Practice Midterm 2 - DSC 152, Fall 2026

Instructions:

- This is a practice midterm. It provides an example of the midterm exam, but this does not constitute any guarantee for the midterm exam.
- This practice midterm consists of 8 questions. You have **50 minutes** to complete it.
- Please write clearly in the provided answer boxes; work elsewhere may not be graded. Completely fill in bubbles and square boxes; unclear selections may lose points.
  ◯ A bubble means select only one choice.
  ☐ A square box means select all that apply.
- Show all work and R code where requested. Partial credit may be awarded.
- You may use one double-sided handwritten sheet of notes. No calculators, computers, phones, or other devices.
- Assume all necessary `library()` calls in R have already been run.

Additional Practice Midterm notes:

- This practice midterm is intended to reflect the format of the previous practice midterm while sampling important concepts from Lectures 6 through 11.
- It is not intended to cover every concept or course activity. Use it as practice, then check that you can explain why each answer is correct.

A campus library wants to know which self-checkout screen helps people check out five books faster. It is considering three screen layouts: Standard, Guided, and Compact. Smaller checkout times are better. All studies and data on this exam are fictional.

The `kiosk` data frame comes from an **observational study**. Each of 72 volunteers chose a layout and completed one checkout session. There were 24 sessions for each layout and no missing values. The columns are:

- `session_id` (integer): unique identifier for each session
- `layout` (factor): `"Standard"`, `"Guided"`, or `"Compact"`, with Standard as the reference level
- `prior_uses` (integer): number of previous visits to a self-checkout kiosk
- `checkout_sec` (double): checkout time in seconds
- `order` (integer): order in which the sessions were recorded

The first five rows of `kiosk` are shown below. Not all columns will be needed on this midterm.

	session_id	layout	prior_uses	checkout_sec	order
1	51001	Standard	4	55	1
2	51002	Guided	6	53	2
3	51003	Compact	5	48	3
4	51004	Standard	3	56	4
5	51005	Compact	4	42	5

The library wants to assess whether changing the checkout layout affects checkout time and whether the effect depends on a screen prompt. Use $\alpha = 0.05$ throughout. Assume the conditions for each test hold unless asked to assess them.

TSN: _____

Question 1

In a randomized study, 24 volunteers are assigned to use the Standard or Guided layout, 12 per group. The `trial` data frame contains `layout` and `checkout_sec`, with one row per participant. Let μ_G and μ_S denote the population mean checkout times under the two layouts.

a) Write the null and alternative hypotheses for a two-sided test of whether the mean times differ.

H_0 : H_A :

b) Write one line of R code to run the default two-sample t-test on `trial` and output the two-sided p-value.

c) Suppose the p-value is 0.080. A teammate says: *This means there is an 8% chance that the two population means are equal.* Is this interpretation correct? Why or why not?

d) Name two conditions that support the nominal Type I Error rate of the default two-sample t-test, especially with 12 observations per group. Briefly explain why each matters. Must the population variances be equal?

TSN: _____

Question 2

Suppose we fit a simple linear regression using only the 24 Standard-layout sessions in `kiosk`. Write x for `prior_uses`. The fitted line is $\hat{y} = 60 - 2x$, and the standard error of the slope estimate is 1.

a) A participant with four prior uses takes 55 seconds. Find the fitted time and the residual.

b) For the test of $H_0 : \beta_1 = 0$, compute the t-statistic and give its degrees of freedom.

c) A colleague runs the following R code to obtain the p-value for the slope:

```
fit <- lm(checkout_sec ~ prior_uses, data = kiosk,
          subset = layout == "Standard")
summary(fit)$coefficients["prior_uses", 4]
## [1] 0.0579957
```

State the decision at $\alpha = 0.05$. Briefly explain what the p-value means in this study.

TSN: _____

Question 3

In `trial`, the observed Guided-minus-Standard mean difference is -6 seconds. We use a two-sided randomization test of the null that the layout changes no participant's checkout time.

a) Of 1,000 shuffled differences, 38 are at most -6 and 32 are at least 6. What is the estimated p-value and decision at $\alpha = 0.05$? Select one.

◯ A. 0.038; reject H_0.

◯ B. 0.070; fail to reject H_0.

◯ C. 0.140; fail to reject H_0.

◯ D. 0.032; reject H_0.

b) Fill in the blanks in the R code below. The function `diff_means(y, g)` computes the Guided mean minus the Standard mean using labels `g`.

```
obs <- -6
null_stats <- numeric(1000)
for (i in 1:1000) {
  g <- sample(trial$layout, replace = __(a)__)
  null_stats[i] <- diff_means(trial$checkout_sec, __(b)__)
}
pval <- mean(__(c)__)
```

(a): _____ (b): _____

(c): _____

c) In one sentence, explain what the resulting value of `pval` estimates.

TSN: _____

Question 4

Use the Standard and Guided sessions in `kiosk`. The unadjusted Guided mean is 3 seconds lower than the Standard mean. After adjusting for prior uses, the fitted model is

$$\hat{y} = 60 + 3G - 2x,$$

where $G = 1$ for Guided and $G = 0$ for Standard, and x is the number of prior uses.

a) Interpret the coefficient of G in context.

b) Guided users have three more prior uses on average. Explain how this can account for the reversal in the comparison.

c) We refit the model using $x_c = x - 4$ in place of x. Give the new intercept and interpret it in context.

d) Does the adjusted comparison establish that the Guided layout causes longer checkout times? Explain.

TSN: _____

Question 5

Use all 72 sessions in `kiosk`. The full model includes layout and prior uses. The reduced model keeps prior uses and omits layout. Both models include an intercept and use the same observations. The residual sums of squares are $RSS_{\text{reduced}} = 3112$ and $RSS_{\text{full}} = 2720$.

Degrees of freedom (df_1, df_2)	$(1, 68)$	$(2, 68)$	$(3, 68)$	$(2, 69)$
5% upper-tail cutoff	3.98	3.13	2.74	3.13

a) Compute the partial F-statistic and give both degrees of freedom.

b) State the null hypothesis in terms of the adjusted layout means. Use the table above to give the decision at $\alpha = 0.05$ and the conclusion in context.

TSN: _____

Question 6

Researchers want to estimate power for the two-sided slope test of $H_0 : \beta_1 = 0$. Hold the vector `x` of prior-use counts fixed across repetitions. Assume the true mean checkout time is $60 + \beta_1 x$ and the errors are independent and normal with standard deviation `sd`.

a) Complete the function below to return a simulated estimate of power. The argument `beta1` is the true slope used to generate the data.

```
sim_power <- function(x, beta1, sd, alpha = 0.05,
                      reps = 10000) {
  count <- 0
  for (i in 1:reps) {
    y <- 60 + __(a)__ * x + rnorm(n = __(b)__,
                                  mean = 0, sd = __(c)__)
    fit <- lm(y ~ x)
    pval <- summary(fit)$coefficients[2, 4]
    if (__(d)__) count <- count + 1
  }
  return(__(e)__)
}
```

(a): _____ (b): _____

(c): _____ (d): _____

(e): _____

b) If `sim_power(rep(0:7, each = 3), beta1 = -1, sd = 6)` returns approximately 0.43, what does this number represent in the kiosk study?

TSN: _____

Question 7

For a regression on a separate sample of kiosk sessions, the residuals stay centered near zero, but their vertical spread increases as fitted time increases. A Shapiro-Wilk test on the residuals gives a p-value of 0.42.

a) At $\alpha = 0.05$, what is the decision for the Shapiro-Wilk test?

◯ Reject H_0. ◯ Fail to reject H_0.

b) A teammate argues that this p-value makes the usual regression coefficient p-values reliable. In 2–3 sentences, identify the LINE condition called into question by the residual plot and explain why the Shapiro-Wilk result does not resolve the concern.

Question 8

In a separate study, 48 volunteers are randomly assigned to four combinations of layout and an on-screen prompt, 12 per combination. In `fact`, `guided` (G) is 1 for Guided and 0 for Standard; `prompt` (P) is 1 if the prompt is on and 0 if it is off. Sample mean checkout times in seconds are:

	Prompt off ($P = 0$)	Prompt on ($P = 1$)
Standard ($G = 0$)	52	49
Guided ($G = 1$)	50	41

The model `full <- lm(checkout_sec ~ guided * prompt, data = fact)` gives a p-value of 0.391 for the `guided` coefficient.

a) What is the estimated coefficient of the interaction term GP? Select one.

◯ A. -2 seconds.

◯ B. -6 seconds.

◯ C. -8 seconds.

◯ D. $+6$ seconds.

b) A colleague says the p-value of 0.391 shows that Guided has no effect at either prompt setting. Which statement correctly describes the null tested by this p-value? Select one.

◯ A. It tests the Guided effect when the prompt is on.

◯ B. It tests the Guided effect when the prompt is off, leaving the prompt and interaction coefficients unrestricted.

◯ C. It jointly tests the Guided effect at both prompt settings.

◯ D. It tests the interaction while assuming the Guided and prompt coefficients are zero.

c) State the joint null that Guided has no effect at either prompt setting. Write R code to fit the reduced model and compare it with `full`. This partial F-test gives $p = 0.0037$. State the conclusion in context.