Statistics 7.2.2 Exercise 3

Normal Distributions and Hypothesis Testing

Lucas Liona

Table of contents

Problem Background

This exercise examines data from Dai et al. (1988) on testosterone levels in smokers and nonsmokers. We are given that:

- Nonsmoking men (ages 35-57) have testosterone levels with mean 620.6 ng/dL and standard deviation 241.5 ng/dL
- A sample of 28 heavy smokers had a mean testosterone level of 795.1 ng/dL
- We want to determine if the distributions are the same for smokers and nonsmokers

Part (a): Sample Mean Probability

We need to find the probability that the sample mean for a sample of size 28 taken from the nonsmoker distribution would be 795.1 ng/dL or greater.

First, we need to find the distribution of the sample mean. By the Central Limit Theorem:

$\bar{X} \sim \text{Normal}\,(\mu_{\bar{X}}, \sigma_{\bar{X}})$

Where: - $\mu_{\bar{X}} = \mu = 620.6$ ng/dL - $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{241.5}{\sqrt{28}} \approx 45.6392$ ng/dL

Now we can calculate $P(\bar{X} \geq 795.1)$:

```r
# Given values
mu <- 620.6        # Mean testosterone level for nonsmokers
sigma <- 241.5     # Standard deviation for nonsmokers
n <- 28            # Sample size
sample_mean <- 795.1  # Observed sample mean for smokers

# Calculate standard error of the mean
se <- sigma / sqrt(n)

# Calculate z-score
z <- (sample_mean - mu) / se

# Calculate probability
prob <- 1 - pnorm(z)

cat("Mean of sampling distribution:", mu, "ng/dL\n")
```

```
Mean of sampling distribution: 620.6 ng/dL
```

```r
cat("Standard error of the mean:", se, "ng/dL\n")
```

```
Standard error of the mean: 45.63921 ng/dL
```

```r
cat("Z-score:", z, "\n")
```

```
Z-score: 3.823467
```

```r
cat("Probability of observing 795.1 ng/dL or higher:", prob, "\n")
```

```
Probability of observing 795.1 ng/dL or higher: 6.579419e-05
```

```r
cat("Probability in scientific notation:", format(prob, scientific = TRUE), "\n")
```

```
Probability in scientific notation: 6.579419e-05
```

The probability of observing a sample mean of 795.1 ng/dL or higher from the nonsmoker distribution is approximately 0.0001, which is extremely small (about 1 in 10,000).

This suggests that the distributions of testosterone levels for smokers and nonsmokers are likely different. It would be extremely unlikely to observe such a high sample mean if smokers and nonsmokers had the same testosterone distribution. The evidence strongly indicates that smokers tend to have higher testosterone levels than nonsmokers.

Part (b): Individual Probability

Now we need to find the proportion of individual nonsmokers who have testosterone levels above 795.1 ng/dL.

Assuming testosterone levels in nonsmokers follow a Normal distribution with mean 620.6 ng/dL and standard deviation 241.5 ng/dL, we can calculate:

```
# Calculate z-score for individual measurement
z_individual <- (795.1 - mu) / sigma

# Calculate probability for individual measurement
prob_individual <- 1 - pnorm(z_individual)

cat("Z-score for individual:", z_individual, "\n")
```

```
Z-score for individual: 0.7225673
```

```
cat("Proportion of nonsmokers with testosterone above 795.1 ng/dL:", prob_individual, "\n")
```

```
Proportion of nonsmokers with testosterone above 795.1 ng/dL: 0.2349729
```

```
cat("Percentage of nonsmokers with testosterone above 795.1 ng/dL:", prob_individual * 100,
```

```
Percentage of nonsmokers with testosterone above 795.1 ng/dL: 23.49729 %
```

The proportion of nonsmokers with testosterone levels above 795.1 ng/dL is approximately 0.2350 or 23.50%.

Part (c): Comparing Probabilities

There is no conflict between the answers to parts (a) and (b). The probabilities are calculating different things:

1. In part (a), we calculated the probability of a sample mean being 795.1 ng/dL or higher. This is very small (0.0001) because sample means vary much less than individual measurements. The Central Limit Theorem tells us that the distribution of sample means has the same mean as the original distribution but a much smaller standard deviation (reduced by a factor of $\sqrt{n}$).

2. In part (b), we calculated the probability of an individual nonsmoker having a testosterone level of 795.1 ng/dL or higher. This is much larger (0.2350 or 23.5%) because individual measurements have much more variability.

The key difference is that part (a) deals with the behavior of an average (sample mean) across 28 individuals, while part (b) deals with a single individual measurement. Sample means are much more precise estimators of the population mean than individual measurements.

This is why we can be very confident that the smoker population has a higher mean testosterone level, even though a substantial proportion (23.5%) of individual nonsmokers have testosterone levels above the smoker sample mean of 795.1 ng/dL.