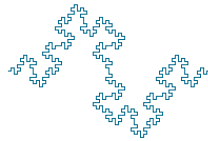


High-Throughput Sequencing Course

Statistical Inference: Part II

Biostatistics and Bioinformatics



Summer 2018

TWO-SAMPLE MODEL: INFERENCE

- ▶ The mRNA abundance level in the untreated population is μ_0
- ▶ The mRNA abundance level in the treated population is μ_1
- ▶ Assumed model:
 - ▶ Untreated Population: $Y = \mu_0 + \epsilon$
 - ▶ Treated Population: $X = \mu_1 + \epsilon'$
- ▶ Statistical Hypotheses
 - ▶ $H_0 : \mu_0 = \mu_1$ (no treatment effect)
 - ▶ $H_a : \mu_0 \neq \mu_1$ (treatment effect)

TWO-SAMPLE MODEL: ESTIMATION

- ▶ What is often of interest is estimate the unknown parameters or quantities
- ▶ Examples
 - ▶ Mean level for the untreated group μ_0
 - ▶ Mean level for the treated group μ_1
 - ▶ Fold-change $\rho = \frac{\mu_1}{\mu_0}$
 - ▶ Standardized difference $\Delta = |\mu_1 - \mu_0|/\sigma$
- ▶ Two types of estimates
 - ▶ Point estimate
 - ▶ Interval estimate

CONFIDENCE INTERVALS

- ▶ Example: The sample mean (the average of the observations) is a point estimate of the population (true) mean
- ▶ It is either equal to the true value of the parameter or is not
- ▶ As it is a single number it does not provide any direct measure of accuracy
- ▶ An interval estimate incorporates some measure of accuracy
- ▶ Thus it is generally more appropriate to present an interval estimate
- ▶ A common example of an interval estimate is the confidence interval

ESTIMATION EXAMPLE (ONE-SAMPLE MODEL)

- ▶ Truth: The RNA abundance follows a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$
- ▶ Assumption: The RNA abundance follows a normal distribution with *unknown* mean μ and *unknown* standard deviation σ
- ▶ Goal: The population mean μ is to be estimated on the basis of sample of size $n = 7$
- ▶ Objectives:
 - ▶ Produce point estimate of μ
 - ▶ Produce a 95% confidence interval of μ

ESTIMATION EXAMPLE (SIMULATE DATA)

```
mu <- 0
sigma <- 1
n <- 7
set.seed(12131)
x <- rnorm(n, mu, sigma)
x

## [1]  1.5227356 -2.7829224  0.3571897  0.5478351  1.2733071  0.5166791
## [7] -1.3890287
```

POINT ESTIMATOR

- ▶ A point estimator of μ is the so called sample mean
- ▶ The sample mean \bar{x}_n is obtained by simply averaging all the observations
- ▶ Note that an alternative is to use the sample median (rather than sample mean)
- ▶ The sample median is obtained by first sorting the observations (in say ascending order)
- ▶ The median is the middle observation (among the sorted observation)
- ▶ The median is more robust against outliers

POINT ESTIMATES

- ▶ The data

```
x
## [1]  1.5227356 -2.7829224  0.3571897  0.5478351  1.2733071  0.5166791
## [7] -1.3890287
```

- ▶ The sample mean

```
mean(x)
## [1] 0.006542226
```

- ▶ The data sorted in ascending order

```
sort(x)
## [1] -2.7829224 -1.3890287  0.3571897  0.5166791  0.5478351  1.2733071
## [7]  1.5227356
```

- ▶ Sample median

```
median(x)
## [1] 0.5166791
```

CONFIDENCE INTERVAL ESTIMATORS

- ▶ To construct a confidence interval for μ we need to deal with the nuisance parameter σ
- ▶ We can estimate it using the sample standard deviation s_n (details omitted)
- ▶ A 95% confidence interval for μ is obtained as

$$[\bar{x}_n - \frac{s_n}{\sqrt{n}}t(0.975, n-1), \bar{x}_n + \frac{s_n}{\sqrt{n}}t(0.975, n-1)]$$

- ▶ $t(0.975, n-1)$ is the 0.975 quantile of a t distribution with $n-1=6$ degrees of freedom
- ▶ $\frac{s_n}{\sqrt{n}}$ is called the standard error
- ▶ $\frac{s_n}{\sqrt{n}}t(0.975, n-1)$ is called the margin of error
- ▶ The confidence interval is obtained as the point estimate plus or minus the margin of error

SIMULATE EXPERIMENT 1

- Calculate the sample mean

```
xbar <- mean(x)
xbar
## [1] 0.006542226
```

- Calculate standard deviation

```
s <- sd(x)
s
## [1] 1.544261
```

- Calculate standard error

```
se <- s/sqrt(n)
se
## [1] 0.5836759
```

- Calculate margin of error

```
me <- qt(0.975, df = n - 1) * se
me
## [1] 1.428204
```

- Calculate 95% CI

```
c(xbar - me, xbar + me)
## [1] -1.421661 1.434746
```

COVERED OR NOT COVERED

- The goal is to estimate μ
- If μ (the true but unknown parameter) is contained in the confidence interval, we say that it is covered
- Otherwise, it is not covered
- Note that when doing a simulation study, we can ascertain if μ is covered or not.
- Why?
- In real data analysis, we cannot ascertain if μ is covered by the confidence interval
- Why?
- We can only state that we are 95% *confident* that μ is covered by the interval estimate based on the data from our experiment
- More on "confidence" later

REPEAT THE EXPERIMENT

```
set.seed(12301)
makeest <- function(b, n, mu, sigma, alpha) {
  x <- rnorm(n, mu, sigma)
  xbar <- mean(x)
  s <- sd(x)
  me <- qt(1 - alpha/2, df = n - 1) * s/sqrt(n)
  lcl <- xbar - me
  ucl <- xbar + me
  cover <- ifelse(mu >= lcl && mu <= ucl, TRUE, FALSE)
  data.frame(exp = b, n, mu, sigma, xbar, s, lcl, ucl, cover, len = ucl - lcl)
}
res <- foreach(b = 1:10, .combine = rbind) %do% {
  makeest(b, n, mu, sigma, 0.05)
}
```

REPEAT THE EXPERIMENT 10 TIMES

exp	n	mu	sigma	xbar	s	lcl	ucl	cover	len
1	7	0	1	0.48	0.42	0.09	0.87	FALSE	0.78
2	7	0	1	0.34	0.88	-0.47	1.15	TRUE	1.63
3	7	0	1	-0.51	1.18	-1.60	0.58	TRUE	2.18
4	7	0	1	-0.87	0.67	-1.49	-0.25	FALSE	1.24
5	7	0	1	-0.09	0.95	-0.97	0.78	TRUE	1.76
6	7	0	1	0.30	1.62	-1.20	1.80	TRUE	3.00
7	7	0	1	-0.68	0.52	-1.15	-0.20	FALSE	0.96
8	7	0	1	0.06	1.30	-1.15	1.26	TRUE	2.41
9	7	0	1	0.28	1.02	-0.66	1.23	TRUE	1.89
10	7	0	1	-0.31	0.48	-0.76	0.14	TRUE	0.89

CONFIDENCE INTERVAL: COMMON MISUNDERSTANDING

- ▶ A (not the) 95% CI for the mean based on the first experiment was $(0.09, 0.87)$
- ▶ A (not the) 95% CI for the mean based on the second experiment was $(-0.47, 1.15)$
- ▶ It is wrong to say that the probability that the first CI does not contain the true value $\mu = 0$ is 95%
- ▶ It is also wrong to say that the probability that the second CI contains the true value $\mu = 0$ is 95%
- ▶ We conduct one and only one experiment
- ▶ Based on the first experiment, we can say that we are 95% confident that it contains the true value
- ▶ Note that μ is *not* covered by the first experiment
- ▶ If we repeated the experiment a large number of times, 95% of the CIs would cover the true value
- ▶ We are 95% confident that the first experiment is among these (which it is not)

RECAP: ASSUMPTIONS

- ▶ We do not need to make distributional assumptions (e.g., normality) on the sample mean for the purpose of point estimation
- ▶ The sample mean, however, is not robust against outliers
- ▶ Why did 1984 UNC geography graduates have high average salary?
- ▶ We made distributional assumptions for using the confidence interval
- ▶ The margin of error was based on a t distribution

A MORE COMPLICATED EXAMPLE: OUTLINE

- ▶ Suppose that you are measuring a quantity that is between 0 and θ
- ▶ How would you estimate θ ?
- ▶ Would you take the sample average?
- ▶ How about the sample mean?
- ▶ If the measurements are uniformly distributed, it turns out that the maximum observation is an "optimal" estimator
- ▶ It is also intuitively speaking a "reasonable" estimator
- ▶ Why?

A MORE COMPLICATED EXAMPLE: SIMULATION

- ▶ Simulate data from a uniform distribution on $[0, 1]$

```
n <- 10
theta <- 1
set.seed(2313)
x <- runif(n, 0, theta)
x

## [1] 0.34807917 0.12084940 0.11035999 0.03917718 0.79590237 0.72536724
## [7] 0.80347454 0.95498314 0.62601926 0.19549397
```

- ▶ Sample mean

```
mean(x)

## [1] 0.4719706
```

- ▶ Sample median

```
median(x)

## [1] 0.4870492
```

- ▶ Maximum observation

```
max(x)

## [1] 0.9549831
```

A MORE COMPLICATED EXAMPLE: RECAP

- ▶ An estimator is "valid" if it depends only on the data and no unknown quantities (including the parameter to be estimated)
- ▶ Why?
- ▶ Both the sample mean and median are *valid* estimators of θ
- ▶ There are, however, not good estimators
- ▶ In fact, in this case, the sample mean and median should be close to 0.5
- ▶ Why?
- ▶ The maximum observation is not only a valid estimator but also intuitively reasonable estimator
- ▶ This example has a rich history

QUICK NOTE: ESTIMATE VERSUS ESTIMATOR

- ▶ We use the terms estimate and estimators interchangeably
- ▶ There is a subtle but important distinction
- ▶ Suppose that you decide to estimate the population mean using the sample mean (once you get the data)
- ▶ The sample mean is the estimator
- ▶ Its outcome is random before you collect the data
- ▶ Once you collect the data and plug them into the estimator you get an (not the) estimate