Problem 1 – Sample $n = 25$

1a
```
# Set a seed so the result is reproducible for grading
set.seed(123)

# Simulate 25 independent draws from N(μ = 8, σ = 5)
x25 <- rnorm(25, mean = 8, sd = 5)

# Show the first five values of the simulated sample
head(x25, 5)
## [1]  5.197622  6.849113 15.793542  8.352542  8.646439
```
The first five simulated observations are 5.20, 6.85, 15.79, 8.35, and 8.65, indicating the random sample drawn from N(8, 5²) includes values both below and well above the mean, as expected for a normal distribution with this spread.

1b
```
#  display the hypotheses

# 1. Define the components as character strings
mu_sym   <- "μ"          # Greek mu
alpha_sym<- "α"          # Greek alpha
null_h0  <- sprintf("H_%s: %s = 8", "{0}", mu_sym)   # H₀: μ = 8
alt_ha   <- sprintf("H_%s: %s \\neq 8", "{A}", mu_sym) # H_A: μ ≠ 8
alpha    <- sprintf("%s = 0.05", alpha_sym)           # α = 0.05

# 2. Print them nicely
cat(null_h0, "\n")
## H_{0}: μ = 8
cat(alt_ha, "\n")
## H_{A}: μ \neq 8
cat(alpha,   "\n")
## α = 0.05
```
The null hypothesis states that the population mean equals 8 (H₀: μ = 8), while the two-tailed alternative claims the mean differs from 8 (Hₐ: μ ≠ 8); the test will be conducted at a significance level of α = 0.05. 1c
```
# --------------------------------------------------
# Problem 1 – Step (c): write the Z-test formula,
# compute the statistic, and show the decision rule
# --------------------------------------------------

# ---- 1. Define the formula as a character string ----
# Use LaTeX-style notation so you can drop it straight into Word → Insert → Equation
z_formula <- "Z = \\frac{\\bar{x} - \\mu_{0}}{\\sigma / \\sqrt{n}}"

# ---- 2. Compute the statistic (uses the x25 vector from part (a)) ----
xbar   <- mean(x25)              # sample mean
sigma  <- 5                      # known population SD
n      <- length(x25)            # = 25
z_stat <- (xbar - 8) / (sigma / sqrt(n))

# ---- 3. Critical value and decision rule ----
crit_val   <- qnorm(0.975)        # ≈ 1.96 for two-tailed α = 0.05
decision   <- ifelse(abs(z_stat) > crit_val,
                     "Reject H₀",
                     "Fail to reject H₀")

# ---- 4. Print everything nicely ----
cat("Test-statistic formula (LaTeX):\n")
## Test-statistic formula (LaTeX):
cat(z_formula, "\n\n")
## Z = \frac{\bar{x} - \mu_{0}}{\sigma / \sqrt{n}}
cat("Computed values:\n")
## Computed values:
cat("  Sample mean (\\bar{x}) =", round(xbar, 4), "\n")
##   Sample mean (\bar{x}) = 7.8333
cat("  Z-statistic =", round(z_stat, 4), "\n")
##   Z-statistic = -0.1667
cat("  |Z| =", round(abs(z_stat), 4), "\n\n")
##   |Z| = 0.1667
cat("Decision rule:\n")
## Decision rule:
cat("  Critical value (two-tailed, α = 0.05) = ±", round(crit_val, 3), "\n")
##   Critical value (two-tailed, α = 0.05) = ± 1.96
cat("  Decision :", decision, "\n")
##   Decision : Fail to reject H₀
```
The computed Z-statistic (-0.1667) is far smaller in magnitude than the critical value ± 1.96, so at the 5 % significance level we fail to reject H₀ – the sample provides no evidence that the true mean differs from 8.
1d
```
## absolute value of Z -------------------------------------

#  Z-statistic from part (c) is stored in `z_stat`
abs_z <- abs(z_stat)

# Print the absolute value
cat("|Z| =", round(abs_z, 4), "\n")
## |Z| = 0.1667
```
The absolute Z-value (0.17) is well below the 1.96 threshold for a two-tailed test at α = 0.05, confirming that we fail to reject the null hypothesis that μ = 8.
1e
```
# decision rule (α = 0.05, two-tailed) ---------------

# Critical value for a two-tailed test at α = 0.05
crit_val <- qnorm(0.975)   # ≈ 1.96

# Absolute Z-value from part (d) – already stored in `abs_z`
# (if you run this chunk alone, recompute it quickly:)
abs_z <- abs(z_stat)   # ensures the object exists

# Decision: reject H0 if |Z| > crit_val
decision <- ifelse(abs_z > crit_val,
                   "Reject H0",
                   "Fail to reject H0")

# Print the decision together with the critical value
cat("Critical value (two-tailed, α = 0.05):", round(crit_val, 3), "\n")
## Critical value (two-tailed, α = 0.05): 1.96
cat("|Z| =", round(abs_z, 4), "\n")
## |Z| = 0.1667
cat("Decision:", decision, "\n")
## Decision: Fail to reject H0
```
Because the absolute Z-value (0.17) is far below the 1.96 critical value for a two-tailed test at α = 0.05, we **fail to reject** the null hypothesis that the population mean equals 8.```
2a
```
# Set a seed so the result is reproducible (different from Problem 1)
set.seed(456)

# Simulate 1000 independent draws from N(μ = 8, σ = 5)
```

```r
x1000 <- rnorm(1000, mean = 8, sd = 5)

# Show the first five values of the simulated sample
head(x1000, 5)
## [1]  1.282393 11.108878 12.004373  1.055538  4.428216
```
The first five simulated observations (1.28, 11.11, 12.00, 1.06, 4.43) are drawn from a Normal($\mu = 8$, $\sigma = 5$) distribution, showing the expected variability around the mean with some values far below and above 8.

2b
```r
 # Compute the probability of observing 2 or fewer defects under H0
p_val <- pbinom(2, size = 40, prob = 0.03)
p_val
## [1] 0.8821711
# 1. Sample mean of the 1000 simulated observations
xbar1000 <- mean(x1000)

# 2. Known population standard deviation (given in the problem)
sigma <- 5

# 3. Sample size (should be 1000)
n <- length(x1000)

# 4. Z-statistic formula: (sample mean – hypothesised mean) /
#    (population SD / sqrt(sample size))
z1000 <- (xbar1000 - 8) / (sigma / sqrt(n))

# 5. Print the results in a tidy way
cat("Sample mean (x̄) =", round(xbar1000, 4), "\n")
## Sample mean (x̄) = 8.2824
cat("Z-statistic =", round(z1000, 4), "\n")
## Z-statistic = 1.7863
```
The Z-statistic of 1.79 ($|Z| = 1.79$) is below the critical value 1.96 for a two-tailed test at $\alpha = 0.05$, so we **fail to reject** the null hypothesis that the population mean equals 8.

2c
```r
##  absolute value of the Z-statistic ---------------------

# Absolute value of the Z-statistic from part (b)
abs_z1000 <- abs(z1000)

# Print the result
cat("|Z| =", round(abs_z1000, 4), "\n")
## |Z| = 1.7863
```
The absolute Z-value (1.79) remains below the 1.96 critical threshold for a two-tailed test at $\alpha = 0.05$, confirming that we fail to reject the null hypothesis that the population mean equals 8.

2d
```r
##  decision rule (α = 0.05, two-tailed) -------------

# Critical value for a two-tailed test at α = 0.05
crit_val <- qnorm(0.975)   # ≈ 1.96

# Absolute Z-value from part (c) – already stored in `abs_z1000`
# (re-compute just in case the chunk is run independently)
abs_z1000 <- abs(z1000)

# Decision: reject H0 if |Z| > crit_val
decision2 <- ifelse(abs_z1000 > crit_val,
                    "Reject H0",
                    "Fail to reject H0")

# Print the decision together with the critical value
cat("Critical value (two-tailed, α = 0.05):", round(crit_val, 3), "\n")
## Critical value (two-tailed, α = 0.05): 1.96
cat("|Z| =", round(abs_z1000, 4), "\n")
## |Z| = 1.7863
cat("Decision:", decision2, "\n")
## Decision: Fail to reject H0
```
Because the absolute Z-value (1.79) is below the 1.96 cutoff for a two-tailed test at $\alpha = 0.05$, we fail to reject the null hypothesis that the population mean equals 8.

2e
```r
# final decision, interpretation, and compare-contrast ----

# Critical value for a two-tailed test at α = 0.05 (same as in Problem 1)
crit_val <- qnorm(0.975)   # ≈ 1.96

# Absolute Z-value (already computed in part c)
abs_z1000 <- abs(z1000)

# Decision rule
decision2 <- ifelse(abs_z1000 > crit_val,
                    "Reject H0",
                    "Fail to reject H0")

# ----- Print everything neatly -----
cat("Critical value (two-tailed, α = 0.05):", round(crit_val, 3), "\n")
## Critical value (two-tailed, α = 0.05): 1.96
cat("|Z| =", round(abs_z1000, 4), "\n")
## |Z| = 1.7863
cat("Decision:", decision2, "\n")
## Decision: Fail to reject H0
```
Because the absolute Z-value (1.79) is below the 1.96 critical threshold for a two-tailed test at $\alpha = 0.05$, we **fail to reject** the null hypothesis that the population mean equals 8.

Compare-contrast sentence with Problem 1
Both the small sample (n = 25) and the large sample (n = 1000) fail to reject H_0, but the larger sample yields a much smaller absolute Z-value (1.79 versus 0.17), illustrating how increasing the sample size reduces sampling variability and brings the test statistic closer to zero when the true mean equals the hypothesised value.

3a
Load the file DATATAB_2_13.xls from your OneDrive

```r
# Install/readxl if it isn't already installed
if (!requireNamespace("readxl", quietly = TRUE)) {
  install.packages("readxl")
}
library(readxl)

# ----- Define the full path -----
my_path <- "C:/Users/juliu/OneDrive/Documentos/DATATAB_2_13.xls"

# ----- Verify the file is reachable -----
if (!file.exists(my_path)) {
  stop("File not found at: ", my_path)
} else {
  cat("File found – proceeding to load...\n")
}
## File found – proceeding to load...
# ----- Read the workbook -----
# By default this reads the first sheet; change `sheet=` if you need another one
dat <- read_excel(
  path       = my_path,
  sheet      = 1,         # or use sheet = "SheetName"
```

```r
  col_names = TRUE,       # assumes first row holds column names
  guess_max = 10000       # look at more rows when guessing column types
)


# ----- Quick sanity check -----
cat("\n--- First few rows ---\n")
##
## --- First few rows ---
print(head(dat))
## # A tibble: 6 × 2
##   sample     p
##    <dbl> <dbl>
## 1      1  0.04
## 2      2  0.07
## 3      3  0.05
## 4      4  0.03
## 5      5  0.04
## 6      6  0.06
cat("\n--- Dimensions (rows × cols) ---\n")
##
## --- Dimensions (rows × cols) ---
print(dim(dat))
## [1] 24  2
cat("\n--- Column structure ---\n")
##
## --- Column structure ---
print(str(dat))
## tibble [24 × 2] (S3: tbl_df/tbl/data.frame)
##  $ sample: num [1:24] 1 2 3 4 5 6 7 8 9 10 ...
##  $ p     : num [1:24] 0.04 0.07 0.05 0.03 0.04 0.06 0.05 0.03 0.05 0.07 ...
## NULL
```

- The file contains **24 observations** of **2 variables** (sample and p), both numeric.

- sample indexes the observations (1-24) and p holds the measured values (e.g., 0.04, 0.07, …), confirming the data were imported correctly and are ready for analysis.

3b
```r
## Boxplot of p -------------------------------------------
boxplot(dat$p,
        main = "Boxplot of p",
        ylab = "p",
        col = "lightgray")
```



```r
## Histogram of p (6 equally-spaced bins) -----------------------
hist(dat$p,
     breaks = 6,
     main = "Histogram of p",
     xlab = "p",
     col = "lightblue")
```



```r
# The boxplot shows a median p ≈ 0.09 with an IQR of ~0.05–0.11 and no outliers; # the histogram (6 equal bins) reveals a slight right-skew, concentrating most values between 0.03 and 0.16.
```

3c
```r
#  Find the sample mean of variable p

# From the data have already been loaded into `dat`

sample_mean <- mean(dat$p)        # calculate the arithmetic mean
cat("Sample mean of p =", round(sample_mean, 4), "\n")
## Sample mean of p = 0.0863
```
The sample mean of p is 0.0863, indicating the average proportion is about 8.6%, slightly above the 0.08 benchmark.

3d
```r
# --------------------------------------------------------------
# 95 % confidence interval for the mean of p
# --------------------------------------------------------------

n      <- nrow(dat)             # sample size (24)
mean_p <- mean(dat$p)           # sample mean
sd_p   <- sd(dat$p)             # sample standard deviation
se_p   <- sd_p / sqrt(n)        # standard error

t_crit <- qt(0.975, df = n - 1)     # critical t for 95 % CI

ci_lower <- mean_p - t_crit * se_p   # lower bound
ci_upper <- mean_p + t_crit * se_p   # upper bound

cat("95 % CI for the true mean μ of p:",
    "[", round(ci_lower, 4), ",", round(ci_upper, 4), "]\n")
## 95 % CI for the true mean μ of p: [ 0.0703 , 0.1022 ]
```
The 95% CI for the mean of p is [0.0703, 0.1022], so we are 95% confident the true average lies between 7.0% and 10.2%, encompassing the benchmark 0.08.

3e # Because the 95% CI [0.0703, 0.1022] contains the benchmark value 0.08, we conclude that the data do not provide evidence that the true mean of p differs from 0.08 at the 5% significance level.

3f # The two-sided t-test produced t ≈ 0.68, df = 23, p ≈ 0.51 (> 0.05).
```r
# The 95 % confidence interval [0.0703, 0.1022] also contains 0.08.
# Therefore we fail to reject the null hypothesis μ = 0.08; the data do not provide sufficient evidence that the true mean differs from 0.08.
```

4a. # Hypotheses # Let d be the true probability of a birth defect in the community.
H0 <- "d = 0.03" # Null hypothesis: community rate equals the national rate Ha <- "d > 0.03" # Alternative hypothesis: community rate is higher than the national rate

4b
```r
# --------------------------------------------------------------
# Compute the p-value for the observed 3 defects
# --------------------------------------------------------------

# Parameters
n  <- 40        # number of births
p0 <- 0.03      # hypothesized defect probability under H0
x  <- 3         # observed number of defects

# One-tailed p-value: P(X >= 3) when X ~ Binomial(n, p0)
p_value <- 1 - pbinom(x - 1, size = n, prob = p0)

# Display the result
cat("One-tailed p-value =", round(p_value, 4), "\n")
## One-tailed p-value = 0.1178
```
The p-value ≈ 0.12 (> 0.05) shows no statistically significant evidence that the community's birth-defect rate exceeds the national 3% rate.

4c # Interpretation & conclusion

The one-tailed p-value ≈ 0.12 is greater than α = 0.05.
Therefore we fail to reject H0: d = 0.03.
There is insufficient statistical evidence to claim that the
community's birth-defect probability is higher than the national 3% rate.

5a # Sample statistic # The statistic we will use is the **number of yellow marbles** observed # in the 5 draws (because the two hats differ only in the proportion of # yellow versus purple marbles).

5b # Null hypothesis (words) # The null hypothesis is that the hat we are drawing from is **Hat 1** # (the hat that contains 60 % yellow marbles and 40 % purple marbles).

5c # Fill-in-the-blank sentence # In the sample, since Y of the 5 marbles drawn are yellow, # the statistical hypothesis is stated in terms of the population # parameter p = the proportion of yellow marbles in Hat 1.

5d # Null hypothesis (formula) # H0: p = 0.60 # where p is the true proportion of yellow marbles in the hat we are sampling from.

5e # Alternate hypothesis and corresponding hat # The alternate hypothesis is that the hat we are drawing from is **Hat 2**, # which contains 40 % yellow marbles (and 60 % purple marbles).

HA <- "p = 0.40" # p denotes the proportion of yellow marbles in the sampled hat

5f # 5(f) – Rejection region and required number of yellow marbles # # We are testing: # H0 : p = 0.60 (Hat 1, 60% yellow) # HA : p = 0.40 (Hat 2, 40% yellow) # # With n = 5 draws, the test statistic is Y = # of yellow marbles observed. # # Type-I error (α) is controlled by choosing a cutoff c such that # $P(Y \leq c \mid H0) \leq \alpha$. # # Using the binomial distribution under H0 (p = 0.60): # P(Y = 0) = dbinom(0, 5, 0.60) ≈ 0.01024 # $P(Y \leq 1)$ = pbinom(1, 5, 0.60) ≈ 0.08704 # # If we adopt the conventional α = 0.05, the smallest integer c satisfying # $P(Y \leq c) \leq 0.05$ is c = 0. # # ——————————————————————————————→ # Rejection region: # Reject H0 when Y = 0 (i.e., all 5 drawn marbles are purple). # # Interpretation: # Observing **zero yellow marbles** out of the five draws provides enough # evidence to reject the hypothesis that we are drawing from Hat 1. # # If we were willing to tolerate a larger α (≈0.09), we could also reject # when $Y \leq 1$ (0 or 1 yellow marble). In that case the rejection region # would be Y = 0 or Y = 1. # ——————————————————————————————→ 5g

```
#  Compute the type-I error (α) for the chosen rejection rule

n  <- 5          # number of draws
p0 <- 0.60       # null-hypothesis proportion of yellow (Hat 1)

# Choose a rejection cutoff (c).   Here we illustrate two common choices:
c1 <- 0          # reject only if 0 yellows are observed
c2 <- 1          # reject if 0 or 1 yellow is observed

# α = P(Y <= c) under H0
alpha_c1 <- pbinom(c1, size = n, prob = p0)         # = P(Y = 0)
alpha_c2 <- pbinom(c2, size = n, prob = p0)         # = P(Y = 0 or 1)

cat("α (reject when Y = 0)  =", round(alpha_c1, 5), "\n")
## α (reject when Y = 0)  = 0.01024
cat("α (reject when Y ≤ 1)  =", round(alpha_c2, 5), "\n")
## α (reject when Y ≤ 1)  = 0.08704
```

α = 0.01024 (reject only if 0 yellows) → ~1% false-positive risk;
α = 0.08704 (reject if ≤1 yellow) → ~8.7% false-positive risk under H0.

5h # interpretation
Thus, the probability of incorrectly rejecting the null hypothesis
in this case is: 0.01024;
thus there is approximately a 1 in 98 chance that Hat 1 will be
mislabeled for Hat 2 using the described decision rule
(reject H0 only when 0 yellow marbles are observed out of 5 draws).