# STA 104 - Assignment 1

Xiaowei Zeng

October 8, 2023

## Exercise 3 (P21)

The data in the table are the yearly rainfall totals in Scranton, Pa., for the years 1951-1984.

Rainfall Totals (inches) for Scranton, Pa., 1951–1984

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 21.3 | 28.8 | 17.6 | 23.0 | 27.2 | 28.5 | 32.8 | 28.2 | 25.9 | 22.5 | 27.2 | 33.1 | 28.7 | 24.8 | 24.3 | 27.1 | 30.6 |
| 26.8 | 18.9 | 36.3 | 28.0 | 17.9 | 25.0 | 27.5 | 27.7 | 32.1 | 28.0 | 30.9 | 20.0 | 20.2 | 33.5 | 26.4 | 30.9 | 33.2 |

a. Make a 95% confidence interval for the median.

b. Make 90% confidence intervals for the 20th and 80th percentiles.

c. The confidence interval procedure assumes that the observations are independent and identically distributed. Do you think this is a reasonable assumption for the rainfall data? If not, what could cause this assumption to be invalid?

***Solution:***

a. Denote the $i$th order statistic as $X_{(i)}, i = 1, 2, ..., N, N = 34$. We wish to find an interval $(X_{(a)}, X_{(b)})$ such that

$$P(X_{(a)} < \theta_{.5} < X_{(b)}) = 95\%.$$

$X_{(a)} < \theta_{.5} < X_{(b)}$ means at least $a$ of the observations must fall less than $\theta_{.5}$ and at most $b-1$ of the observations must fall less than or equal to $\theta_{.5}$.

Therefore, the exact $a$ and $b$ should be computed by

$$\sum_{k=a}^{b-1} \binom{34}{k} (0.5)^{34} = 95\%.$$

However, the binomial distribution is descrete and it may be hard to find the exact limits. Since the sample size is larger than 30, we can obtain $a$ and $b$ by using the normal approximation to the binomial distribution:

$$\frac{a-0.5\times 34}{\sqrt{0.25\times 34}}=-z_{0.975},\quad \frac{b-1-0.5\times 34}{\sqrt{0.25\times 34}}=z_{0.975}.$$

$$\therefore\ a\approx 11.29,\quad b\approx 23.71.$$

Rounding them to the nearest integer, we have $X_{(11)}=25.0$ and $X_{(24)}=28.7$ as the lower and upper 95% confidence limits.

b. Similar as question (a), we wish to find intervals $(X_{(a_1)}, X_{(b_1)})$ and $(X_{(a_2)}, X_{(b_2)})$ such that

$$P(X_{(a_1)}<\theta_{.2}<X_{(b_1)})=90\%,\quad P(X_{(a_2)}<\theta_{.8}<X_{(b_2)})=90\%.$$

Use the normal approximation to the binomial distribution, we have

$$\frac{a_1-0.2\times 34}{\sqrt{0.2\times 0.8\times 34}}=-z_{0.975},\quad \frac{b_1-1-0.2\times 34}{\sqrt{0.2\times 0.8\times 34}}=z_{0.975},$$
$$\frac{a_2-0.8\times 34}{\sqrt{0.8\times 0.2\times 34}}=-z_{0.975},\quad \frac{b_2-1-0.8\times 34}{\sqrt{0.8\times 0.2\times 34}}=z_{0.975}.$$

$$\therefore\ a_1\approx 2.96,\ b_1\approx 11.64, a_2\approx 23.36,\ b_2\approx 32.04.$$

Rounding them to the nearest integer, $X_{(3)}=18.9$, $X_{(12)}=25.9$ are the lower and upper 90% confidence limits for the 20th percentiles, and $X_{(23)}=28.5$, $X_{(32)}=33.2$ are the lower and upper 90% confidence limits for the 80th percentiles.

c. I think the assumption of identically distributed rainfalls is reasonable, but the observations may not be independent because there may exist time dependence between adjacent years.

## Exercise 4 (P22)

Suppose we test the hypotheses $H_0:\theta_{.5}=75$ versus $H_a:\theta_{.5}>75$ and, regardless of the data, we reject $H_0$.

a. What is the probability of a Type I error?

b. What is the power of the test for values of $\theta_{.5}>75$?

***Solution:***

a. $P(\text{Type I Error})=P(\text{reject } H_0|H_0)$. Since we reject $H_0$ regardless of the data, $P(\text{Type I Error})$ should be 1.

b. power $=P(\text{reject } H_0|H_a)$. Since we reject all $H_0$, the power should also be 1.

## Exercise 5 (P22)

Suppose we assume that the population distribution under $H_0$ is symmetric so that $\theta_{.5} = \mu$. Without looking at the data to check the validity of this assumption, we apply the binomial test and the CLT test. Suppose it turns out that 39 data values that are equal to 75.1 and the 40th one is equal to 90.

a. What decision is reached using the binomial test to test $H_0 : \theta_{.5} = 75$ versus $H_a : \theta_{.5} > 75$?

b. What decision is reached using the CLT test to test $H_0 : \mu = 75$ versus $H_a : \mu > 75$, where the statistics is computed using the sample standard deviation $S$ in place of the unknown population standard deviation $\sigma$?

c. Based on the results of parts a and b, what types of distributions that satisfy the alternative hypothesis are particually easy for the binomial test to detect in comparison to the CLT test?

d. Replace 90 by other values such as 80, 78 and 76 that are closer to the null hypothesis. Note what happens to the value of $Z_\mu$. Does this correspond to intuition?

***Solution:*** Suppose the significance level is 0.05.

a. Let $B = \#$ of data values that are higher than $75 = 40$.

$$\text{p-value} = \binom{40}{40}(0.5)^{40} \approx 9.09 \times 10^{-13} < 0.05.$$

We should reject $H_0$ using the binomial test.

b. The sample mean $\bar{X} = 75.4725$, and the sample standard deviation $S \approx 2.3559$. By CLT and using $S$ to estimate $\sigma$, under $H_0$, we have $\bar{X} \sim N(75, S/\sqrt{40})$.

$$Z_\mu = \frac{\bar{X} - 75}{S/\sqrt{40}} \approx 1.268$$

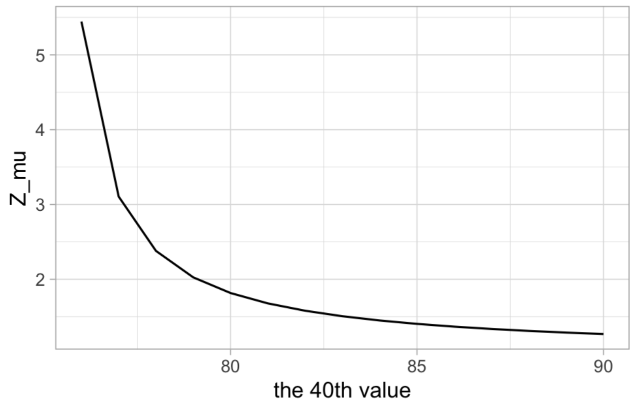
$$\text{p-value} = 1 - \Phi(Z_\mu) \approx 0.102 > 0.05.$$

We cannot reject $H_0$ using the CLT test.

c. The distributions that are highly skewed or heavily tailed.

d. $Z_\mu \approx \begin{cases} 1.268, & \text{the 40th value is 90} \\ 1.816, & \text{the 40th value is 80} \\ 2.379, & \text{the 40th value is 78} \\ 5.444, & \text{the 40th value is 76} \end{cases}.$

I also plot a figure for the variation of $Z_\mu$ with the decrease of the 40th value.




It turns out that more closer the 40th value is to 75 (the null hypothesis $H_0$), the larger $Z_\mu$ is. It corresponds to our intuition because a larger $Z_\mu$ means a larger p-value and thus a smaller possibility to reject $H_0$. When the 40th value gets closer to 75, the data is more centralized, less skewed and the sample mean gets closer to the null hypothesis, so it's reasonable that we have a larger $Z_\mu$.

## Exercise 6 (P22)

Refer to Section 1.3.3. No computation are required to answer the following question.

a. What is the value of the power of the binomial test when $\mu = 75$?

b. What happens to the power as $\mu$ gets large?

c. How does increasing the sample size affect the power of the binomial test?

***Solution:***

a. If $\mu = 75$, then $p = 0.5$ ($p$ is the probability that an observation is greater than 75 for a given value of $\mu$), and we have

$$Z_B = \frac{p - 0.5}{\sqrt{p(1-p)/N}}, \quad \eta := 1.645\sqrt{\frac{0.25}{p(1-p)}} - Z_B, \tag{1}$$
$$\text{power of binomial test} = 1 - \Phi(\eta) = 1 - \Phi(1.645) = 0.05.$$

But actually, power of a test is only defined when $H_0$ is not true, so when $\mu = 75$

which is the same to the null hypothesis, power shouldn't exist.

b. When $\mu$ gets large, $p$ also gets large (more likely that an observation $> 75$), and thus the probability of rejecting $H_0$ under $H_a$, namely the power, increases.

c. The power of the binomial test will increase if the sample size $N$ gets larger.

In equation (1), with $\mu$ fixed, we know when $N$ increases, $\sqrt{p(1-p)/N}$ decreases, $Z_B$ increases, $\eta$ decreases, $\Phi(\eta)$ decreases, and thus $1-\Phi(\eta)$, the power, increases.

## Exercise 7 (P22)

Suppose we test $H_0 : \theta_{.5} = \theta_H$ versus $H_a : \theta_{.5} > \theta_H$ using the binomial test with a sample size $n = 10$.

a. If we reject $H_0$ when $B \geq 8$, use the binomial Table A1 to determine the exact probability of a Type I error.

b. Suppose we observe a value of $B = b_{obs}$. The p-value is the probability that $B \geq b_{obs}$ given that $H_0$ is true. Find the p-values for $b_{obs} = 5, 6, 7, 8, 9, 10$.

***Solution:***

a. $B \sim \text{Binomial}(10, 0.5)$, so

$$\begin{aligned} P(\text{Type I Error}) &= P(\text{reject } H_0 | H_0) \\ &= P(B = 8) + P(B = 9) + P(B = 10) \\ &= 0.0439 + 0.0098 + 0.0010 \\ &= 0.0547. \end{aligned}$$

b. p-value $= P(B \geq b_{obs} | H_0) = \sum_{k=b_{obs}}^{10} P(B = k)$

Using the binomial Table 1 $(n = 10, p = 0.5)$,

| $k$ | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|
| $P$ | 0.2461 | 0.2051 | 0.1172 | 0.0439 | 0.0098 | 0.0010 |

for $b_{obs} = 5, 6, 7, 8, 9, 10$, we have p-value $= \begin{cases} 0.6231, & b_{obs} = 5 \\ 0.3770, & b_{obs} = 6 \\ 0.1719, & b_{obs} = 7 \\ 0.0547, & b_{obs} = 8 \\ 0.0108, & b_{obs} = 9 \\ 0.0010, & b_{obs} = 10 \end{cases}.$

## Exercise 8 (P22)

Refer to the derivations of the power functions in Section 1.3. Evaluate and sketch the power functions of the Statistics $Z_\mu$ and $Z_B$ for values of the mean between 75 and 77 assuming that the populations have normal distributions.

a. Using your sketch, determine the maximum difference between power functions.

b. Repeat this procedure for the Laplace population distribution.

***Solution:***

a. As defined in Exercise 5, 6, we have

$$Z_\mu = \frac{\mu - 75}{2.5/\sqrt{40}},$$
$$Z_B = \frac{p - 0.5}{\sqrt{p(1-p)/40}}.$$

The power function for CLT test is

$$1 - \Phi(1.645 - Z_\mu),$$

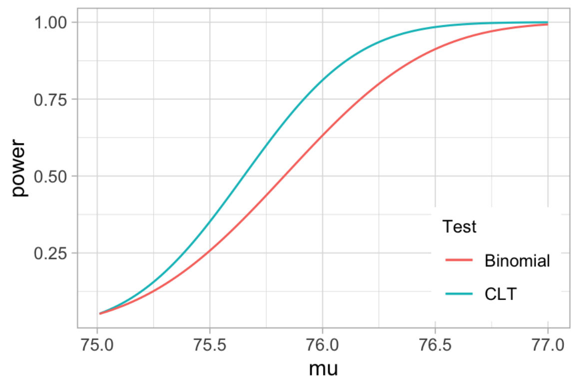
and the power funtion for binomial tests is

$$1 - \Phi(1.645\sqrt{\frac{0.25}{p(1-p)}} - Z_B).$$

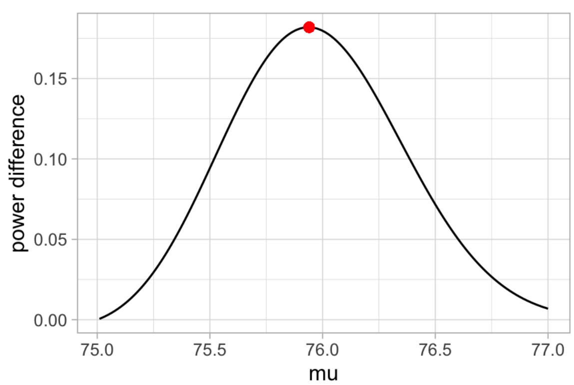
Use R to sketch the power functions of $Z_\mu$ and $Z_B$, and then determine the maximum difference (absolute value) between them.

```r
# Define the functions of Z scores.
Z_mu = function(mu){(mu - 75) / (2.5 / sqrt(40))}
Z_B = function(p){(p - 0.5) / sqrt(p * (1 - p) / 40)}

# Define the power functions.
pw_mu = function(mu){1 - pnorm(1.645 - Z_mu(mu))}
pw_B = function(mu){
  p = pnorm((mu - 75) / 2.5)
  1 - pnorm(1.645 * sqrt(0.25 / (p * (1 - p))) - Z_B(p))}
pw_Diff = function(mu){abs(pw_mu(mu) - pw_B(mu))}

# Set the values of mu, and calculate corresponding powers.
mu_list = seq(75.01, 77, 0.01)
data = data.frame('mu' = mu_list,
                  'CLT' = pw_mu(mu_list),
```

```r
                        'Bin' = pw_B(mu_list),
                        'Diff' = pw_mu(mu_list) - pw_B(mu_list))

# Plot the two power functions.
ggplot(data = data) +
  geom_line(aes(y = CLT, x = mu, color = 'CLT')) +
  geom_line(aes(y = Bin, x = mu, color = 'Binomial')) +
  labs(x = 'mu', y = 'power') +
  guides(color = guide_legend(title = 'Test')) +
  theme(legend.position = c(0.85, 0.2)) +
  theme_light()
```




```r
# Plot the difference between them (CLT minus Binomial).
ggplot() +
  geom_line(aes(y = data[, 2] - data[, 3], x = mu_list)) +
  geom_point(aes(y = Diff[94], x = mu[94]), color = 'red') +
  labs(x = 'mu', y = 'power difference') +
  theme_light()
```




From the sketch above, the maximum difference should be located near $\mu = 75.9$. By the optimization function in R,

```r
# Optimize the difference function.
optimise(pw_Diff, lower = 75 + 1e-5, upper = 77, maximum = T)
```

```
$maximum
[1] 75.93741
$objective
[1] 0.1818704
```

the maximum is reached when $\mu \approx 75.94$, with the objective value being 0.182.

b. If the population distribution changes from Normal to Laplace, then the only difference is the determination of $p$ in binomial test.

$$\begin{aligned} p = P(X > 75 \mid \mu) &= \int_{75}^{\infty} f(x \mid \mu) dx \\ &= \frac{1}{2\lambda} \int_{75}^{\infty} e^{-\frac{|x-\mu|}{\lambda}} dx \\ &(\text{let } y = \frac{x-\mu}{\lambda}) \\ &= \frac{1}{2} \int_{\frac{75-\mu}{\lambda}}^{\infty} e^{-|y|} dy \quad \mu \in (75, 77] \\ &= \frac{1}{2} + \frac{1}{2} \int_{0}^{\frac{\mu-75}{\lambda}} e^{-y} dy \\ &= \frac{1}{2} + \frac{1}{2}(1 - e^{-\frac{\mu-75}{\lambda}}) \\ &= 1 - 0.5e^{-\frac{\mu-75}{\lambda}} \end{aligned}$$

Suppose the standard deviation $\sigma$ of Laplace distribution is also 2.5.

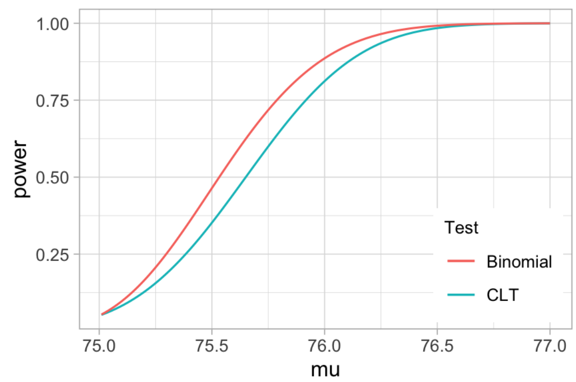
Notice that $\lambda \neq \sigma$! For the Laplace distribution, $D(X) = 2\lambda^2 = 2.5^2$, and thus $\lambda = \sqrt{3.125}$.

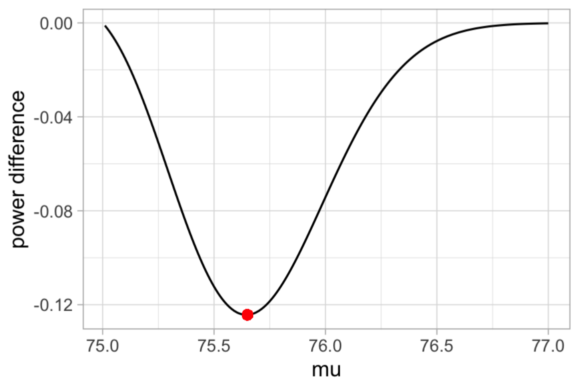
Or we could use $p = 1 - 0.5e^{-\sqrt{2}|x-\mu|/\sigma}$ derived on Page 21.

Change the power function of binomial test in R.

```r
pw_B = function(mu){
  p = 1 - 0.5 * exp(- (mu - 75) / sqrt(3.125))
  # p = 1 - 0.5 * exp(- sqrt(2) * (mu - 75) / 2.5)
  1 - pnorm(1.645 * sqrt(0.25 / (p * (1 - p))) - Z_B(p))}
```

Then repeat the plot procedures in question (a).




From the sketch above, the maximum difference should be located near $\mu = 76.0$. By the optimization function in R,

```r
# Optimize the difference function.
optimise(pw_Diff, lower = 75 + 1e-5, upper = 77, maximum = T)
```

```
$maximum
[1] 75.64519
$objective
[1] 0.124344
```

the maximum is reached when $\mu \approx 75.65$, with the objective value being 0.124.

***Summary:***

In conclusion, the power of the binomial test is less than that of the CLT test in the case of the Normal population, but greater in the case of the Laplace. Generally, the binomial test will have higher power than the CLT test for heavier-tailed population distributions, but the opposite will be true for lighter-tailed distributions.