Inference Concept Check

Chapter 4, Lab 4: Solutions

OpenIntro Biostatistics

Topics

- Relationship between hypothesis tests and confidence intervals
- One-sided hypothesis tests and confidence intervals
- Choosing between one-sided and two-sided tests
- Definition of α as the acceptable error probability

The previous labs in this unit discussed the mechanics of calculating confidence intervals and conducting hypothesis tests. This lab focuses on examining some conceptual details of inference.

The material in this lab corresponds to Sections 4.3.3 - 4.3.5 of *OpenIntro Biostatistics*.

Hypothesis Testing and Confidence Intervals

Suppose we would like to assess whether the mean weight in the `yrbss` "population" is different from 60 kilograms, using the information from a random sample. The YRBSS data was introduced in the first lab in this unit (Chapter 4, Lab 1).

Run the following code chunk to take a random sample of size 100 from `yrbss.complete`, the version of `yrbss` that has no missing weight values.

```
#load the dataset
library(oibiostat)
data("yrbss")

#remove rows with missing values
yrbss.complete = yrbss[complete.cases(yrbss$weight), ]

#set parameters
sample.size = 100

#set seed for pseudo-random sampling
set.seed(5011)

#obtain random sample of row numbers
sample.rows = sample(1:nrow(yrbss.complete), sample.size)

#create yrbss.sample
yrbss.sample = yrbss.complete[sample.rows, ]
```

1. Calculate a 95% confidence interval for μ_{weight} using the data in `yrbss.sample`. Does it seem like the population mean weight will be different from 60 kilograms?

   The 95% confidence level is (63.238, 70.257) kg. It seems like population mean weight will be different from 60 kilograms.

```
#calculate confidence interval
t.test(yrbss.sample$weight, conf.level = 0.95)$conf.int
```

```
## [1] 63.23826 70.25754
## attr(,"conf.level")
## [1] 0.95
```

2. Test $H_0 : \mu_{weight} = 60$ kg versus $H_A : \mu_{weight} \neq 60$ kg at the $\alpha = 0.05$ significance level.

   a) What point on the t-distribution (with $df = 99$) has area of 0.025 to the left? What point on the t-distribution (with $df = 99$) has area of 0.025 to the right?

   These points mark off the *rejection region* as defined by α; the t-statistic calculated from $\bar{x}$ must lie within the tail areas bounded by these points in order to constitute sufficient evidence against H_0.

   The points -1.98 and 1.98 mark the rejection regions for $\alpha = 0.05$ on a t-distribution that has 99 degrees of freedom.

```
#t-point marking the upper tail
qt(0.975, df = 99)
```

```
## [1] 1.984217
```

```
#t-point marking the lower tail
qt(0.025, df = 99)
```

```
## [1] -1.984217
```

   b) Calculate the t-statistic from `yrbss.sample`. Confirm that it lies within the rejection region and that the associated p-value is less than α.

   The t-statistic calculated from weights in `yrbss.sample` is 3.82, which is within the upper tail, beyond the t point of 1.98. The p-value of 0.0002 is much smaller than α.

```
t.test(yrbss.sample$weight, mu = 60, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  yrbss.sample$weight
## t = 3.815, df = 99, p-value = 0.0002371
## alternative hypothesis: true mean is not equal to 60
## 95 percent confidence interval:
##  63.23826 70.25754
## sample estimates:
## mean of x
##   66.7479
```

3. The relationship between a hypothesis test and the corresponding confidence interval is defined by α. Suppose that a two-sided test is conducted at significance level α; the confidence level of the matching interval is (1 - α)%. For example, a 95% confidence interval can be compared to a two-sided hypothesis test with $\alpha = 0.05$.

   a) What is the margin of error, m, from the confidence interval calculated in Question 1?

      The margin of error is half the width of a confidence interval ($\bar{x} \pm m$). Thus, the margin of error in the confidence interval (63.238, 70.257) kg is $(70.257 - 63.238)/2 = 3.510$.

```
#use r as a calculator
x.bar = mean(yrbss.sample$weight)
s = sd(yrbss.sample$weight)
n = 100
t.star = qt(0.975, df = 99)

m = t.star * (s/sqrt(n))
m
```

```
## [1] 3.509638
```

```
x.bar - m; x.bar + m
```

```
## [1] 63.23826
```

```
## [1] 70.25754
```

b) What values of $\bar{x}$ would correspond to the t-points that mark off the rejection region? In other words, what weight values for the sample mean would be considered 'extreme' enough to warrant rejecting H_0?

The weight values of 63.51 kg and 56.49 kg would correspond to the t-points that mark off the rejection regions in the upper and lower tails, respectively. A sample mean would be considered extreme if it were greater than 63.51 or lower than 56.49.

```
#use r as a calculator
mu.0 = 60
s = sd(yrbss.sample$weight)
n = length(yrbss.sample$weight)

t.upper.bound = qt(0.975, df = 99)
t.lower.bound = qt(0.025, df = 99)

x.upper = mu.0 + (s/sqrt(n))*t.upper.bound
x.lower = mu.0 + (s/sqrt(n))*t.lower.bound

x.upper; x.lower
```

```
## [1] 63.50964
```

```
## [1] 56.49036
```

c) How far apart are the two values calculated in part b) from each other? How does the distance relate to m, as calculated in part a)?

The two values calculated in part b) are 7.02 units away from each other, which is exactly twice the margin of error m calculated in part a).

The purpose of this question is to illustrate how a hypothesis test and confidence level with the same α have the quantity $m = t^\star(s/\sqrt{n})$ in common.

- **Hypothesis Test.** How far does $\bar{x}$ need to be from μ_0 in order to be considered extreme? For a two-sided test, m units away in either direction. Note how in part b), the t-points marking off the rejection region are equal to the $t^\star$ value used in the confidence interval; the +/- t-points account for the $\pm$ structure in the confidence interval.
- **Confidence Interval.** What are the plausible range of values for μ_0 around $\bar{x}$? The range is defined as $(\bar{x} - m, \bar{x} + m)$; if μ_0 is plausible, it can at most be m units away in either direction from $\bar{x}$. Does the interval contain μ_0? If not, μ_0 is implausible according to α and H_0 is rejected.
- A matter of perspective: The hypothesis testing approach asks whether $\bar{x}$ is far enough away from μ_0, and the confidence interval approach asks whether μ_0 is close enough to $\bar{x}$. In both cases, "far enough" and "close enough" are defined by α, which determines the $t^\star$ used in calculating m.

One-sided hypothesis tests and confidence intervals

One-sided confidence intervals for a population mean provide either a lower bound or upper bound, but not both. One-sided confidence intervals have the form:

$$(\bar{x} - m, \infty) \text{ or } (-\infty, \bar{x} + m), \text{where } m = \frac{s}{\sqrt{n}} t^\star.$$

In general, for a confidence interval of $(1 - \alpha)100\%$, $t^\star$ is the point on a t distribution with $n - 1$ degrees of freedom that has area 1 - α to the left. For a 95% confidence interval, $\alpha = 0.05$; $t^\star$ is the point on a t distribution with area $1 - 0.05 = 0.95$ to the left.

A one-sided hypothesis test with $\alpha = 0.05$ and $H_A : \mu > \mu_0$ corresponds to a one-sided 95% confidence interval that has a lower bound, but no upper bound: $(\bar{x} - m, \infty)$. If instead, $H_A : \mu < \mu_0$, the test corresponds to a one-sided 95% interval with no lower bound: $(-\infty, \bar{x} + m)$.

4. Suppose that in a random sample of 150 adults from the US population, the average amount of nightly sleep is 6.80 hours, with standard deviation 1.60 hours.

   a) Is there evidence that on average, American adults sleep less than 7 hours per night? Use $\alpha = 0.10$.

   The null hypothesis is that American adults sleep 7 hours a night, $H_0 : \mu = 7.0$ hours. The alternative hypothesis is that American adults sleep less than 7 hours a night, $H_A : \mu < 7.0$ hours. The t-statistic is -1.53, and the associated p-value is $P(T \leq -1.53) = 0.06$. Since the one-sided alternative hypothesis is $\mu < \mu_0$, the p-value is the lower tail area; i.e., the area to the left of the t-statistic.

Since p is less than α, there is sufficient evidence to reject the null hypothesis and accept the alternative that American adults sleep less than 7 hours per night. The low p-value indicates that if the true mean hours of sleep per night in the American population were 7 hours, then it would be unusual to observe a sample with mean 6.80 hours.

```
#use r as a calculator
x.bar = 6.80
mu.0 = 7
s = 1.60
n = 150

#calculate the t-statistic
t = (x.bar - mu.0)/(s/sqrt(n))
t
```

```
## [1] -1.530931
```

```
#calculate the p-value
pt(t, df = n - 1, lower.tail = TRUE)
```

```
## [1] 0.0639534
```

b) Calculate a one-sided lower 90% confidence interval for the average amount of sleep per night among US adults; i.e., of the form $(-\infty, \bar{x} + m)$. Compare the information obtained from a confidence interval versus a hypothesis test and assess whether the results of the test are practically significant, as opposed to statistically significant.

The upper 90% bound is 6.97 hours. Based on the data in this sample, we can be 90% confident that the population mean hours of sleep per night among US adults is at most 6.97 hours.

The interval does not contain the μ_0 value of 7.0 hours; 7.0 hours is not a plausible value for μ_0 according to the confidence interval, which agrees with the conclusion from part a). While the interval provides a plausible range of values for the population mean hours sleep per night, the p-value from the test gives a measure of the strength of the evidence against H_0.

Note that while the results of the test are statistically significant, but not necessarily practically significant. 6.97 hours is only about 2 minutes less than 7 hours. Thus, from a practical standpoint that 2 minutes of sleep is a negligible amount, this sample does not provide convincing evidence that American adults sleep a meaningful amount less than 7 hours per night.

```
#use r as a calculator
x.bar = 6.80
s = 1.60
n = 150
t.star = qt(0.90, df = n - 1)

m = s/(sqrt(n)) * t.star
ci.upper.bound = x.bar + m
```

```
ci.upper.bound
```

```
## [1] 6.968167
```

c) Suppose we were instead interesting in testing whether American adults sleep more than 7 hours per night on average. Calculate the p-value for the test and the corresponding 90% confidence interval.

The t-statistic remains the same: -1.53. However, the p-value is in the upper tail since the direction of the alternative has changed; the area corresponding to the p-value is in the direction specified by the alternative. Thus, $p = P(T \geq -1.53) = 0.94$.

The upper 90% bound is 6.63 hours. We can be 90% confident that the population mean hours of sleep per night among American adults is at least 6.63 hours.

```
#use r as a calculator
pt(t, df = n - 1, lower.tail = FALSE)
```

```
## [1] 0.9360466
```

```
ci.lower.bound = x.bar - m
ci.lower.bound
```

```
## [1] 6.631833
```

Choosing between one-sided and two-sided tests

5. A standard test for diabetes involves measuring blood sugar levels after an overnight fast. Someone without diabetes typically has a fasting blood sugar level of around 5.31 mmol/L. High fasting blood sugar levels are indicative of diabetes or prediabetes.

   Neighborhood and community-level factors are known to influence diabetes risk. People living in poorer neighborhoods tend to be at higher risk of diabetes; these neighborhoods often lack grocery stores, recreational facilities, and green space.

   a) For the following scenarios, choose whether to conduct a one-sided or two-sided test. Formulate null and alternative hypotheses.

      i. Suppose that you are interested in learning whether mean fasting blood sugar levels in Neighborhood A, a middle-class neighborhood, is different from 5.31 mmol/L.

      A two-sided alternative seems reasonable; there is not information to suggest that population fasting blood sugar levels in this neighborhood is higher or lower than normal.

      $H_0 : \mu = 5.31$ mmol/L, $H_A : \mu \neq 5.31$ mmol/L.

      ii. Neighborhood B is primarily inhabited by low-income families. If the mean fasting blood sugar level in Neighborhood B is higher than normal, you will start recruiting participants for an antidiabetic drug trial from the neighborhood.

      It is reasonable to expect that if the null hypothesis is not true, then the mean fasting blood sugar level in this neighborhood is higher than 5.31 mmol/L.

$H_0 : \mu = 5.31$ mmol/L, $H_A : \mu > 5.31$ mmol/L.

Additionally, the consequences in this scenario are the same whether we fail to reject H_0 or if it is actually the case that mean fasting blood sugar level is less than 5.31 mmol/L; we will not choose to recruit participants for an antidiabetic drug trial.

b) The dataset `sugar.levels.A` contains simulated fasting blood sugar levels for 100 individuals from Neighborhood A; `sugar.levels.B` contain the simulated data from Neighborhood B. Both datasets are in the `oibiostat` package.

i. Use `t.test()` to test the hypotheses from part a). Use $\alpha = 0.05$. Summarize your conclusions.

The p-value for the Neighborhood A data is 0.06; there is not sufficient evidence at the $\alpha = 0.05$ significance level to reject the null hypothesis that the mean fasting blood sugar in this neighborhood is 5.31 mmol/L.

The p-value for the Neighborhood B data is 0.03, which is less than $\alpha = 0.05$. There is sufficient evidence to reject the null and accept the alternative hypothesis that the mean fasting blood sugar in this neighborhood is higher than 5.31 mmol/L.

```
#load datasets
data("sugar.levels.A")
data("sugar.levels.B")

#conduct hypothesis tests
t.test(x = sugar.levels.A, mu = 5.31, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  sugar.levels.A
## t = 1.9122, df = 99, p-value = 0.05874
## alternative hypothesis: true mean is not equal to 5.31
## 95 percent confidence interval:
##  5.298065 5.955605
## sample estimates:
## mean of x
##  5.626835
```

```
t.test(x = sugar.levels.B, mu = 5.31, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  sugar.levels.B
## t = 1.9122, df = 99, p-value = 0.02937
## alternative hypothesis: true mean is greater than 5.31
## 95 percent confidence interval:
##  5.35172     Inf
## sample estimates:
```

```
## mean of x
##  5.626835
```

ii. The values in the two datasets are actually identical. Why are the p-values from the two tests different from each other?

The p-value for a one-sided test is exactly half the p-value for a two-sided test conducted at the same significance level. For a two-sided test, $p = P(T \geq |t|) = 2P(T \geq t)$. The p-value for a one-sided test is only $P(T \geq t)$ or $P(T \leq t)$, depending on the direction of the alternative. A result needs to be more extreme to be within the rejection region of a two-sided test.

The definition of α as the acceptable error probability

The significance level α can be thought of as the value that quantifies how rare or unlikely an event must be in order to represent sufficient evidence against the null hypothesis. For example, an α level of 0.05 means that an event occurring with probability lower than 5% will be considered sufficient evidence against H_0.

It is also possible to think about α in the context of decision errors. Hypothesis tests can potentially result in incorrect decisions. Rejecting the null hypothesis when the null is true is referred to as a **Type I** error. A Type I error occurs with probability α, since α determines the cutoff point for rejecting the null hypothesis. If $\alpha = 0.05$, there is a 5% chance of incorrectly rejecting H_0.

To explore this idea, let's return to the `yrbss` repeated sampling simulation.

Recall that `yrbss.complete` is our artificial "population", where the mean weight, μ_{weight}, is 67.91 kg. Let 67.91 kg be μ_0 and test $H_0 : \mu = 67.91$ kg versus $H_A : \mu \neq 67.91$ kg. A simulation can be run in which a hypothesis test is performed on each sample drawn from the population; since in this setting, it is known that μ truly is 67.91 kg, each instance of rejecting H_0 represents an instance of making a Type I error.

6. Run the code chunk shown in the template to take 1,000 random samples of size 100 from `yrbss.complete`. The code calculates the t-statistics from each sample and returns `TRUE` if the t-statistic from a sample is within the rejection region.

```
#set parameters
sample.size = 100
replicates = 1000
alpha = 0.05

#set seed
set.seed(2017)

#create empty lists
t.stat = vector("numeric", replicates)

#calculate t-statistics from each sample
for (k in 1:replicates) {
  sample.rows = sample(nrow(yrbss.complete), sample.size)
```

```r
  #define constants
  mu = mean(yrbss.complete$weight)
  sample.mean = mean(yrbss.complete$weight[sample.rows])
  sample.sd = sd(yrbss.complete$weight[sample.rows])

  #calculate t statistics
  t.stat[k] = (sample.mean - mu) / (sample.sd / sqrt(sample.size))
}

#define upper and lower bounds of rejection region
reject.ub = qt(1 - (alpha)/2, df = sample.size - 1)
reject.lb = qt(alpha/2, df = sample.size - 1)

#is the t-stat in the rejection region?
in.rejection.region = (t.stat >= reject.ub) | (t.stat <= reject.lb)

table(in.rejection.region)
```

```
## in.rejection.region
## FALSE  TRUE
##   947    53
```

a) With $\alpha = 0.05$, what percentage of samples result in the (incorrect) conclusion that the population mean is not 67.91 kg?

53/1000 = 5.3% of the samples result in the incorrect conclusion that the population mean is not 67.91 kg.

b) What happens when α is changed? Test $\alpha = 0.10$ and $\alpha = 0.01$. Compare your results to those in part a).

When $\alpha = 0.10$, 10.2% of the samples result in rejecting H_0. When $\alpha = 0.01$, 1.1% of the samples result in rejecting H_0.[1] As α decreases, the percentage of samples that return an incorrect conclusion decreases.

This is directly related to what was observed in the confidence interval simulation from Lab 2 of this unit: α is the percentage of confidence intervals that did not contain the population mean μ_{weight}. Thus, α can be thought of as the error probability for both hypothesis tests and confidence intervals.

```
## in.rejection.region
## FALSE  TRUE
##   898   102
```

```
## in.rejection.region
## FALSE  TRUE
##   989    11
```

[1] Note that the observed proportion 1.1% is different from 1% only due to random chance; with a larger number of simulations, the observed proportion will approach 1%.

c) Why not eliminate the chance of Type I error completely by letting $\alpha = 0$? What implications would this have for hypothesis testing?

If α were equal to 0, then the null hypothesis would never be rejected, no matter how far the sample mean was from the hypothesized population mean μ_0. Hypothesis testing simply wouldn't work – there would be no "testing" happening!

It is necessary to balance the significance level such that the chance of a Type I error occurring is relatively rare, but α remains large enough to prevent the null hypothesis from almost never being rejected.