MAST20005/MAST90058: Week 4 Lab

Goals: (i) Compute and interpret confidence intervals; (ii) Assess properties of estimators using numerical simulations.

Data for Section 1: Weight of primary tumour in mice (`PTweight.txt`). Observations represent measurements on primary tumour weight (in micro-grams) in mice randomly assigned to stress and control groups. Measurements are taken after a follow-up period. The goal of the study is to compare tumour development in the two groups of mice. (Data source: Sloan Biology Lab, Monash University)

Data for Section 2: Ivadopodia data (`invadopodia.txt`). Invadopodia are protrusions in cells which are often associated with cancer invasiveness and metastasis. The first dataset consists of counts for the number of cells susceptible of invadopodia development from mice tissue samples. The first column (`Condition`) denotes treatment group (1 = no treatment, 2 = isopropile (iso), 3 = propile (pro), 4 = iso + pro), while the second column (`Total`) denotes cell counts. (Data source: Sloan Biology Lab, Monash University)

1 Confidence intervals

1. While constructing confidence intervals, we will often use R to find quantiles of common distributions, including χ^2-, t- and F-distributions. For each of the following distributions, find the 0.9, 0.95 and 0.975 quantiles.

   (a) Standard normal, $\mathrm{N}(0, 1)$

   ```
   p <- c(0.9, 0.95, 0.975)
   qnorm(p)

   ## [1] 1.281552 1.644854 1.959964
   ```

   (b) $\mathrm{N}(5, 3^2)$

   ```
   qnorm(p, 5, 3)
   ```

   (c) t_5

   ```
   qt(p, 5)
   ```

   (d) χ^2_1

   ```
   qchisq(p, 1)
   ```

   (e) χ^2_5

   ```
   # your turn...
   ```

   (f) $\mathrm{F}_{12,4}$

   ```
   qf(p, 12, 4)
   ```

2. Load the primary tumour data into a data frame called `PTweight` (how?). Find an approximate 90% confidence interval for the mean tumour weight in the control group.

```
x <- PTweight[1:5, 2]
n <- length(x)
x.bar <- mean(x)
s <- sd(x)
t <- qt(0.95, n - 1)
x.bar + c(-1, 1) * t * s / sqrt(n)

## [1] 189.0001 791.1999
```

Notice that in the last line the vector $(-1, 1)$ is multiplied and added to scalars. This produces a two-element vector corresponding to $\bar{x} \pm t_{n-1}^{-1}(0.05)s/\sqrt{n}$. Next compare your result with the in-built function `t.test`.

```
t.test(x, conf.level = 0.90)

##
##  One Sample t-test
##
## data:  x
## t = 3.47, df = 4, p-value = 0.02558
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  189.0001 791.1999
## sample estimates:
## mean of x
##     490.1
```

The last command is what we use in practice. We will look at the hypothesis testing part of the output later in the course.

3. Construct a 95% confidence interval for the difference of means $\mu_X - \mu_Y$ in the case and control groups. Can you conclude that stress and tumour growth are associated? (Assume equal variances in the two groups, i.e. $\sigma_X^2 = \sigma_Y^2$)

```
y <- PTweight[6:10, 2]  # stress group data
y.bar <- mean(y)
s.p <- sqrt((4 * var(x) + 4 * var(y)) / 8)  # pooled sample sd
x.bar - y.bar + c(-1, 1) * qt(0.975, df = 8) * s.p * sqrt(1 / 5 + 1 / 5)

## [1] -662.5751  153.2951
```

The value $\mu_X - \mu_Y < -0$ is inside the interval. Thus, there is not enough evidence in these data to claim that the means in the two groups are different.

Compare now with the result from `t.test`.

```r
t.test(x, y, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = -1.4394, df = 8, p-value = 0.188
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -662.5751  153.2951
## sample estimates:
## mean of x mean of y
##    490.10    744.74
```

4. Is it reasonable to assume that the two variances are the same in the two groups? This can be checked formally by computing a confidence interval for the variance ratio σ_X^2/σ_Y^2 as follows:

```r
var.test(x, y)
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 1.7583, num df = 4, denom df = 4, p-value = 0.598
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   0.183065 16.887192
## sample estimates:
## ratio of variances
##           1.758253
```

Since the value $\sigma_X^2/\sigma_Y^2 = 1$ is inside the CI, there is not enough evidence to suggest that the variances are unequal.

2 Comparing estimators by simulation

1. Let $X_1, \ldots, X_n$ be a random sample of size n from a Poisson distribution with mean λ. Let $\bar{X}$ and S^2 denote the sample mean and sample variance, respectively. Note that both are unbiased estimators for λ since

$$\mathbb{E}(\bar{X}) = \mathbb{E}(S^2) = \lambda.$$

While we already know that $\text{var}(\bar{X}) = \lambda/n$ (recall how), an expression for $\text{var}(S^2)$ is harder to compute. In such situations computer simulations can help us compare estimators. First let us check that the two estimators are unbiased.

```r
lambda <- 10
B <- 1000  # simulation runs
n <- 10    # sample size
xbar <- 1:B  # we will collect results in these two vectors
s2   <- 1:B
for (b in 1:B) {  # repeat B times for b = 1,...,B
    x <- rpois(n, lambda)
    xbar[b] <- mean(x)  # compute and store X-bar
    s2[b]   <- var(x)   # compute and store S^2
}
```

The LLN for iid variables $Z_1, \ldots, Z_B$ states that $B^{-1} \sum_{b=1}^{B} Z_b \to \mathbb{E}(Z_1)$ as $B \to \infty$. This can be used to approximate the $\mathbb{E}(\bar{X})$ and $\mathbb{E}(S^2)$ from our simulations as follows

```r
mean(xbar)

## [1] 9.9721

mean(s2)

## [1] 9.8617
```

Both estimators seem to be unbiased for λ.

Using the LLN we can also approximate the variance of the two estimators as follows:

```r
var(xbar)

## [1] 1.054046

var(s2)

## [1] 20.92035
```

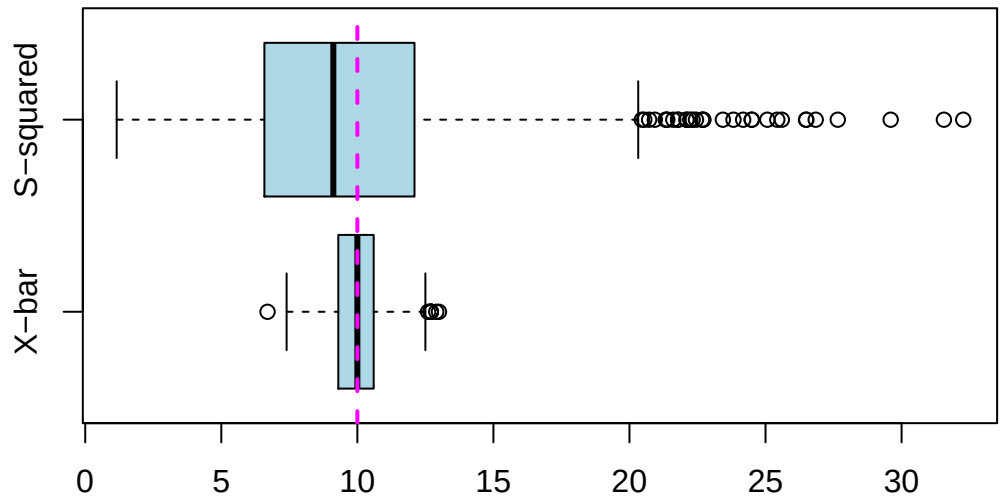
While we already know that $\text{var}(\bar{X}) = \lambda/n$, a more involved calculation shows that $\text{var}(S^2) = [\lambda(2n\lambda + n - 1)]/[n(n-1)]$. The values obtained from the above simulation are quite close to these theoretical values:

```r
lambda / n

## [1] 1

lambda * (2 * n * lambda + n - 1) / (n * (n - 1))

## [1] 23.22222
```

Both estimators seem to be unbiased for λ but the first estimator is clearly superior in terms of variance. Try different values of the true parameter λ and n in the above code and see what happens. Also, check what happens when you increase B.

2. Compare the accuracy of the estimators graphically, for example using boxplots.

```
boxplot(xbar, s2, names = c("X-bar", "S-squared"),
        col = "lightblue", horizontal = TRUE)

# Draw a dashed vertical line showing the true parameter value.
abline(v = lambda, lty = 2, lwd = 2, col = "magenta")
```



From the box plots, the distribution of both estimators is clearly centred around the true λ, but the distribution of $\bar{X}$ has a much smaller spread. Note also that the distribution of S^2 is not symmetric.

3. Load the invadopodia data. Assume a Poisson model with mean λ_j, $j = 1, \ldots, 4$ with different means corresponding to treatment groups and estimate λ_j. For the first two groups we have:

```
invadopodia <- read.table("invadopodia.txt")
x1 <- invadopodia[invadopodia$Condition == 1, 2]
x2 <- invadopodia[invadopodia$Condition == 2, 2]
x.bar1 <- mean(x1)
x.bar2 <- mean(x2)
```

4. Construct approximate 95% confidence intervals for λ recalling that $\hat{\lambda} = \bar{X}$ obeys the Central Limit Theorem. Specifically, $\hat{\lambda} \approx \mathrm{N}(\lambda, \lambda/n)$, so approximate 95% confidence intervals can be computed using:

$$\hat{\lambda}_j \pm 1.96 \times \sqrt{\frac{\hat{\lambda}_j}{n}}, \quad j = 1, \ldots, 4.$$

```r
# 95% CI for group 1.
x.bar1 + c(-1, 1) * 1.96 * sqrt(x.bar1 / length(x1))

## [1] 4.871097 6.684459

# 95% CI for group 2.
x.bar2 + c(-1, 1) * 1.96 * sqrt(x.bar2 / length(x2))

## [1] 3.057910 4.293441
```

Note that 1.96 in the above code gives 95% confidence intervals. To compute 90% or 99% CIs, replace 1.96 by the appropriate standard normal quantiles:

```r
qnorm(0.95)   # use this to obtain a 90% CI

## [1] 1.644854

qnorm(0.995)  # use this to obtain a 99% CI

## [1] 2.575829
```

5. The CIs do not overlap, suggesting that the two means are actually different. To carry out proper inference for the difference of means, $\lambda_1 - \lambda_2$, we need to derive a specific CI. Since the first two groups are independent we have,

$$\hat{\lambda}_1 - \hat{\lambda}_2 \approx \mathrm{N}\left(\lambda_1 - \lambda_2, \frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}\right)$$

```r
# 95% CI for difference between groups 1 and 2.
x.bar1 - x.bar2 + c(-1, 1) * 1.96 *
    sqrt(x.bar1 / length(x1) + x.bar2 / length(x2))

## [1] 1.004967 3.199237
```

The interval is clearly above the value $\lambda_1 - \lambda_2 = 0$, suggesting that the true difference is likely to be different from zero. Hence, we conclude that we have evidence that the isopropile treatment reduces the number of susceptible cells in mice tissues.

3 Simulating discrete distributions

The function `sample()` carries out sampling from a discrete distribution in a few different ways. For example, to simulate 10 coin tosses you would use:

```r
sample(c("tails", "heads"), 10, replace = TRUE)

##  [1] "tails" "tails" "heads" "heads" "tails" "heads" "heads" "heads"
##  [9] "tails" "tails"
```

The first argument specifies the possible values to observe. For example, we could pick a random day of the week using:

```
days <- c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")
sample(days, 1)

## [1] "Fri"
```

By default, all items are given equal probability. This can be changed by specifying different values using the `prob` argument. For example, biased coin tosses can be obtained with:

```
sample(c("tails", "heads"), 10, prob = c(0.2, 0.8), replace = TRUE)

##  [1] "heads" "heads" "heads" "heads" "heads" "tails" "heads" "heads"
##  [9] "heads" "tails"
```

The vector passed to `prob` should have the same length as the set of possible values (the first argument).

The argument `replace` specifies whether to do sampling with or without replacement[1]. The default is without replacement, which means we always need to write `replace = TRUE` to get iid samples.

Exercises

1. Give estimates and 95% CIs for λ_3 and λ_4.
2. The investigators are wondering whether isopropile treatment affects the number of susceptible cells in mice tissues when propile treatment is already provided. What analysis answers that question? Carry out this analysis.
3. Consider question 1 from the tutorial problems. Derive a 75 % CI for the population mean.
4. Consider question 2 from the week 3 tutorial problems. This involved a random sample of n observations on X having the following pmf:

x	0	1	2
$p(x)$	$1-\theta$	$3\theta/4$	$\theta/4$

   For the case of $n = 10$ and $\theta = 0.6$, use numerical simulations to show:

   (a) $T_1 = (4/5)\bar{X}$ and $T_2 = 1 - n^{-1}\sum_{i=1}^{n} I(X_i = 0)$ are unbiased

   (b) $\text{var}(T_1) > \text{var}(T_2)$

[1] 'Replacement' refers to putting the sampled item back into the pool of possible values before taking the next sample.