1} > t^*)$
Conclusion and interpretation	• If $p\text{-value} < \alpha$, reject H_0. "We have strong evidence to conclude that the true mean of ... is (less than / greater than / different from) μ_0." • If $p\text{-value} \geq \alpha$, failed to reject H_0. "We do not have enough evidence to conclude that the true mean of ... is (less than / greater than / different from) μ_0."		

66. Equivalence between confidence intervals and two-sided hypothesis tests for population means:

- The two-sided hypothesis test for population mean with the null hypothesis $H_0 : \mu = \mu_0$ and alternative hypothesis $H_a : \mu \neq \mu_0$ results in the rejection of H_0 at the significance level α precisely when μ_0 is NOT in the $100(1 - \alpha)\%$ confidence interval for μ .
- The equivalence does NOT hold for one-sided hypothesis tests, since the two-sided tests and the confidence intervals truncate both tails of the sampling distribution, but the one-sided tests only truncate one tail of the sampling distribution.
- The equivalence does NOT hold for the hypothesis tests for population proportions, since the standard error of $\hat{p}$ for confidence intervals is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, but that for hypothesis tests is $\sqrt{\frac{p_0(1-p_0)}{n}}$.

67. Types I and II errors:

		Truth	
		H_0 is true	H_a is true
Action	Failed to reject H_0	Correct decision	Type II error
	Reject H_0	Type I error	Correct decision

- The probability of making a type I error under H_0 is α (the significance level.)
- The probability of making a type II error under H_a is more difficult to calculate. However, in general, the probability of a type II error decreases as the probability of a type I error increases.

68. Caveats about hypothesis tests:

- Statistical significance does NOT imply practical significance. A statistical significance suggests that there might be a difference between the true parameter value and the hypothesized value, but does NOT suggest that such a difference would cause a practical impact.
- Failing to reject H_0 is NOT the same as accepting H_0 . Failing to reject H_0 implies that p_0 (or μ_0) is one of the feasible values for p (or μ), but does NOT imply that p_0 (or μ_0) is THE ONLY feasible value for p (or μ .)
- The p -value is the probability of obtaining a more extreme test statistic than the one observed under H_0 , NOT the probability that H_0 is true given the observed test statistic.

Chapter 10

Section 10.1 – Comparing two independent population proportions

69. The point estimate of $p_1 - p_2$, the difference between two independent population proportions, is $\hat{p}_1 - \hat{p}_2$, the difference between the two sample proportions.

70. Sampling distribution of $\hat{p}_1 - \hat{p}_2$:

- The mean of $\hat{p}_1 - \hat{p}_2$ is $p_1 - p_2$.
- The standard deviation of $\hat{p}_1 - \hat{p}_2$ is $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.
- When both sample sizes n_1 and n_2 are sufficiently large (rule of thumb: at least 10 expected successes and 10 expected failures in each sample), the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal: $\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$.

71. Confidence interval for $p_1 - p_2$:

- Assumptions: Two independent random samples, each with at least 10 successes and 10 failures.
- A $100(1 - \alpha)\%$ confidence interval for $p_1 - p_2$ is given by $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, where the critical z -value $z_{\alpha/2}^*$ is the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution.
- Interpretation of the confidence interval:
 - If 0 is in the confidence interval, it is plausible that $p_1 = p_2$.
“We are $100(1 - \alpha)\%$ confident that the difference between the true proportions of ... and ... is between $(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ and $(\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.”
 - If all values in the confidence interval are greater than 0, it is likely that $p_1 > p_2$; if all values in the confidence interval are less than 0, it is likely that $p_1 < p_2$.
“We are $100(1 - \alpha)\%$ confident that the true proportion of ... is between $\left|(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right|$ and $\left|(\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right|$ greater than the true proportion of ...”

72. Hypothesis test for $p_1 - p_2$:

Assumptions	- Two independent random samples - At least 5 successes and 5 failures in each sample		
Hypotheses	$H_0 : p_1 = p_2$ $H_a : p_1 < p_2$	$H_0 : p_1 = p_2$ $H_a : p_1 > p_2$	$H_0 : p_1 = p_2$ $H_a : p_1 \neq p_2$
Test statistic	$z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, where $\hat{p} = \frac{\text{combined number of successes}}{\text{combined sample size}} = \frac{X_1 + X_2}{n_1 + n_2}$ When H_0 is true, the test statistic approximately follows the z -distribution.		
p -value	$\Pr(Z < z^*)$	$\Pr(Z > z^*)$	$2\Pr(Z > z^*)$
Conclusion and interpretation	- If p-value $< \alpha$, reject H_0. “We have strong evidence to conclude that the true proportion of ... is (less than / greater than / different from) that of ...” - If p-value $\geq \alpha$, failed to reject H_0. “We do not have enough evidence to conclude that the true proportion of ... is (less than / greater than / different from) that of ...”		

Section 10.2 – Comparing population means for matched pairs

73. Two samples are called *matched pairs* if each observation in one sample is matched with an observation in the other sample.
- (a) Examples of a matched pair:
 - i. Measurements of a subject before and after a treatment.
 - ii. Measurements of two subjects after receiving different treatments, where the two subjects were matched up by their characteristics.
 - (b) Reason of using matched pairs: Controlling for extra sources of variability in the response variable.
74. Method of analyzing matched pairs:
 Compute the difference for each matched pair. Then proceed the paired differences using one-sample t -procedures. Note that the sample size n_D in the t -procedures is the number of *pairs* in the sample. (See item 58 for confidence intervals and item 65 for hypothesis tests.)

Section 10.3 – Comparing two independent population means

75. The point estimate of $\mu_1 - \mu_2$, the difference between two independent population means, is $\bar{x}_1 - \bar{x}_2$, the difference between the two sample means.
76. Sampling distribution of $\bar{x}_1 - \bar{x}_2$:
- (a) The mean of $\bar{x}_1 - \bar{x}_2$ is $\mu_1 - \mu_2$.
 - (b) The standard deviation of $\bar{x}_1 - \bar{x}_2$ is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
 - (c) When both populations are normal or when sample sizes n_1 and n_2 are large, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal: $\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$.
 - (d) However, since σ_1 and σ_2 in the standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ is unknown, we estimate the standard deviation by the standard error $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. The distribution of $\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ approximately follows the t -distribution with the Welch's degrees of freedom. For simplicity, we replace the Welch's degrees of freedom by its lowest possible value $\min\{n_1 - 1, n_2 - 1\}$. That is:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim T_{\min\{n_1 - 1, n_2 - 1\}}.$$

77. Confidence interval for $\mu_1 - \mu_2$:
- (a) Assumptions: Two independent random samples; both populations are approximately normal (violation of the normality assumption can be tolerated when n_1 and n_2 are large and there is no outlier)
 - (b) A $100(1-\alpha)\%$ confidence interval for the population mean difference $\mu_1 - \mu_2$ is given by $(\bar{x}_1 - \bar{x}_2) \pm t_{\min\{n_1 - 1, n_2 - 1\}, \alpha/2}^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, where the critical t -value $t_{\min\{n_1 - 1, n_2 - 1\}, \alpha/2}^*$ is the upper $\frac{\alpha}{2}$ quantile of the t -distribution with $\min\{n_1 - 1, n_2 - 1\}$ degrees of freedom.
 - (c) Interpretation of the confidence interval:
 - i. If 0 is in the confidence interval, it is plausible that $\mu_1 = \mu_2$.
 “We are $100(1-\alpha)\%$ confident that the difference between the true means of ... and ... is between $(\bar{x}_1 - \bar{x}_2) - t_{\min\{n_1 - 1, n_2 - 1\}, \alpha/2}^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and $(\bar{x}_1 - \bar{x}_2) + t_{\min\{n_1 - 1, n_2 - 1\}, \alpha/2}^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.”

- ii. If all values in the confidence interval are greater than 0, it is likely that $\mu_1 > \mu_2$; if all values in the confidence interval are less than 0, it is likely that $\mu_1 < \mu_2$.

“We are $100(1 - \alpha)\%$ confident that the true mean of ... is between

$\left| (\bar{x}_1 - \bar{x}_2) - t_{\min\{n_1-1, n_2-1\}, \alpha/2}^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right|$ and $\left| (\bar{x}_1 - \bar{x}_2) + t_{\min\{n_1-1, n_2-1\}, \alpha/2}^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right|$ greater than the true mean of ...”

78. Hypothesis test for $\mu_1 - \mu_2$:

Assumptions	• Two independent random samples • Both population are approximately normal or the sample sizes are large		
Hypotheses	$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 < \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 > \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 \neq \mu_2$
Test statistic	$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ When H_0 is true, the test statistic approximately follows the t -distribution with $\min\{n_1 - 1, n_2 - 1\}$ degrees of freedom.		
p -value	$\Pr(T_{\min\{n_1-1, n_2-1\}} < t^*)$	$\Pr(T_{\min\{n_1-1, n_2-1\}} > t^*)$	$2\Pr(T_{\min\{n_1-1, n_2-1\}} > t^*)$
Conclusion and interpretation	• If p-value $< \alpha$, reject H_0. “We have strong evidence to conclude that the true mean of ... is (less than / greater than / different from) that of ...” • If p-value $\geq \alpha$, failed to reject H_0. “We do not have enough evidence to conclude that the true mean of ... is (less than / greater than / different from) that of ...”		

Chapter 14

79. Main idea of the Analysis of Variance (ANOVA):

The goal of the *Analysis of Variance (ANOVA)* is to test the equality of 3 or more population means.

If all groups have the same population mean, then the *variation between groups* should be of similar size to the *variation within groups* (see item 81 below); if some groups have higher population means than some other groups, then the variation between groups is expected to be greater than the variation within groups.

Therefore, we estimate the between-group and within-group variations, and compute their ratio. If the ratio is around 1, then it is feasible that all groups have the same population mean; if the ratio is much greater than 1, then it is likely that some groups have higher means than some other groups.

80. Drawbacks of using pairwise t -tests to test the equality of 3 or more population means:

- There are as many as $\binom{g}{2} = \frac{g(g-1)}{2}$ pairs of means to test, where g is the number of groups. However, the probability of making at least one type I error increases as the number of tests increases.
- Each t -test only allows us to compare 2 means at a time, but we would like to test whether all means are equal using a single test.

81. Estimating the between-group variation, the within-group variation, and their ratio:

- Estimating variability between groups:

- i. Sum of Squared deviations for Groups ($SSG = \sum_{i=1}^g n_i (\bar{y}_i - \bar{y})^2$)
 - ii. Mean Squared deviation for Groups ($MSG = \frac{SSG}{g-1}$, a measure of the variation of the group means)
- (b) Estimating variability within groups:
- i. Sum of Squared deviations for Error ($SSE = \sum_{i=1}^g (n_i - 1) s_i^2 = \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$).
 - ii. Mean Squared deviation for Error ($MSE = \frac{SSE}{N-g}$, a measure of the variation of the observations within groups; also serves as an estimate of σ^2 , the common population variance of each group)
- (c) Estimating overall variability:
- Total sum of squares ($SST = \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = SSG + SSE$, which is not directly used in ANOVA, but allows us to compute SSG from SSE or vice versa)
- (d) The F -statistic:
- $F = \frac{MSG}{MSE}$, the ratio of the between-group variation to the within-group variation

82. The ANOVA table:

Source of variation	df	SS	MS	F	p -value
Group	$g - 1$	$SSG = \sum_{i=1}^g n_i (\bar{y}_i - \bar{y})^2$	$MSG = \frac{SSG}{g-1}$	$F^* = \frac{MSG}{MSE}$	$\Pr(F_{g-1, N-g} > F^*)$
Error	$N - g$	$SSE = \sum_{i=1}^g (n_i - 1) s_i^2$	$MSE = \frac{SSE}{N-g}$		
Total	$N - 1$	$SST = SSG + SSE$			

83. The ANOVA F -test:

Assumptions	• g independent random samples from g populations • Each population is approximately normal • The standard deviation of each population is approximately equal
Hypotheses	$H_0 : \mu_1 = \mu_2 = \dots = \mu_g$ H_a : At least 2 of the population means are unequal
Test statistic	$F^* = \frac{MSG}{MSE}$ When H_0 is true, the test statistic follows the F -distribution with $g - 1$ and $N - g$ degrees of freedom (see item 84.)
p -value	$\Pr(F_{g-1, N-g} > F^*)$
Conclusion and interpretation	• If p-value $< \alpha$, reject H_0. "We have strong evidence to conclude that the true mean of ... is different for at least two groups." • If p-value $\geq \alpha$, failed to reject H_0. "We do not have enough evidence to conclude that the true mean of ... is different for at least two groups."

84. The F -distribution:

- (a) is a family of distributions; its degrees of freedom is specified by two numbers df_1 and df_2 .
- (b) is continuous, non-negative, and right-skewed.
- (c) has a mean of approximately 1, and a mode of also approximately 1.

85. Tukey's Honest Significant Difference (HSD) test:

If the ANOVA results in the rejection of the null hypothesis, we can only conclude that there are at least two group means are significantly different. To determine which pairs of means are significantly different while controlling the overall type I error probability, we use the Tukey's HSD test to obtain the $100(1 - \alpha)\%$ family-wise confidence interval for the difference between each of the $\binom{g}{2}$ pairs of means.

The Tukey's HSD test is an example of *multiple comparison* (see item 86 below.)

86. Multiple comparison:

A multiple comparison at the significance level α performs several statistical analyses simultaneously and produces a confidence interval for each analysis, such that there is a $100(1 - \alpha)\%$ confidence that ALL the intervals contain their respective true parameter values.

Chapter 3

87. Association between two variables:

Two variables are *associated* if some values of the first variable occurs more frequently together with some specific values of the second variable (see also item 89 for the independence between two variables.)

88. Response and explanatory variables:

- (a) The *response variable* in a study is the outcome variable to be measured and compared.
- (b) The *explanatory variable* is the variable that is used to predict or explain the response variable.

Chapter 11 Chi-squared test of independence

89. Independence between two variables:

Two variables are *independent* if the distribution of one variable does not depend on the observed value of the other variable (see also item 87 for the association between two variables.)

90. Main idea of the chi-squared test:

The goal of the *chi-squared test* is to test whether there is an association between two categorical variables.

In the chi-squared test, we compute the expected distribution of the outcome combinations of the two variables given that the two variables are independent. Then we compare the expected distribution against the observed distribution. If their discrepancy is large, then it is likely that there is an association between the two variables.

91. Contingency table:

A *contingency table* displays the joint distribution of two categorical variables. Each row of the table represent an outcome of the first variable, and each column represent an outcome of the second variable. Each entry in the table represents the number of observations with the corresponding combination of outcomes of the two variables.

92. Expected cell count:

For the cell on the i -th row and j -th column of the contingency table, its *expected cell count* is

$$E_{ij} = \frac{(\text{total of row } i)(\text{total of column } j)}{\text{total of all cells}}.$$

93. The chi-squared test:

Assumptions	• Random sample • Sufficiently large sample size so that all expected cell counts are at least 5
Hypotheses	H_0 : The two variables are independent H_a : The two variables are associated
Test statistic	$\chi^{2*} = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, where O_{ij} and E_{ij} are the observed and expected cell counts for the (i, j) -th cell of the contingency table, respectively. When H_0 is true, the test statistic approximately follows the chi-squared distribution with $(r - 1)(c - 1)$ degrees of freedom (see item 94.)
p -value	$\Pr(\chi_{(r-1)(c-1)}^2 > \chi^{2*})$
Conclusion and interpretation	• If p-value $< \alpha$, reject H_0. “We have strong evidence to conclude that there is a significant association between ... and ...” • If p-value $\geq \alpha$, failed to reject H_0. “We do not have enough evidence to conclude that there is a significant association between ... and ...”

94. The chi-squared (χ^2) distribution:

- is a family of distributions; its degrees of freedom k determines the exact shape of the distribution.
- is continuous, non-negative, and right-skewed.
- has a mean of k .
- has a median of approximately $k - \frac{2}{3}$.
- has a standard deviation of $\sqrt{2k}$, and is therefore more spread out when the degrees of freedom is large.

95. Risk and relative risk:

- Risk:
 - The *risk* of an outcome is its probability of occurrence.
 - We estimate the risk of an outcome by its proportion of occurrence in the sample $\frac{X}{n}$.
- Relative risk:
 - The ratio of the risks of an outcome for two groups is called the *relative risk*.
 - The relative risk is a measure of the strength of the association between two categorical variables.
 - We estimate the relative risk among two groups by the ratio of the estimated risks of the two groups $\frac{X_1/n_1}{X_2/n_2}$.
 Interpretation: “We estimate that ... is $\frac{X_1/n_1}{X_2/n_2}$ times more likely to occur in group 1 than in group 2.”

Chapter 12

96. Main idea of the regression analysis:

The goal of the *regression analysis* is to study the association between 2 quantitative variables. Specifically,

- (a) Summarize the sample data graphically and numerically.
- (b) Based on the sample data, make inference about the true relationship between the two variables in the population.

Section 12.0 – Exploring the data

97. Scatterplot:

- (a) The x -axis represents the explanatory variable.
- (b) The y -axis represents the response variable.
- (c) Each point in the graph represents an observation in the data, where the (x, y) coordinate of the point corresponds to the values of the explanatory and response variables of the observation.

98. Interpretations of the scatterplot:

- (a) Overall pattern: The overall trend of the data points.
e.g. linear, curved
- (b) Strength: How close the data points are to a perfect pattern.
e.g. weak, moderate, strong
- (c) Direction:
 - i. Positive association: Two quantitative variables are positively associated if one of them tend to increase as the other increases.
 - ii. Negative association: Two quantitative variables are negatively associated if one of them tend to decrease as the other increases.
- (d) Outliers: Data points that are far away from the cluster formed by the rest of the data points.

99. Correlation: The correlation between the paired samples $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ is

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

100. Properties of the correlation r :

- (a) does not change if the roles of x and y are reversed.
- (b) does not change under shifting and positive scaling of x and/or y .
- (c) only reflects the strength and direction of a linear relationship.
- (d) is not robust to outliers.

101. Interpretations of the correlation r :

- (a) Direction:
 - i. $r > 0$: positive association
 - ii. $r < 0$: negative association
 - iii. $r = 0$: uncorrelated (no linear relationship, but not necessarily independent)
- (b) Strength:
 - i. $|r| \leq 0.5$: weak linear relationship

- ii. $0.5 < |r| \leq 0.8$: moderate linear relationship
- iii. $0.8 < |r| \leq 1$: strong linear relationship
- iv. $r = \pm 1$: perfect linear relationship, all data points are on a straight line
(see also item 109, the properties and interpretations of r^2)

102. Regression line:

A regression line describes the linear relationship between the response variable y and the explanatory variable x . It has the form

$$\hat{y} = a + bx$$

and can be used to predict the value of y given the value of x .

103. Properties of b , the slope of the regression line:

- (a) For each unit of increase in x , we expect y to change by b units.
- (b) The slope b and the correlation r always have the same sign.

104. Residual:

The residual for an observation (x, y) is the signed difference between the response variable (y) and the value predicted by the regression line

$$residual = y - \hat{y} = y - (a + bx).$$

Note: The order of y and $\hat{y}$ cannot be switched!

105. Interpretations of the residual:

- (a) $residual < 0 \Rightarrow y < \hat{y}$, the predicted value is too high
- (b) $residual > 0 \Rightarrow y > \hat{y}$, the predicted value is too low
- (c) $residual = 0 \Rightarrow$ the predicted value is spot on

106. Residual sum of squares (RSS):

The residual sum of squares is a measure for the goodness of fit of a regression line. It is defined as the sum of the squared residuals

$$RSS = \sum (residual)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

107. Least square regression line (LSL):

The least square regression line for the paired samples $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ is the line $\hat{y} = a + bx$ where a and b are chosen such that the residual sum of squares is minimized. Specifically, the choices for a and b are

$$\begin{aligned} b &= r \left(\frac{s_y}{s_x} \right), \\ a &= \bar{y} - b\bar{x}. \end{aligned}$$

108. Properties of the least square regression line:

- (a) Changing the roles of x and y will yield a different regression line.
- (b) The sum of residuals is 0: $\sum residual = 0$.
- (c) The RSS of the least square regression line is smaller than that of any other line.

- (d) The least square regression line always passes through the point $(\bar{x}, \bar{y})$, that is, we always have $\bar{y} = a + b\bar{x}$.

109. Properties and interpretations of r^2 , the square of the correlation:

- (a) r^2 is the proportion of variation in y that is explained by its linear relationship with x .
- (b) $0 \leq r^2 \leq 1$.
- (c) The higher the value of r^2 , the better x can predict y through a linear relationship.
- (d) $r^2 = 1$ is achieved if and only if all variability in y is explained by its linear relationship with x , which happens when all the data points are on a straight line.

Section 12.1 – Regression analysis

110. Assumptions of the population regression model:

- (a) Linearity: A linear relationship between x and the expected value of y given x , that is, $E[Y | X = x] = \mu_x = \alpha + \beta x$.
- (b) Normality: The distribution of y given x is $Y | X = x \sim N(\mu_x, \sigma) = N(\alpha + \beta x, \sigma)$.
- (c) Constant variance: The standard deviation of y given x is a constant σ for all values of x .

Section 12.2 – Inference about the population regression model

111. Point estimates of α and β in the population regression model:

- (a) a is an unbiased estimate of α
- (b) b is an unbiased estimate of β

(see item 110 for the assumptions of the population regression model, and see item 107 for the computations of a and b)

112. Sampling distribution of b :

The quantity $\frac{b - \beta}{\text{se}(b)}$ follows the t -distribution with $n - 2$ degrees of freedom, where n is the sample size.

113. Confidence interval for β :

- (a) Assumptions: Random sample, linearity, normality, and constant variance (see item 110 for details.)
- (b) A $100(1 - \alpha)\%$ confidence interval for β is given by $b \pm t_{n-2, \alpha/2}^* \text{se}(b)$, where $t_{n-2, \alpha/2}^*$ is the upper $\frac{\alpha}{2}$ quantile of the t -distribution with $n - 2$ degrees of freedom.
- (c) Interpretation of the confidence interval:
 - i. If 0 is in the confidence interval, it is plausible that x and y are independent.
 - ii. If all values in the confidence interval are greater than 0, it is likely that x and y have a positive linear association.
 - iii. If all values in the confidence interval are less than 0, it is likely that x and y have a negative linear association.
- (d) Interpretation of the confidence interval: “We are $100(1 - \alpha)\%$ confident that for each unit of increase in x , y increases between $b - t_{n-2, \alpha/2}^* \text{se}(b)$ and $b + t_{n-2, \alpha/2}^* \text{se}(b)$ units in average.”

114. Hypothesis test for β :

Assumptions	• Random sample • Linearity • Normality • Constant variance (see item 110 for details)
Hypotheses	$H_0 : \beta = 0$ $H_a : \beta \neq 0$
Test statistic	$t^* = \frac{b}{\text{se}(b)}$ When H_0 is true, the test statistic follows the t -distribution with $n - 2$ degrees of freedom.
p -value	$2\Pr(T_{n-2} > t^*)$
Conclusion and interpretation	• If $p\text{-value} < \alpha$, reject H_0. “We have strong evidence to conclude that there is a significant linear association between ... and ...” • If $p\text{-value} \geq \alpha$, failed to reject H_0. “We do not have enough evidence to conclude that there is a significant linear association between ... and ...”

115. R and R^2 :

- R (population correlation): A measure of the strength and direction of the linear association between 2 quantitative variables in the population. We estimate R by r .
- R^2 : Proportion of variation in y explained by its linear relationship with x in the population. We estimate R^2 using r^2 .

Section 12.3 – Correlation and regression: A cautionary tale

116. Caveats about correlation and regression:

- Correlation and regression are only appropriate for describing linear relationships.
- Correlation and regression are not resistant to outliers.
- Regression lines should not be extrapolated far outside the range of the observed data (see item 117 for the definition of extrapolation.)

117. Extrapolation:

Extrapolation is the use of a regression line to predict y for the values of x that are far outside the observed range of the data. It is unreliable because we have no assurance that the linear trend continues beyond the observed range of the data.

118. Lurking variable:

A *lurking variable* refers to a variable that is not included in the study, but it strongly affects the relationship among the variables of primary interest. It may either falsely suggest a strong relationship or hide an important existing relationship.

119. Steps of establishing a causation:

- Set up a carefully designed experiment that controls for the potential lurking variables.

- (b) The association should be strong and consistent across different studies.
- (c) The alleged cause should be plausible and preceded the effect in time.
- (d) Higher doses should evoke stronger responses.

Chapter 13

120. Main idea of the multiple regression:

The goal of the *multiple regression* is to use multiple explanatory variables simultaneously to predict a single response variable. The reasons of using the multiple regression are:

- (a) Multiple regression uses more information to improve the prediction.
- (b) Multiple regression allows us to analyze the association between 2 variables while controlling for other factors.

Section 13.1 – The multiple regression model

121. Assumptions of the population regression model:

- (a) Linearity: There is a linear relationship between y and each explanatory variable x_i .
- (b) Normality: The distribution of y given the explanatory variables $(x_1, x_2, \dots, x_k)$ is $N(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k, \sigma)$.
- (c) Constant variance: The standard deviation of y given each set of explanatory variables $(x_1, x_2, \dots, x_k)$ is a constant σ .

Section 13.2 – Estimation of the multiple regression model

122. Least square regression:

We estimate the population regression equation $\mu_{Y|X_1, X_2, \dots, X_k} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$ (see item 121 for the assumptions) by the sample regression equation $\hat{y} = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$, where $a, b_1, b_2, \dots, b_k$ are chosen to minimize the residual sum of squares $\sum_{i=1}^n (y_i - \hat{y}_i)^2$.

123. Use the diagnostic plots to check the model assumptions:

- (a) To check for the linearity assumption, examine the scatterplots of y vs each x_i in the scatterplot matrix.
- (b) To check for the normality assumption, examine the normal Q-Q plot for the residuals.
- (c) To check for the constant variance assumption, examine the residual vs fitted values plot (residual plot.)

124. Prediction using the least square regression equation:

For the given set of explanatory variables $(x_1, x_2, \dots, x_k)$, the value of y is predicted as $\hat{y} = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$.

125. Estimates of β_i 's and their interpretations:

For each $i = 1, 2, \dots, k$, we estimate β_i using b_i . Interpretation: We expect y to change by b_i units for each unit of change in x_i , given that all other explanatory variables are fixed.

126. Residual and its interpretation:

As in the case of the simple linear regression, the residual for an observation (x, y) is the signed difference between the response variable (y) and the value predicted by the regression line

$$residual = y - \hat{y} = y - (a + b_1x_1 + b_2x_2 + \cdots + b_kx_k).$$

Note: The order of y and $\hat{y}$ cannot be switched.

Interpretations of the residual:

- (a) $residual < 0 \Rightarrow y < \hat{y}$, the predicted value is too high
- (b) $residual > 0 \Rightarrow y > \hat{y}$, the predicted value is too low
- (c) $residual = 0 \Rightarrow$ the predicted value is spot on

Section 13.3 – Inference for the multiple regression model

127. Coefficient of determination:

- (a) The *population coefficient of determination*, R^2 , is the proportion of the variation in y accounted for by its linear relationship with $x_1, x_2, \dots, x_k$ in the population.
- (b) We estimate R^2 by the *sample coefficient of determination* r^2 (see item 128 below.)

128. Properties of r^2 :

- (a) $0 \leq r^2 \leq 1$
- (b) r^2 equals the square of the sample correlation between $\hat{y}$ and y .

$$r^2 = \left(\frac{1}{n-1} \sum_{i=1}^n \left(\frac{\hat{y}_i - \bar{y}}{s_{\hat{y}}} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right)^2$$

- (c) r^2 increases, or at least stays the same, when any new explanatory variable is added to the model, even if the new explanatory variable is not helpful for predicting y through its linear relationship.

129. The F -test for testing that at least one β_i is not 0:

Assumptions	• Random sample • Linearity • Normality • Constant variance (see item 121 for details)
Hypotheses	$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$ H_a : At least one β_i is not 0
Test statistic	$F^* = \frac{\text{Mean Square for Regression}}{\text{Mean Square for Error}}$ When H_0 is true, the test statistic follows the F -distribution with k and $n - k - 1$ degrees of freedom.
p -value	$\Pr(F_{k,n-k-1} > F^*)$
Conclusion and interpretation	• If $p\text{-value} < \alpha$, reject H_0. “We have strong evidence that at least one explanatory variable is significantly associated with the response variable.” Run the individual t-tests (see item 130 below) for each β_i to determine which explanatory variables are significantly associated with the response. • If $p\text{-value} \geq \alpha$, failed to reject H_0. “We do not have enough evidence that at least one explanatory variable is significantly associated with the response variable.”

130. The t -test for testing the significances of the individual β_i 's:

Assumptions	• Random sample • Linearity • Normality • Constant variance (see item 121 for details)
Hypotheses	$H_0 : \beta_i = 0$ $H_a : \beta_i \neq 0$
Test statistic	$t^* = \frac{b_i}{\text{se}(b_i)}$ When H_0 is true, the test statistic follows the t -distribution with $n - k - 1$ degrees of freedom.
p -value	$2\Pr(T_{n-k-1} > t^*)$
Conclusion and interpretation	• If $p\text{-value} < \alpha$, reject H_0. “We have strong evidence to conclude that there is a significant linear association between the i-th explanatory variable and the response, after accounting for the linear effects of all other explanatory variables.” • If $p\text{-value} \geq \alpha$, failed to reject H_0. “We do not have enough evidence to conclude that there is a significant linear association between i-th explanatory variable and the response, after accounting for the linear effects of all other explanatory variables.”

131. Confidence interval for β_i :

- (a) Assumptions: Random sample, linearity, normality, and constant variance (see item 121 for details.)
- (b) A $100(1 - \alpha)\%$ confidence interval for β_i is given by $b_i \pm t_{n-k-1, \alpha/2}^* \text{se}(b_i)$, where $t_{n-k-1, \alpha/2}^*$ is the upper $\frac{\alpha}{2}$ quantile of the t -distribution with $n - k - 1$ degrees of freedom.
- (c) Interpretation of the confidence interval:
 - i. If 0 is in the confidence interval, it is plausible that x_i and y are independent after accounting for the linear effects of all other x_j 's.
 - ii. If all values in the confidence interval are greater than 0, it is likely that x_i and y have a positive linear association after accounting for the linear effects of all other x_j 's.
 - iii. If all values in the confidence interval are less than 0, it is likely that x_i and y have a negative linear association after accounting for the linear effects of all other x_j 's.
- (d) Interpretation of the confidence interval: "We are $100(1 - \alpha)\%$ confident that for each unit of increase in the i -th explanatory variable, the response increases between $b_i - t_{n-k-1, \alpha/2}^* \text{se}(b_i)$ and $b_i + t_{n-k-1, \alpha/2}^* \text{se}(b_i)$ units in average, after accounting for the linear effects of all other explanatory variables."

132. Confounding variable:

- (a) A *confounding variable* is a variable included in the study which is associated with both the explanatory and response variables. It would therefore affect the relationship between the explanatory and response variables.
- (b) It is good to include confounding variables in a multiple regression analysis. It allows the model to describe the relationship between the response variable and the explanatory variables of interest while imagining that the confounding variables are held fixed.

133. Difference between a confounding variable and a lurking variable:

A lurking variable is a variable which is not included in the study but could potentially be confounding if it were taken into account.