Statistical_Inference

Laura Bennett

July 9, 2017

Statistical Inference for Data Science: Chapter 8 Exercises

1. For iid Guassian data, the statistic $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ must follow what type of distribution?

- **t distribution**

2. Paired differences T confidence intervals are useful when:

- **Pairs of observations are linked, such as when there is subject level matching or in a study with baseline and follow up measurements on all participants.**

3. The assumption that the variances are equal for the independent group T interval means that:

- **The population variances are identical, but the sample variances may be different.**

4. Load the data set mtcars in the datasets R package. Calculate a 95% confidence interval to the nearest MPG for the variable mpg.

-

```
r   library(datasets)   data(mtcars)   t.test(mtcars[['mpg']])$conf.int
```

```
## [1] 17.91768 22.26357   ## attr(,"conf.level")   ## [1] 0.95
```

5. Suppose that standard deviation of 9 paired differences is 1. What value would the average difference have to be so that the lower endpoint of a 95% students t confidence interval touches zero?

- **Recall that the confidence interval for a t-distribution may be caluclated as $\bar{X} \pm t_{n-1}\frac{S}{\sqrt{n}}$. In this case, $S = 1$, $\sqrt{n} = \sqrt{9} = 3$, and $t_{n-1} \to t_{0.975}$. This particular t-quartile is chosen because we are interested in the lower endpoint of a 95% confidence interval; 2.5% of the density of the pdf will be on either side of the bell curve to make up the interval - thus, 97.5% of the area will be above the lower endpoint. In order for the lower endpoint to touch zero, it needs to be the case that $\bar{X} - t_{0.975,8}\frac{1}{\sqrt{9}} = 0$, and therefore it must be that $\bar{X} = t_{0.975,8}\frac{1}{\sqrt{9}}$ ($\bar{X}$ being our average difference). We can then calculate the average difference needed to be:**

```
r   qt(0.975,8)/3
```

```
## [1] 0.768668
```

6. An independent group Student's T interval is used instead of a paired T interval when:

- **The observations between the groups are naturally assumed to be statistically independent.**

7. Consider the mtcars dataset. Construct a 95% T interval for MPG comparing 4 to 7 cylinder cars (subtracting in the order of 4-6). Assume a constant variance.

-

“‘r mpg4<-mtcars$mpg[mtcars$cyl==4] mpg6<-mtcars$mpg[mtcars$cyl==6]

```
 t.test(mpg4,mpg6,var.equal=TRUE)$conf.int
```

‘‘‘

```
## [1]  3.154286 10.687272    ## attr(,"conf.level")    ## [1] 0.95
```

8. If someone put a gun to your head and said “your confidence interval must contain what it’s estimating or I’ll pull the trigger,” what would be the smart thing to do?

- **Make the interval as wide as possible**

9. Refer back to comparing MPG for 4 versus 6 cylinders. What do you conclude?

- **The interval was above 0, suggesting that 4 cylinders gets better mpg than 6 cylinders (since we are subtracting 4 and 6, this suggests the average mpg of 4 is higher)**

10. Suppose that 18 obese subjects were randomized, 9 each, to a new diet pill and a placebo. Subjects’ BMI were measured at a baseline and again after having received the treatment or placebo for four weeks. The average differnece from follow-up to the baseline (followup-baseline) was 3kg/m2 for the treated group and 1 kg/m2 for the placebo group. The corresponding standard deviations of the differences was 1.5 kg/m2 for the treatment group and 1.8 kg/m2 for the placebo group. The study aims to answer whether the change in BMI over the four week period appear to differ between the treated and placebo groups. What is the pooled variance estimate? +**The pooled variance in the case of groups with equal n is simply the average of their individual variances. Therefore:**

```
r   (1.5^2+1.8^2)/2
```

```
## [1] 2.745
```