STAT 110 Practice Examination

Information for Questions 1 to 2

We observe data $y = (39.7, 41.3, 44.4, 39.0, 45.5)$.

1. The sample mean $\bar{y}$ is closest to

A. 41.3 B. 39.7 C. 42.0 D. 52.5 E. 209.9

2. The sample standard deviation s is closest to

A. 1.0 B. 8.2 C. 94.0 D. 2.9 E. 1.7

---

3. While on holiday we watch a match of T20 cricket. The total number of runs scored across both innings was 325. Using only the information provided, select the best option below:

A. It is unlikely the total number of runs in a T20 match ever exceeds 500.

B. Since this was a randomly chosen game, a total score of 325 can be considered typical for a T20 match.

C. We cannot determine if 325 runs is unusual unless we know something about the range of likely scores in a T20 match.

D. It is normal that more than 300 runs are scored in a T20 match.

E. Every T20 game will have a total of 325 runs scored.

---

4. Are variation and uncertainty important concepts in statistics?

A. Yes, we use statistical models to describe variation, and we quantify uncertainty wherever possible.

B. No, statistics is about finding exact population level parameters, which are known without error.

C. Yes, it helps make the data analysis look more sophisticated.

D. No, with our interconnected world, variation and uncertainty are no longer important as we can collect and use large samples in many instances.

E. No, once data has been collected, there is no longer any uncertainty.

Information for Questions 5 to 7

In a genetic study, researchers investigate copy number variation (CNV) at a specific gene locus. For a certain population, the number of copies this gene can take is one of four values: 0 (gene deletion), 1 (single copy), 3 (partial duplication), or 4 (full duplication). Let Y represent the number of copies in a randomly selected individual. The probability distribution is shown below.

i	1	2	3	4
y_i	0	1	3	4
$\Pr(Y = y_i)$	0.05	0.15	0.50	?

5. What is the probability of observing 3 or 4 copies of the gene, i.e. $\Pr(Y = 3 \text{ or } 4)$?

A. 0.30 B. 0.80 C. 0.20 D. 0.50 E. 1.00

6. Find the quantity $E[Y]$

A. 2.25 B. 2.85 C. 2.50 D. 3.00 E. 2.80

7. Which of the following is the best description of $E[Y]$?

A. A randomly selected person from this population will have $E[Y]$ copies of the gene.

B. It is the expected, or average, number of copies of the gene for a randomly selected person.

C. Half of the population will have fewer than $E[Y]$ copies of the gene.

D. An individual with exactly $E[Y]$ copies of the gene is at increased risk for genetic disorders.

E. The gene copy number is normally distributed with mean $E[Y]$.

Information for Questions 8 to 11

A marketing team wants to understand preferences for plant-based burgers. We let V be the event that a customer is a vegetarian, and B be the event that a customer chooses a plant-based burger. We have

- $\Pr(V) = 0.30$
- $\Pr(B|V) = 0.600$
- $\Pr(B|V^{\complement}) = 0.15$

8. What is the best interpretation of $\Pr(B \mid V^{\complement})$?

A. The probability that a customer chooses a plant-based burger, given that they are not vegetarian.

B. The probability that a customer is not vegetarian, given that they chose a plant-based burger.

C. The probability that a customer chooses a plant-based burger, given that they are vegetarian.

D. The probability that a customer is not vegetarian and chooses a plant-based burger.

E. The probability that a customer chooses a plant-based burger, regardless of whether they are vegetarian.

9. The probability $\Pr(V|B)$ is closest to

A. 0.632 B. 0.600 C. 0.180 D. 0.285 E. 0.105

10. The probability that a customer is either vegetarian or buys a plant-based burger is closest to

A. 0.405 B. 0.180 C. 0.300 D. 0.285 E. 1.000

11. The probability that a customer buys a burger is closest to

A. 0.180 B. 0.300 C. 0.285 D. 0.105 E. 1.000

---

12. What is the best description of a random variable?

A. A variable that summarises both the population and sample.

B. A variable that is normally distributed.

C. It is a random process with a numerical outcome.

D. A variable whose value is fixed but unknown.

E. A variable that takes certain values depending on the observed sample.

---

13. Which of the following would best be modeled using a continuous random variable?

A. The number of eggs in a bird's nest.

B. The number of website visitors who purchase an item.

C. The pH value of seawater.

D. The number of voters who support a given candidate.

E. The number of tasks completed in a fixed time.

Information for Questions 14 to 15

A food scientist is developing a nutrient score based on lab measurements of protein and saturated fat, both measured in grams per serving. Each serving has random variation in nutrient content due to processing.

Let X be the amount of protein, with $E[X] = 10$, and $Var(X) = 1.5$. Let Y be the amount of saturated fat, with $E[Y] = 4$, and $Var(Y) = 0.5$. The proposed nutrient score is $2X - 3Y$.

14. What is the expected nutrient score per serving, $E[2X - 3Y]$?

A. 4 B. 20 C. 12 D. 8 E. 0

15. What is the standard deviation of the nutrient score, $\text{sd}(2X - 3Y)$ closest to? You should assume that X and Y are independent.

A. 10.50 B. 1.22 C. 3.24 D. 2.12 E. 1.50

Information for Questions 16 to 19

Working memory span refers to the amount of information a person can temporarily hold and manipulate in their mind while performing a cognitive task. A score of working memory span has been developed that is normally distributed with mean $\mu = 40$ and standard deviation $\sigma = 8$ for healthy adults in the population.

16. A randomly selected healthy adult has a working memory score that is 1.5 standard deviations below the mean ($z = -1.5$). Their working memory score is closest to

A. -1.5 B. 38.5 C. 52.0 D. 28.0 E. 40.0

17. Which of the following options calculates the probability that a randomly selected healthy adult has a score above 48?

A. `1-pnorm(1.0)` B. `1-pnorm(48)` C. `1-pnorm(-1.0)` D. `pnorm(1.0)` E. `pnorm(48)`

18. If we were to collect a sample of $n = 64$ healthy adults and calculate their working memory score, select the option below that best describes the sampling distribution of the sample mean $\bar{y}$:

A. It is normally distributed with mean 40 and standard deviation 8

B. It is normally distributed with mean 40 and standard deviation 1.

C. It is normally distributed with mean 0 and standard deviation 1.

D. It has a t-distribution with 63 degrees of freedom.

E. There is no way to know, or even approximate, what the sampling distribution is.

19. As a summer research project we develop a new working memory score that is not normally distributed but still has mean $\mu = 40$ and standard deviation $\sigma = 8$ (we can assume it is not excessively skewed). If we were to collect a sample of $n = 64$ healthy adults and calculate their working memory score, select the option below that best describes the sampling distribution of $\bar{y}$:

A. It is approximately normally distributed with mean 40 and standard deviation 8.

B. It is approximately normally distributed with mean 40 and standard deviation 1.

C. It is approximately normally distributed with mean 0 and standard deviation 1.

D. It has a t-distribution with 63 degrees of freedom.

E. There is no way to know, or even approximate, what the sampling distribution is.

Information for Questions 20 to 22

A food scientist is analyzing the fat content of a popular brand of yogurt. She randomly selects $n = 80$ containers from a production lot and measures the fat content (in grams, g) of each.

20. To find a confidence interval for the population mean we use

$$\text{estimate} \pm \text{multiplier} \times \text{standard error}.$$

Which formula is appropriate for constructing a 95% confidence interval for the population mean fat content?

A. $\hat{p} \pm z_{0.975} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

B. $\bar{y} \pm t_{79,0.975} \times \frac{\sigma}{\sqrt{n}}$.

C. $\mu \pm z_{0.975} \times \frac{\sigma}{\sqrt{n}}$.

D. $\bar{y} \pm z_{0.975} \times \frac{s}{\sqrt{n}}$.

E. $\bar{y} \pm t_{79,0.975}\, \frac{s}{\sqrt{n}}$.

21. What happens to the confidence interval if we increase the confidence level from 95% to 99%?

A. The estimate changes; the effect on the interval is unclear.

B. The standard error changes; the interval is wider.

C. The standard error changes; the interval is narrower.

D. The multiplier changes; the interval is wider.

E. The multiplier changes; the interval is narrower.

22. The resulting 99% confidence interval for μ is (3.54, 3.82). Select the best interpretation below.

A. 99% of yogurt containers in the sample have fat content between 3.54 g and 3.82 g.

B. There is a 99% chance that the fat content of a randomly chosen yogurt container is between 3.54 g and 3.82 g.

C. There is a probability of 0.99 that the population mean lies between 3.54 g and 3.82 g.

D. If we repeated this process many times, 99% of the sample means would fall between 3.54 g and 3.82 g.

E. We are 99% confident that the true mean fat content of the yogurt in this production lot lies between 3.54 g and 3.82 g.

---

23. Suppose we carry out a one-sample hypothesis test for the population mean μ:

$$\mathrm{H}_0 : \mu = \mu_0; \qquad \mathrm{H}_A : \mu \neq \mu_0.$$

If $\alpha = 0.01$ and we find a p-value of 0.000001, which of the following statements is most appropriate?

A. Since p-value is very small, there must be a large effect. The sample mean $\bar{y}$ will be a long way away from μ_0.

B. We can be certain that $\mu \neq \mu_0$.

C. Since p-value $< \alpha$, the data observed are unlikely if μ truly was μ_0.

D. The probability that $\mu = \mu_0$ is 0.000001.

E. The probability that our calculations are correct is 0.000001.

---

24. Select the option that best describes estimation of a statistical model?

A. Estimation is where we assume the population mean is the sample mean.

B. Estimation is where we find the standard error by taking the standard deviation divided by the square root of the sample size.

C. Estimation is the process of finding the statistic using population data.

D. Estimation is the process of guessing data values so that the model fits as well as possible.

E. Estimation is using a statistic to make an educated guess about an unknown parameter.

Information for Questions 25 to 26

The R object `penguin` contains information on a random sample of chinstrap penguins from the Palmer archipelago. There are two variables: `bill`, the bill length (mm), and `flipper`, the flipper length (mm). We consider the R code below:

```
mean(penguin$bill)
sd(penguin$flipper)
```

25. What is being evaluated in the first line of R code: `mean(penguin$bill)`?

A. The sample mean of the bill length.

B. The sample standard deviation of the bill length

C. The sample mean of the flipper length.

D. The sample standard deviation of the flipper length.

E. The sample median bill length.

26. What is being evaluated in the second line of R code: `sd(penguin$flipper)`?

A. The sample mean of the bill length.

B. The sample standard deviation of the bill length

C. The sample mean of the flipper length.

D. The sample standard deviation of the flipper length.

E. The sample median bill length.

Information for Questions 27 to 30

A study is carried out in a forest ecosystem to assess the effect of a controlled burn on the soil nitrogen content. The research selects 30 random locations, and collects a soil sample before a scheduled controlled burn, and another after the burn has occurred. The data are in the R object `burn`. There are two variables `before`, which gives the nitrogen levels (mg/kg) before the burn, and `after`, which gives the nitrogen levels after the burn. We can assume that the data are approximately normally distributed. The first few observations are given below:

```
head(burn)
```

```
##   before after
## 1   4.30  4.18
## 2   4.63  4.52
## 3   4.61  4.74
## 4   4.58  4.53
## 5   4.84  5.00
## 6   4.48  4.22
```

27. The R code below carries out the hypothesis test:

$$\mathrm{H}_0 : \mu_d = 0; \qquad \mathrm{H}_A : \mu_d \neq 0,$$

where μ_d is the mean difference in the nitrogen levels (after-before). If $\alpha = 0.05$, select the best interpretation below:

```
burn$diff = burn$after - burn$before
t.test(burn$diff)
```

```
##
##  One Sample t-test
##
## data:  burn$diff
## t = -2.2573, df = 29, p-value = 0.0317
```

```
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.154391169 -0.007608831
## sample estimates:
## mean of x
##    -0.081
```

A. Since the p-value $< \alpha$, the observed data would be unusual if there were truly no difference in mean nitrogen levels.

B. Since p-value $< \alpha$, the observed data would be unusual if the true mean nitrogen level after the burn was 0.

C. Since p-value $< \alpha$, the observed data would not be unusual if there were truly no difference in mean nitrogen levels.

D. Since p-value $> \alpha$, the correlation between the before and after observations is 0.

E. Since p-value $< \alpha$, there should be no more controlled burns in this area.

28. Which model is used in Question 27?

A. A normal model with paired data.

B. A normal model with two (independent) groups.

C. A chi-squared test.

D. A linear regression model.

E. A binomial model.

29. In the context of the hypothesis test carried out in Question 27, what is power?

A. The probability that the null hypothesis is true.

B. The probability that we reject the null hypothesis when there truly is no difference in mean nitrogen levels before and after the burn.

C. The probability of rejecting the null hypothesis when there truly is a difference in mean nitrogen levels before and after the burn.

D. The power is the same thing as the p-value.

E. The probability that the p-value is less than 0.05.

30. Which of the following options will definitely increase power?

A. Using `t.test` with the option `paired = true`.

B. Increasing α and decreasing sample size.

C. Decreasing α.

D. Replicating the study.

E. Increasing the sample size.

Information for Questions 31 to 33

A clinical trial is conducted to compare the effectiveness of two blood pressure medications: Drug A and Drug B. Researchers recruit 120 adult participants with high blood pressure. They randomly assign 60 participants to receive Drug A, and 60 to receive Drug B. After 8 weeks, the reduction in systolic blood pressure is measured for each participant. The data are in the R object `bp`. There are two variables `reduction` and `drug`. A `reduction` of 10 means that the systolic blood pressure reduced by 10 units across the 8 weeks. For the questions below, we take group 1 to be those receiving drug A, and take group 2 to be those receiving drug B. The first few observations are shown below:

```
head(bp)
```

```
##   reduction drug
## 1      16.0    A
## 2      21.6    A
## 3      33.6    A
## 4       9.1    A
## 5       4.7    A
## 6      23.8    A
```

31. The sample mean reduction for drug A is $\bar{y}_1 = 19.00$ with sample standard deviation $s_1 = 13.579$. The sample mean reduction for drug B is $\bar{y}_2 = 15.95$ with sample standard deviation $s_2 = 9.054$. The estimated standard error for $\bar{y}_1 - \bar{y}_2$ is closest to:

A. 2.11 B. 4.44 C. 0.61 D. 2.92 E. 16.32

32. We obtain separate data frames for drug A and B in R:

```
drugA = subset(bp, drug == "A")
drugB = subset(bp, drug == "B")
```

Which of the following options should we use to find a 95% confidence interval for $\mu_1 - \mu_2$?

A. `t.test(drugA$reduction, drugB$reduction)`

B. `t.test(drugA$reduction, drugB$reduction, paired = TRUE)`

C. `t.test(drugA$drug, drugB$drug)`

D. `prop.test(drugA$reduction, drugB$reduction)`

E. `t.test(bp$reduction)`

33. The R output of a suitable model is below. Select the best interpretation:

```
...
## t = 1.4452, df = 102.81, p-value = 0.1514
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.133774  7.223774
## sample estimates:
## mean of x mean of y
##    18.995    15.950
```

A. We are 95% confident that the true mean reduction in systolic blood pressure when taking drug A is between -1.13 and 7.22.

B. We are 95% confident that drug A is more effective than drug B at reducing systolic blood pressure preferred since $\bar{y}_1 > \bar{y}_2$ since $19.00 > 15.95$.

C. We are 95% confident that a randomly selected individual would experience a reduction in systolic blood pressure of exactly $\bar{y}_1 - \bar{y}_2 = 19.00 - 15.95 = 3.05$ if they were to move from drug B to drug A.

D. We are 95% confident that neither drug A nor drug B lead to a reduction in mean systolic blood pressure since the confidence interval (-1, 7) spans 0.

E. We are 95% confident that the true difference in mean systolic blood pressure reduction between drug A and drug B (drug A - drug B) is between -1.13 and 7.22.

Information for Questions 34 to 36

Suppose that we know for small to moderate mammals (those weighing up to 120 kg), the resting heart rate (beats per minute, bpm) can be described as a linear relationship with body mass (kg). The true relationship is

$$\texttt{resting heart rate} = 150 - 0.75 \times \texttt{body mass} + \varepsilon,$$

where ε is normally distributed with mean 0 and a standard deviation of 10.

34. Based on the true model, which of the following options is not correct?

A. Among the subpopulation of mammals with body mass of 50 kg, the expected resting heart rate is 112.5 bpm.

B. Larger mammals have a lower resting heart rate, on average.

C. On average, mammals with body mass 75 kg have a resting heart rate that is 0.75 bpm lower than mammals with body mass 74 kg.

D. All mammals with body mass of 100 kg have a lower resting heart rate than those with body mass of 50 kg.

E. Among the subpopulation of mammals with body mass of 100 kg, the resting heart rate is normally distributed with mean 75 bpm and standard deviation of 10.

35. Which of the following is correct for the subpopulation of mammals that have body mass of 20 kg?

A. The distribution of resting heart rate is normally distributed with mean 135 bpm and standard deviation of 10.

B. The distribution of resting heart rate is normally distributed with mean 148.5 bpm and standard deviation of 10.

C. They will all have a resting heart rate of 135 bpm.

D. They will all have a resting heart rate of 148.5 bpm.

E. It is impossible to say anything about this subpopulation.

36. What is the best interpretation of β_1?

A. The expected heart rate decreases by 0.75 bpm for a one kg increase in mammal weight.

B. The expected heart rate increases by 0.75 bpm for a one kg increase in mammal weight.

C. An individual mammal's heart rate will decrease by 0.75 bpm if they increase their weight by 1 kg.

D. An individual mammal's heart rate will increase by 0.75 bpm if they increase their weight by 1 kg.

E. The expected heart rate of a mammal with body mass of 0 kg is 150 bpm.

Information for Questions 37 to 46

Data on aptitude and speaking age of 21 children is contained in the R object `speak`. The two variables are `age`, the age at which the child first speaks (months), and `aptitude`, the outcome of a Gesell aptitude test that the child completes later in life. The variable `age` ranges from 7 months to 42 months. The variable `aptitude` ranges from 57 to 121. The first few observations are shown below:

```
head(speak)
```

```
##   age aptitude
## 1  15       95
## 2  26       71
## 3  10       83
## 4   9       91
## 5  15      102
## 6  20       87
```

37. The correlation between aptitude and age is:

```r
cor(speak$age, speak$aptitude)
```

```
## [1] -0.64029
```

Select the most appropriate description:

A. The sample correlation is negative. When age is larger than average, aptitude is likely to be below average.

B. The sample correlation is positive. When age is larger than average, aptitude is likely to be larger than average.

C. There is a no apparent linear relationship between aptitude and age. When age is larger than average, it tells us little about the aptitude.

D. There is a positive non-linear relationship between aptitude and age.

E. There is a negative non-linear relationship between aptitude and age.

38. What is the best description of the method used to estimate the parameters in the linear regression model below?

```r
m_speak = lm(aptitude ~ age, data = speak)
```

A. We use a pencil to draw a straight line on the plot and determine the estimates accordingly.

B. The estimates are chosen to make the sum of residuals zero.

C. The computer tries many different lines and selects the one that has the highest (absolute) correlation.

D. The estimates are chosen so that they minimise the sum of squared residuals.

E. The estimates are chosen so the line goes through all the data points exactly.

39. Select the correct expression for the fitted regression model based on the R output below:

```r
summary(m_speak)
```

```
...
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 109.8738     5.0678  21.681 7.31e-15 ***
## age          -1.1270     0.3102  -3.633  0.00177 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.02 on 19 degrees of freedom
```

```
## Multiple R-squared:   0.41,Adjusted R-squared:  0.3789
## F-statistic:  13.2 on 1 and 19 DF,  p-value: 0.001769
```

A. $\widehat{\texttt{aptitude}} = -1.13 + 109.87 \times \texttt{age}$

B. $\widehat{\texttt{aptitude}} = 109.87 - 1.13 \times \texttt{age} + \varepsilon$

C. $\widehat{\texttt{aptitude}} = 109.87 - 1.13 \times \texttt{age}$

D. $\widehat{\texttt{age}} = 109.87 - 1.13 \times \texttt{aptitude}$

E. $\widehat{\texttt{age}} = 109.87 - 1.13 \times \texttt{aptitude} + \epsilon$

40. The p-value in the hypothesis test with $\mathrm{H}_0 : \beta_1 = 0$; $\mathrm{H}_A : \beta_1 \neq 0$ is:

A. 0.00000 B. 0.3102 C. 0.41 D. 0.3789 E. 0.00177

41. The standard error for $\hat{\beta}_1$ is 0.310. Which of the following best describes what this standard error represents?

A. The standard error tells us the average error made when predicting $\hat{y}$

B. The standard error is the mean of the sampling distribution of $\hat{\beta}_1$.

C. The standard error is an estimate of the variability in the observed y-values around the regression line.

D. The standard error is the difference between the observed slope and the true slope.

E. The standard error describes the variability in the sampling distribution of $\hat{\beta}_1$.

42. If the appropriate multiplier is 2.093, the 95% confidence interval for β_1 is closest to:

A. (-1.44, -0.82) B. (99.26, 120.48) C. (-1.78, -0.48) D. (-11.74, 9.48) E. (-2.43, 0.17)

43. Does it make sense to interpret $\hat{\beta}_0$ in this application?

A. Yes, because it tells us the expected aptitude score for a child who first speaks at age 0 months.

B. Yes, because the intercept is always meaningful in a regression model.

C. No, because R^2 is too low for any interpretation to be useful.

D. No, because the intercept has high standard error.

E. No, because a child first speaking at age 0 months is outside the range of the data and likely impossible, so the expected aptitude for a child first speaking at 0 months is scientifically meaningless.

44. The code below finds two intervals. The type of interval is hidden (we have replaced the type of interval by `A` and `B`). Select the best description of these intervals:

```
predict(m_speak, newdata = data.frame(age = 12),
        interval = "A")
```

```
##        fit     lwr      upr
## 1 96.34997 72.6853 120.0146
```

```
predict(m_speak, newdata = data.frame(age = 12),
        interval = "B")
```

```
##        fit      lwr      upr
## 1 96.34997 91.08348 101.6165
```

A. Interval A is a 95% confidence interval for mean response; Interval B is a 95% prediction interval.

B. Both intervals are 95% prediction intervals for different values of age. The narrower interval is evaluated at the mean age value.

C. Interval A is for children with below-average aptitude; Interval B is for those above average.

D. Interval A is a 95% prediction interval; Interval B is a 95% confidence interval for mean response.

E. Interval B is narrower because it was computed using a larger sample size than Interval A.

45. The researchers want to use the model to predict the aptitude of a child who first speaks at 60 months. This quantity is closest to:

A. 177 B. 6591 C. 104 D. 42 E. 60

46. Can the prediction from Question 45 be reliably interpreted?

A. Yes. As the model has an R^2 above 0.3, any prediction using this model is reliable.

B. Yes. The estimated standard deviation s_ε is small, which means the model can be used for reliable predictions.

C. No. The negative correlation means that predictions cannot be trusted.

D. No. This is an extrapolation, and the model may not be appropriate outside the range of the observed data.

E. No. Predictions from linear models are never trustworthy in real-world situations.

Information for Questions 47 to 51

Researchers design a study to explore the relationship between metabolic rate (Kcal/day), temperature (degrees Celsius) and activity (hours/day). The data are available in the R object `meta`. There are three variables: `metabolic_rate`, `temperature`, and `activity`. The temperature values ranged from -4.5 to 24.8 degrees. The activity values ranged from 0 to 10 hours. The first few observations are below:

```r
head(meta)
```

```
##   metabolic_rate temperature activity
## 1           1877         0.0      2.9
## 2           1716         8.2      5.0
## 3           1342        24.8      5.8
## 4           1839        22.9      9.7
## 5           1561        -0.1      0.5
## 6           2019        10.2      9.8
```

47. Which of the following statements about multiple linear regression is correct?

A. There can be no more than two predictors in multiple linear regression

B. We are unable to include categorical predictors in multiple linear regression

C. We no longer use `lm` in R. We must use `mlm` instead.

D. Multiple linear regression allows us to assess the effect of each predictor while controlling for the others.

E. It only makes sense to interpret the effect of one predictor, even if we include several.

48. The researchers first fit a model with only temperature (output below). Which of the following statements about the estimate of β_0 is accurate?

```r
m_meta = lm(metabolic_rate ~ temperature, data = meta)
summary(m_meta)
```

```
##
## Call:
## lm(formula = metabolic_rate ~ temperature, data = meta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -480.14 -113.85   25.17  110.61  464.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1873.897     27.583  67.937  < 2e-16 ***
## temperature  -15.394      2.123  -7.253 9.61e-11 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 174.9 on 98 degrees of freedom
```

```
## Multiple R-squared:  0.3493,Adjusted R-squared:  0.3426
## F-statistic:  52.6 on 1 and 98 DF,  p-value: 9.609e-11
```

A. We estimate the expected metabolic rate when temperature is 0 to be 1873.9. It is appropriate to interpret $\hat{\beta}_0$ as temperature $= 0$ is in the range of the data.

B. We do not interpret $\hat{\beta}_0 = 1873.9$. It is biologically unreasonable for small mammals to survive in temperatures at or close to 0 degrees.

C. We should not interpret the intercept $\hat{\beta}_0 = 1873.9$, as it is an extrapolation due to a temperature of 0 being outside the range of the data.

D. We estimate that metabolic rate changes by 1873.9 for a one unit increase in temperature. It is always appropriate to interpret $\hat{\beta}_0$.

E. We never interpret the intercept in a linear regression model.

49. Researchers then fit a model that includes both temperature and activity. Select the option that gives $\hat{\beta}_2$ and the standard error for $\hat{\beta}_2$.

```
m_meta2 = lm(metabolic_rate ~ temperature + activity, data = meta)
summary(m_meta2)
```

```
##
## Call:
## lm(formula = metabolic_rate ~ temperature + activity, data = meta)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -214.327  -62.236    1.634   60.968  197.781
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1623.692     19.378   83.79   <2e-16 ***
## temperature  -18.627      1.046  -17.81   <2e-16 ***
## activity      52.641      2.947   17.86   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.91 on 97 degrees of freedom
## Multiple R-squared:  0.8483,Adjusted R-squared:  0.8451
## F-statistic: 271.1 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
confint(m_meta2)
```

```
##                  2.5 %     97.5 %
## (Intercept) 1585.23153 1662.15219
## temperature  -20.70283  -16.55056
## activity      46.79078   58.49061
```

A. $\hat{\beta}_2 = 1623.69$, $s_{\hat{\beta}_2} = 19.38$.

B. $\hat{\beta}_2 = 52.64$, $s_{\hat{\beta}_2} = 2.95$.

C. $\hat{\beta}_2 = -18.63$, $s_{\hat{\beta}_2} = 1.05$.

D. $\hat{\beta}_2 = 52.64$, $s_{\hat{\beta}_2} = 19.38$.

E. $\hat{\beta}_2 = -18.63$, $s_{\hat{\beta}_2} = 2.95$.

50. Select the best interpretation of the confidence interval for β_2.

A. We are 95% confident that the expected metabolic rate increases by between 46.8 and 58.5 for a one hour increase in activity.

B. We are 95% confident that the expected metabolic rate is between 46.8 and 58.5 for a mammal with one hour of activity.

C. We are 95% confident that the expected metabolic rate increases by between -20.7 and -16.6 for a one hour increase in temperature, holding activity fixed.

D. We are 95% confident that the expected metabolic rate increases by between 46.8 and 58.5 for a one hour increase in activity, holding temperature fixed.

E. We are 95% confident that activity and temperature affect metabolic rate by between -18.6 and 52.6.

51. The R^2 is 84.8%. Which of the statements below is not correct.

A. R^2 is the squared correlation between y and $\hat{y}$.

B. R^2 has to be between 0 and 1.

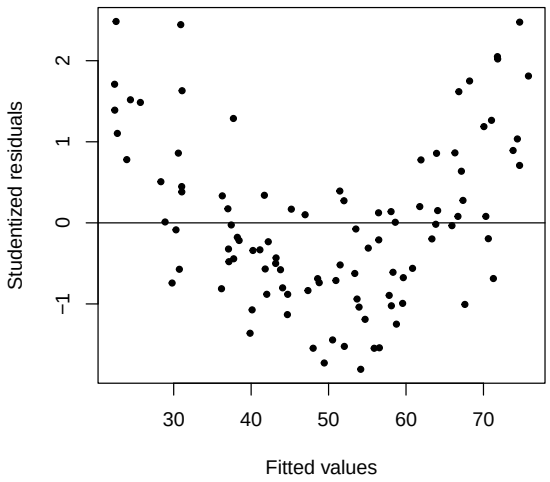
C. R^2 is the proportion of variance in the outcome variable that is explained by the predictors.

D. The R^2 value is provided in the R model output with the label `Multiple R-squared`.

E. A regression model is of little practical use unless $R^2 > 0.5$.

---

52. Suppose that we fit a linear regression model with outcome y and predictor variable x. Based on the plot below, select the option that best describes which regression assumptions, if any, appear to be violated.



A. There appears to be a violation of the linearity assumption.

B. There appears to be a violation of the independence assumption.

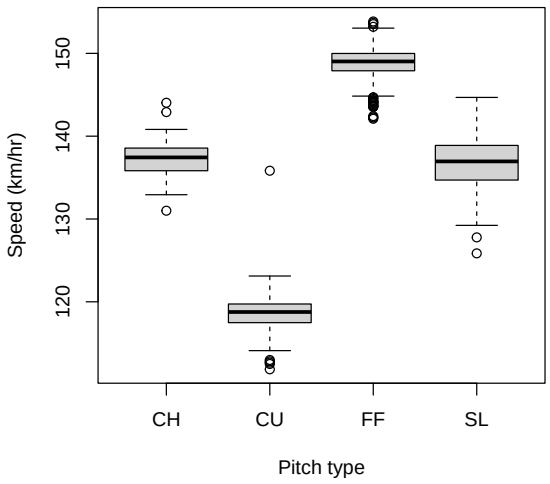
C. There are observations that appear to be outliers.

D. There appears to be non-constant variance.

E. There are no obvious violations to the regression assumptions.

Information for Questions 53 to 59

Data in the R object `baseball` contains information about all 3402 of Clayton Kershaw's pitches in the 2014 Major League Baseball season. There are two variables. The first is `speed`, the speed of each pitch as it leaves Kershaw's hand (kilometers per hour, km/hr). The second is `type`, which gives the type of pitch. There are four possible pitches: 'CH' (changeup), 'CU' (curveball), 'FF' (fastball), and 'SL' (slider). A boxplot of the data, as well as the ANOVA table, is below




```
a_baseball = aov(speed ~ type, data = baseball)
summary(a_baseball)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## type           3 355921  118640   27890 <2e-16 ***
## Residuals   3398  14455       4
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

53. Select the hypotheses that are being tested with ANOVA:

A. $H_0 : \mu_1 - \mu_2 = 0$; $H_A : \mu_1 - \mu_2 \neq 0$.

B. H_0 : the variables are independent; H_A : the variables are related.

C. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$; $H_A : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$.

D. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$; H_A : at least one mean is different.

E. H_0 : the distribution for the groups are the same; H_A: the distribution for the groups all differ.

54. The F-value for the appropriate test in Question 53 is closest to

A. 27890 B. 355921 C. 118640 D. 4 E. 14455

55. Select the option that is not correct with respect to ANOVA:

A. ANOVA compares the variance among the groups to the variance within groups.

B. The F-value is a realisation from an F-distribution if the null hypothesis is true.

C. If the group means explain a lot of the variation in the data, the F-value will be large.

D. The degrees of freedom shown in the ANOVA table are the two values needed to determine the appropriate F-distribution for the test.

E. The ANOVA table tells us which group means differ from one another.

56. For this example, what distribution do we compare the F-value to in order to determine the p-value?

A. A standard normal distribution.

B. A χ^2-distribution with df = 3 degrees of freedom.

C. A F-distribution with df1 = 3, df2 = 3398 degrees of freedom.

D. A χ^2-distribution with df = 3398 degrees of freedom.

E. A t-distribution with df = 3398 degrees of freedom.

57. With $\alpha = 0.01$, select the best interpretation of the p-value from the ANOVA test:

A. As p-value $< \alpha$, the data are unusual if the group means are truly the same.

B. As p-value $< \alpha$, Clayton Kershaw's slider and changeup are thrown at the same speed.

C. As p-value $< \alpha$, all of the group means differ from one another.

D. As p-value $< \alpha$, there is evidence that Clayton Kershaw is a difficult pitcher to face.

E. As p-value $< \alpha$, we can conclude from the ANOVA that Kershaw's curveball is his slowest pitch, while his fastball is the fastest.

58. Select the best description of what the following R code does:

```
TukeyHSD(a_baseball)
```

A. It fits a separate linear model for each pitch type and compares the regression slopes.

B. It tests whether each pitch type has the same variance in speed.

C. It compares the means for each pair of pitch types, accounting for multiple comparisons.

D. It provides an updated ANOVA table comparing pitch types.

E. It tests whether the overall mean pitch speed differs from the mean for each individual pitch type.

59. Select the best interpretation of the row `FF-CU`

```
TukeyHSD(a_baseball)
```

```
...
## $type
##               diff        lwr         upr     p adj
## CU-CH -18.6362165 -19.272387 -18.0000460 0.0000000
## FF-CH  11.6830846  11.089594  12.2765749 0.0000000
```

```
## SL-CH  -0.4597739  -1.069926   0.1503777 0.2128146
## FF-CU  30.3193011  30.036871  30.6017308 0.0000000
## SL-CU  18.1764426  17.860498  18.4923869 0.0000000
## SL-FF -12.1428585 -12.360433 -11.9252841 0.0000000
```

```
tukey_baseball = TukeyHSD(a_baseball)
```

A. We are 95% confident that for every 100 pitches that Clayton Kershaw throws, there will be between 30.04 and 30.60 more fastballs than curveballs.

B. We are 95% confident that the average starting speed of fastballs is between 30.04 and 30.60 km/hr faster than non-fastballs

C. We are 95% confident that for every 100 pitches that Clayton Kershaw throws, there will be between 30.04 and 30.60 strikes.

D. We are 95% confident that the average starting speed of fastballs is between 30.04 and 30.60 km/hr faster than curveballs.

E. There is a probability of 0.95 that the average starting speed of fastballs is between 30.04 and 30.60 km/hr faster than curveballs.

---

60. Which of the following in not an example of binary data?

A. The number of eggs on a nest.

B. Presence or absence of a particular gene.

C. Breeding status (breeder or non-breeder) of dolphin.

D. A variable that describes whether or not a participant completed the task within the time limit.

E. Age status (adult or juvenile) of chinstrap penguin.

---

61. Researchers are studying how frogs respond to a predator cue. They expose individual frogs to the cue and record whether each frog jumps away (yes/no). They continue collecting data until they observe 20 frogs that run away. Which of the binomial assumptions, if any, are violated?

A. There are two possible outcomes: a frog jumps away, or it does not.

B. n, the number of trials is fixed.

C. p, the probability that a frog jumps away is the same in each trial.

D. The outcome is independent among frogs.

E. None of the binomial assumptions are clearly violated.

Information for Questions 63 to 66

Data from 8520 field goals from the NFL (American football) is provided in the contingency table below. The first variable is `distance`. This is classified as either closer than 50 yards ('<50'), or 50 yards and further ('50+'). The second variable is `made` and represents whether the field goal is made ('yes'), or missed ('no').

```
##          made
## distance   no  yes  Sum
##      <50 1299 6430 7729
##      50+  372  419  791
##      Sum 1671 6849 8520
```

62. The sample proportion of field goals made from less than 50 yards is closest to

A. 0.83 B. 0.53 C. 0.94 D. 0.75 E. 0.78

63. The sample proportion of field goals made from 50 yards or further is closest to

A. 0.83 B. 0.53 C. 0.94 D. 0.75 E. 0.78

64. A confidence interval can be found using `prop.test` as below. Select the best description of the parameter being estimated by the confidence interval shown in the output

```
prop.test(x = c(6430, 419), n = c(7729, 791))
```

```
...
## 95 percent confidence interval:
##  0.2657580 0.3386869
...
```

A. $\mu_1 - \mu_2$: the true difference in the mean number of field goals scored ('<50' - '50+' yards).

B. p_1: the true probability of making a field goal from < 50 yards.

C. $p_1 - p_2$: the true difference in the probability of scoring a field goal ('<50' - '50+' yards).

D. p_2: the true probability of making a field goal from 50 yards or longer.

E. p: the true marginal probability of making a field goal.

65. What hypothesis test is being carried out when using `prop.test` as in Question 64?

A. $\mathrm{H}_0 : \mu_1 - \mu_2 = 0$; $\mathrm{H}_A : \mu_1 - \mu_2 \neq 0$.

B. $\mathrm{H}_0 : p = 0.5$; $\mathrm{H}_A : p \neq 0.5$.

C. $\mathrm{H}_0 : p_1 - p_2 \neq 0$; $\mathrm{H}_A : p_1 - p_2 = 0$.

D. $\mathrm{H}_0 : p_1 - p_2 = 0$; $\mathrm{H}_A : p_1 - p_2 \neq 0$.

E. $H_0 : \mu_1 - \mu_2 \neq 0$; $H_A : \mu_1 - \mu_2 = 0$.

66. A friend suggests that we should use `chisq.test` instead of `prop.test` in this application. Select the best description of the difference between them

A. There will be no difference; the test is identical.

B. The `prop.test` test is more powerful, so the p-value from `prop.test` is likely to be lower.

C. The `chisq.test` test is more powerful, so the p-value from `chisq.test` is likely to be lower.

D. They use different test statistics, and will give different p-values. Neither test is better than the other in the long-run. We should use whichever gives a lower p-value.

E. The `chisq.test` test is appropriate for use with contingency tables, whereas `prop.test` should only be used for single proportions.

Information for Questions 67 to 71

A total of 6272 Swedish men were followed for 30 years. Of interest was whether there was any association between the amount of fish in their diet and prostate cancer. The data can be summarized in a contingency table.

```
##           cancer
## diet         no  yes  Sum
##   large     507   42  549
##   moderate 2769  209 2978
##   small    2420  201 2621
##   never     110   14  124
##   Sum      5806  466 6272
```

67. If we assume independence between diet and cancer, the expected count of those with a moderate diet of fish and no cancer is closest to

A. 221.3 B. 2756.7 C. 2426.3 D. 1.4 E. 2769.0

68. If we assume independence between diet and cancer, the expected count of those who never eat fish and have prostrate cancer is closest to

A. 194.7 B. 114.8 C. 0.1 D. 9.2 E. 14.0

69. What is the appropriate hypotheses for the χ^2-test for contingency tables.

A. $H_0 : p_1 - p_2 = 0$; $H_A : p_1 - p_2 \neq 0$.

B. $H_0 : \mu_1 - \mu_2 = 0$; $H_A : \mu_1 - \mu_2 \neq 0$.

C. H_0 : diet and cancer are independent; H_A : diet and cancer are related.

D. H_0 : the distribution for the two groups are the same; H_A: the distribution for the two groups differ.

E. $H_0 : p_1 - p_2 \neq 0$; $H_A : p_1 - p_2 = 0$.

70. What are the degrees of freedom for the χ^2-test?

A. 1 B. 2 C. 3 D. 4 E. 6

71. Select the best interpretation from the χ^2-test below if $\alpha = 0.05$.

```
##
##  Pearson's Chi-squared test
##
## data:  tabfish
## X-squared = 3.6773, df = 3, p-value = 0.2985
```

A. The data suggest that there are more men without cancer than there are with it.

B. The data are unusual if the null hypothesis is correct.

C. The chi-squared test proves there is no relationship between diet and cancer.

D. The data are not unusual given that diet and cancer are truly independent.

E. The sample size is too small to detect any effect.

Information for Questions 72 to 75

Data were collected from a sample of 36 deceased Catholic priests, 17 of whom had mild to moderate Alzheimer's disease, and 19 of whom had no cognitive impairment. The amyloid-beta protein level (pmol/g) was measured for each of the priests. The data are contained in the R object `protein`. There are two variables: `value`, which gives the amyloid-beta level, and `group`, which specifies whether the observation is from a priest with Alzheimer's disease ('AD'), or from a priest with no cognitive impairment ('NCI'). The first few lines of the R code are below.

```
head(protein)
```

```
##   value group
## 1    49   NCI
## 2   537   NCI
## 3   407    AD
## 4   894    AD
## 5  1496    AD
## 6   439    AD
```

72. We are unwilling to rely on either normality or the central limit theorem to compare the amyloid-beta level between the two groups. Instead we use a non-parametric Mann-Whitney test. Select the hypotheses we test with the Mann-Whitney test

A. $H_0 : \mu_1 - \mu_2 = 0$; $H_A : \mu_1 - \mu_2 \neq 0$.

B. $H_0 : \mu_1 - \mu_2 \neq 0$; $H_A : \mu_1 - \mu_2 = 0$.

C. H_0 : the distribution of amyloid-beta protein is the same for Alzheimer's and no impairment groups; H_A: the distribution of amyloid-beta protein differs for Alzheimer's and no impairment groups.

D. $H_0 : p_1 - p_2 = 0$; $H_A : p_1 - p_2 \neq 0$.

E. We cannot test hypotheses with non-parametric approaches.

73. Select the option that best describes how the Mann-Whitney test statistic is found:

A. Comparing the medians of amyloid-beta protein values in the two groups.

B. Comparing the standard deviation of amyloid-beta protein values in the two groups.

C. Comparing the means of amyloid-beta protein values in the two groups.

D. Comparing the number of observations above the mean in each group.

E. Comparing the sum of the ranked amyloid-beta protein values in the two groups.

74. Which of the following is a benefit of using a non-parametric test such as the Mann-Whitney test?

A. It is more powerful than parametric alternatives

B. It is easy to obtain confidence intervals for the mean difference.

C. It is robust to skew and outliers.

D. It determines causality even with observational data.

E. It always has a smaller p-value than the corresponding t-test.

75. Interpret the test carried out below if $\alpha = 0.05$:

```
alz = subset(protein, group == "AD")
nci = subset(protein, group == "NCI")
wilcox.test(alz$value, nci$value)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  alz$value and nci$value
## W = 258, p-value = 0.002336
...
```

A. The difference in means of amyloid-beta protein values between the two groups is statistically significant.

B. There is probability of 0.0023 that the amyloid-beta protein distribution is the same for both groups.

C. The data provide evidence that there is no difference between the two groups.

D. The results suggests that the amyloid-beta protein values are independent of the group variable.

E. The data are unusual if the distribution of amyloid-beta protein is the same for both groups.

---

76. Which of the following best describes stratified sampling?

A. Selecting every 10th person from a list.

B. Dividing the population into groups and selecting everyone who is in a small number of randomly chosen groups.

C. Dividing the population into groups and randomly sampling from each group.

D. Allowing people to volunteer to participate.

E. Randomly assigning people into groups after collecting data.

---

77. Military conscripts (45,570 young men) in Sweden completed a questionnaire that recorded whether or not they had used cannabis. Fifteen years later the national medical database was searched to determine which of these men had gone on to develop schizophrenia. The percentage of cannabis users who developed schizophrenia was compared with the percentage of non-users who developed schizophrenia. If the population of interest is adult Swedish males with military service, which of the following is likely to be the largest source of bias, and why?

A. Selection bias; only participants with prior cannabis use are recruited into the study.

B. Selection bias; the selection criteria exclude females from being in the study.

C. Information bias; people may be unlikely to tell the truth when asked about cannabis use.

D. Information bias; people may not be able to recall events that happened 15 years ago.

E. There are no apparent sources of bias.

---

78. In the 'Warrior Gene' example, which of the following was done by the research team?

A. Allowed the study participants and the communities that they came from, to have control over the narratives arising from the research.

B. Compared the monoamine oxidase (MAO-A) gene frequencies in a Māori cohort, to other ethnicities.

C. Extensive consultation with Māori communities prior to undertaking the research.

D. Co-designing research with Māori community representatives.

E. Designing and conducting the research using a tikanga Māori framework.

---

79. Which of the following is not a characteristic of co-designed research studies?

A. Researchers involve members of the community/ies they are working with in all aspects of study design and conduct.

B. Studies are outcome focused

C. Community representatives have a role determining what the outcomes should be

D. The researchers do not burden the community/ies they are working in with any details of the study, and do not involve them in the research process.

E. Community representatives have a role determining what benefits should be

---

80. In Indigenous data soverignty, the CARE acronym refers to which of the following:

A. Collective benefits, Authority to Control, Responsibility and Ethics.

B. Collective employment contracts, Access to data, Repeatability and Estimation.

C. Collective benefits, Access to data, Repeatability and Estimation

D. Cultural integrity, Access to data, Repeatability and Estimation

E. Corporate relevance, Access to data, Responsibilities to those who wish to access indigenous data, and Ethics

---

81. What is the best description of a placebo group?

A. A control group that receives no treatment.

B. A group that is given a real treatment, but does not understand the purpose of the study.

C. A group given a fake treatment that resembles an actual treatment.

D. A group used in the pilot study to calculate the sample size.

E. A group excluded from the final results because of outliers in the recorded data.

---

82. What is a confounding variable?

A. A variable that affects both the predictor and outcome variables.

B. A variable intentionally excluded because it is thought to be irrelevant or redundant.

C. A variable that is difficult, time-consuming, or too expensive to collect in most studies.

D. A variable associated with the outcome but considered unimportant to the study question.

E. A variable used to inform the posterior distribution in Bayesian inference.

---

83. What is the replication crisis in science primarily about?

A. Many scientific studies cannot be reproduced or validated by others.

B. Studies being published in too many journals.

C. Research that is too theoretical to be useful.

D. Statistics being too complicated for researchers.

E. Overuse of Bayesian inference.

---

84. What is the main danger of performing many statistical tests without adjustment?

A. Mistaking one test for another.

B. Increased chance of a type I error.

C. Increased complexity of analysis.

D. Increased difficulty in interpreting many tests.

E. More accurate estimates of population parameters.

---

85. What is the best description of HARKing?

A. The inappropriate manipulation of data or analysis approaches to enable a statistically significant result.

B. Highlighting all relevant knowledge sources.

C. Presenting exploratory results as if they were a hypothesis determined before data collection.

D. Including a second predictor variable into a linear regression model.

E. Deliberately hiding unfavourable statistical results.

---

86. Which of the following best describes maximum likelihood estimation (MLE)?

A. A method that uses prior beliefs to update the probability of a parameter.

B. An approach for parameter estimation that can only be used if the sample size is not too large.

C. A method that finds the parameter values that are most likely for the observed data.

D. A tool for finding the smallest possible p-value for any given model.

E. A procedure for estimating parameters that is theoretically valuable, but is never used in practice.

---

87. Which of the following is a feature of Bayesian inference?

A. It is not used in applied research.

B. The posterior distribution is based only on the prior beliefs.

C. Probability is used to express uncertainty about parameter values.

D. Bayesian inference cannot be applied to real-world data.

E. The likelihood function is not used in Bayesian inference.

Summary of Formulae

Sample mean and variance

$$\text{Mean: } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{Variance: } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Probability Rules

$$\begin{aligned}
\Pr(A \text{ or } B) &= \Pr(A) + \Pr(B) - \Pr(A \text{ and } B) \\
\Pr(A \text{ and } B) &= \Pr(A)\Pr(B|A) = \Pr(B)\Pr(A|B) \\
\Pr(A|B) &= \frac{\Pr(A \text{ and } B)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|A^{\complement})\Pr(A^{\complement})}
\end{aligned}$$

Random Variables

If X and Y are random variables with means $E[X]$ and $E[Y]$ respectively, then

$$E[aX + bY] = a\,E[X] + b\,E[Y].$$

If X and Y are independent random variables with variances $Var(X)$ and $Var(Y)$ respectively, then

$$Var(aX + bY) = a^2\,Var(X) + b^2\,Var(Y).$$

Discrete Distributions

$$\text{Mean: } E[Y] = \sum_{i=1}^{k} y_i \Pr(Y = y_i) \qquad \text{Variance: } Var(Y) = \sum_{i=1}^{k} (y_i - E[Y])^2 \Pr(Y = y_i)$$

Normal Distribution

A normal random variable, Y, has mean $E[Y] = \mu$ and variance $Var(Y) = \sigma^2$. A standard normal random variable, Z, has mean 0 and variance 1. To transform a normal random variable Y into a standard normal Z (and vice versa):

$$Z = \tfrac{Y-\mu}{\sigma} \qquad Y = Z\sigma + \mu$$

Binomial Distribution

A binomial random variable X has mean $E[X] = np$ and variance $Var(X) = np(1-p)$.

Distributions of Statistics

- The distribution of $\bar{y}$ has mean μ and standard error $\frac{\sigma}{\sqrt{n}}$
  - Estimate of standard error $\frac{s}{\sqrt{n}}$
- The distribution of $\bar{y}_1 - \bar{y}_2$ has mean $\mu_1 - \mu_2$ and standard error $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
  - Estimate of standard error $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.
- The distribution of $\hat{p} = \frac{x}{n}$ has mean p and standard error $\sqrt{\frac{p(1-p)}{n}}$
  - For hypothesis test with $\mathrm{H}_0 : p = p_0$ the standard error is $\sqrt{\frac{p_0(1-p_0)}{n}}$
- The distribution of $\hat{p}_1 - \hat{p}_2$ has mean $p_1 - p_2$ and standard error $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - For hypothesis test with $\mathrm{H}_0 : p_1 - p_2 = 0$ the estimated standard error is $\sqrt{\frac{\hat{p}^*(1-\hat{p}^*)}{n_1} + \frac{\hat{p}^*(1-\hat{p}^*)}{n_2}}$
  - $\hat{p}^* = \frac{x_1+x_2}{n_1+n_2} = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$.

Contingency Tables

- Test statistic: $X^2 = \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}}$
- df $= (R-1)(C-1)$, where R and C are the number of rows and columns respectively
- The expected count: expected $= \frac{\text{row total} \times \text{column total}}{\text{table total}}$

Regression

The simple linear regression model is: $y = \beta_0 + \beta_1 x + \varepsilon$, where ε is normally distributed with mean 0 and variance σ_ε^2. The fitted model is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$