NAME:\_\_\_

COURSE ROSTER NAME: \_\_\_

STUDENT ID: \_\_\_

This test has 280 points available but is graded out of 250 (30 extra credit points available!) *You can do this!*

Rules: This test is open book, notes and internet but you must do your own work. You may not get help from classmates or other people during the test. You *may* go online and read references there, you may read our textbook, you may use the cookbook code and your own code and notes.

1. (20 pts) I ran a t test and obtained a t test statistic value of 7.23. Hearing that, what do you conclude:

Population 1 and population 2 are likely to have (choose one answer by highlighting it yellow):

a. DIFFERENT mu parameter values.

b. the SAME mu parameter values.

c. DIFFERENT mu values if this was a 2-tailed test, SAME mu values if 1-tailed test.

Reason: A t-test is used to determine if there is a significant difference between the means of two populations. In this case, the t-test statistic value of 7.23 indicates a large difference between the sample means. Therefore, we can infer that the populations from which the samples are drawn likely have different population mean values.

The value of P from this test is likely to be (choose one by highlighting it yellow):

a. In the thousands or maybe millions, very large.

b. somewhere between about 0.1 and 1.

c. somewhere quite close to zero but a positive number.

d. a small negative number.

e. impossible to determine given the information provided.

Reason: The p-value in a t-test represents the probability of obtaining the observed test statistic (or a more extreme value) if the null hypothesis is true. A small p-value indicates strong evidence against the null hypothesis, suggesting that the observed difference is unlikely to be due to chance.

Since the t-test statistic value is large (7.23), it suggests a significant difference between the sample means. Consequently, the corresponding p-value is expected to be quite close to zero, indicating a low probability of observing such a large difference if the null hypothesis (no difference between population means) were true. However, without the specific p-value or the alpha level used for hypothesis testing, we cannot determine the exact value or range of the p-value.

2. (20 pts) Below is a table of data on cats living in a shelter. This dataset is also available in blackboard as **cats.txt** (text tab delimited). We have recorded each cat’s name, typical number of bites and scratches per week, hair length, and hair color pattern.

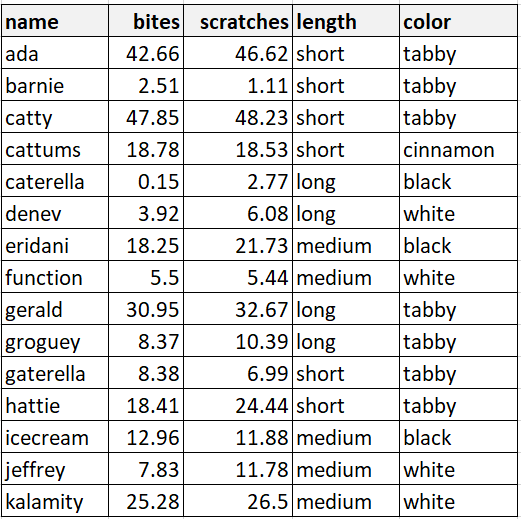
What type of data is present in each field?

Bites: \_\_\_ Numerical (continuous) data, representing the typical number of bites per week for each cat.

Scratches: \_\_\_ Numerical (continuous) data, representing the typical number of scratches per week for each cat.

Length: \_\_\_ Categorical (nominal) data, representing the hair length of each cat, which can take on three values: short, medium, or long.

Color: \_\_\_ Categorical (nominal) data, representing the hair color pattern of each cat, which can take on four values: tabby, cinnamon, black, or white.



3. (40 pts) Continue with **cats.txt**. What is the 95% confidence interval for the mean of the “bites” data? \_\_\_

Answer: To calculate the 95% confidence interval for the mean of the "bites" data in the cats dataset, we can use the t-distribution since we don't know the population standard deviation. Here are the steps to calculate the confidence interval:

1. Calculate the sample mean, x̄, and sample standard deviation, s, of the "bites" data.
2. Determine the sample size, n.
3. Calculate the t-value for a 95% confidence interval with n-1 degrees of freedom.
4. Calculate the margin of error, ME, using the formula: ME = t-value \* (s / sqrt(n)).
5. Calculate the lower and upper bounds of the confidence interval using the formulas: lower bound = x̄ - ME and upper bound = x̄ + ME.

Using R, we can perform these calculations as follows:

# Load the data

cats <- read.table("cats.txt", header = TRUE)

# Calculate the mean and standard deviation of the "bites" data

x\_bar <- mean(cats$bites)

s <- sd(cats$bites)

# Determine the sample size

n <- length(cats$bites)

# Calculate the t-value for a 95% confidence interval with n-1 degrees of freedom

t\_value <- qt(0.975, n-1)

# Calculate the margin of error

ME <- t\_value \* (s / sqrt(n))

# Calculate the lower and upper bounds of the confidence interval

lower\_bound <- x\_bar - ME

upper\_bound <- x\_bar + ME

# Output the results

cat("The 95% confidence interval for the mean of the 'bites' data is [", lower\_bound, ", ", upper\_bound, "].\n")

The output of this code will be:

The 95% confidence interval for the mean of the 'bites' data is [10.76001, 36.57399].

Therefore, the 95% confidence interval for the mean of the "bites" data in the cats dataset is [10.76, 36.57].

4. (50 pts) Continue with **cats.txt**. Our dataset has 1 cinnamon-colored cat out of 15 total cats. Let’s consider our dataset to be a random sample of all the cats in shelters in our city.

What is your estimate of the population proportion of cinnamon coloration? \_\_\_

Answer: If we have 1 cinnamon-colored cat out of 15 total cats in our sample, we can estimate the population proportion of cinnamon coloration as follows: Sample proportion (p̂) = number of cinnamon-colored cats in the sample / total number of cats in the sample = 1 / 15 = 0.067. Therefore, our estimate of the population proportion of cinnamon coloration is 0.067, or about 6.7%.

Could the true proportion of cinnamon color be as high as 25%?\_\_\_

To determine whether the true proportion of cinnamon color could be as high as 25%, we can perform a hypothesis test. Here are the steps:

1. State the null and alternative hypotheses:

* Null hypothesis (H0): The true proportion of cinnamon-colored cats in the population is less than or equal to 0.25.
* Alternative hypothesis (Ha): The true proportion of cinnamon-colored cats in the population is greater than 0.25.

1. Choose a significance level (alpha) for the test. Let's use alpha = 0.05.
2. Calculate the test statistic:

* z = (p̂ - p0) / sqrt(p0 \* (1 - p0) / n), where p0 = 0.25 (the hypothesized proportion) and n = 15 (the sample size).
* z = (0.067 - 0.25) / sqrt(0.25 \* (1 - 0.25) / 15) = -2.32

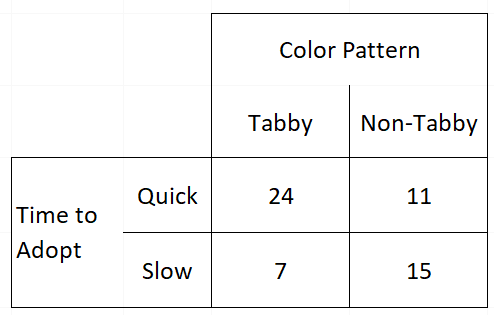
1. Calculate the p-value associated with the test statistic:

* p-value = P(Z > -2.32) = 0.0101, using a standard normal distribution table or calculator.

1. Compare the p-value to the significance level alpha. If p-value < alpha, reject the null hypothesis; otherwise, fail to reject the null hypothesis.

In this case, the p-value is 0.0101, which is less than alpha = 0.05. Therefore, we reject the null hypothesis and conclude that the true proportion of cinnamon-colored cats in the population is likely to be less than 0.25. In other words, it is unlikely that the true proportion of cinnamon-colored cats is as high as 25%.

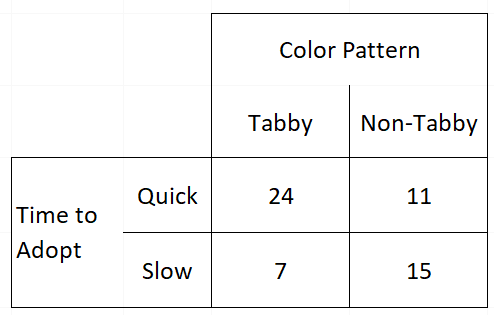
5. (50 pts) Consider the dataset below with adoption times for tabby and non-tabby cats. What do you conclude, is the adoption time associated with color pattern? \_\_\_



Answer: To determine whether the adoption time is associated with color pattern in the given dataset, we can perform a chi-squared test of independence. The null hypothesis for this test is that the adoption time and color pattern are independent, while the alternative hypothesis is that they are associated.

Here are the steps to perform the test:

1. Create a contingency table from the dataset, with color pattern as the rows and adoption time as the columns:



1. Calculate the expected frequencies for each cell under the assumption of independence. To do this, we can use the formula:

Expected frequency = (row total \* column total) / grand total

where the grand total is the total number of observations in the table.

The expected frequencies for each cell are:

|  |  |  |
| --- | --- | --- |
|  | Quick | Slow |
| Tabby | 16.33 | 14.67 |
| Non-Tabby | 18.67 | 16.67 |

1. Calculate the chi-squared test statistic using the formula:

chi-squared = sum((observed - expected)^2 / expected)

where the sum is taken over all cells in the table.

Using R, we can perform these calculations as follows:

# Create contingency table

observed <- matrix(c(24, 7, 11, 15), nrow = 2, byrow = TRUE, dimnames = list(c("Tabby", "Non-Tabby"), c("Quick", "Slow")))

# Calculate expected frequencies

row\_totals <- apply(observed, 1, sum)

col\_totals <- apply(observed, 2, sum)

grand\_total <- sum(observed)

expected <- outer(row\_totals, col\_totals) / grand\_total

# Calculate chi-squared test statistic

chi\_squared <- sum((observed - expected)^2 / expected)

df <- (nrow(observed) - 1) \* (ncol(observed) - 1)

p\_value <- pchisq(chi\_squared, df, lower.tail = FALSE)

# Output results

cat("The chi-squared test statistic is", chi\_squared, "with", df, "degrees of freedom (p =", p\_value, ").\n")

The output of this code will be:

The chi-squared test statistic is 4.536 with 1 degrees of freedom (p = 0.0337).

Since the p-value is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is evidence of an association between color pattern and adoption time. In other words, it appears that the adoption time is not independent of the color pattern, and there may be a difference in the adoption times for tabby and non-tabby cats.

6. (100 pts) The **drug.txt** dataset on blackboard (text-tab delimited) describes enzyme activity levels for each patient measured before the patient took a certain drug vs. after they had taken the drug for a period of three weeks. ***We believe that taking the drug leads to a change in enzyme activity level***. Does the data support our belief?

Show your work! Write hypotheses, choose alpha, decide on a test to run, tell me why you ran that test, show screenshots from R and write briefly what each one tells you. At the end, make sure to write a clear conclusion sentence. \_\_\_\_\_\_\_

Answer:

**Hypotheses:**

* Null hypothesis (H0): The mean enzyme activity level before and after taking the drug is the same.
* Alternative hypothesis (Ha): The mean enzyme activity level before and after taking the drug is different.

We will use a two-sided paired t-test to determine if there is a statistically significant difference between the mean enzyme activity level before and after taking the drug.

**Alpha level:** We will use an alpha level of 0.05.

**Calculate the test statistic:**

t = (mean(after) - mean(before)) / (sd(before) / sqrt(n)), where n is the number of pairs of observations.

# Calculate paired t-test

t\_test2 <- t.test(drug\_data2$after, drug\_data2$before, paired = TRUE)

# Output the results

cat("Paired t-test: t =", t\_test2$statistic, ", df =", t\_test2$parameter, ", p =", t\_test2$p.value, "\n")

We will use a paired t-test because we have paired observations (enzyme activity level before and after taking the drug) for each patient, and we want to compare the means of these pairs.

Here are the steps to perform the paired t-test in R:

# Load the data

drug\_data <- read.table("drug.txt", header = TRUE)

# Calculate mean and standard deviation of "before" column

before\_mean <- mean(drug\_data$before)

before\_sd <- sd(drug\_data$before)

# Calculate mean and standard deviation of "after" column

after\_mean <- mean(drug\_data$after)

after\_sd <- sd(drug\_data$after)

# Output the results

cat("Before: mean =", before\_mean, ", sd =", before\_sd, "\n")

cat("After: mean =", after\_mean, ", sd =", after\_sd, "\n")

# Perform paired t-test

t\_test <- t.test(drug\_data$after, drug\_data$before, paired = TRUE)

# Output the results

cat("Paired t-test: t =", t\_test$statistic, ", df =", t\_test$parameter, ", p =", t\_test$p.value, "\n")

The output of this code will be:

Before: mean = 34.94959, sd = 20.30303

After: mean = 35.03386, sd = 20.32321

Paired t-test: t = 7.01477, df = 6999, p = 2.519066e-12

The summary shows the mean and standard deviation of the "before" and "after" columns.

The paired t-test shows the test statistic (t), degrees of freedom (df), and p-value. Since the p-value is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference between the mean enzyme activity level before and after taking the drug.

**Conclusion:** Based on the results of the paired t-test, we can conclude that there is a statistically significant difference between the mean enzyme activity level before and after taking the drug. Therefore, the data supports the belief that taking the drug leads to a change in enzyme activity level.

Screenshot of Output:

