Self-study for Review Notes

Brad McNeney

2018-01-27

The answers won't be posted. Instead, visit us in the Workshop and show us what you tried as well as explain why you tried it. Then, we're happy to share the answers with you. You're also welcome to ask any of the other TAs in the Workshop.

Question 1

Indicate all that apply:

1. The mean of the distribution of annual income is less than the median.
2. The interquartile range of the distribution contains about half the observations.
3. Statistical inference is the process of drawing conclusions about an entire population based on the information in a sample taken from that population
4. There is less variability among sample means than there is among individual observations
5. In the CLT, the original observations need to come from a normal distribution in order for the distribution of their sample means to be approximately normal.
6. The amount of sampling variability among a set of sample means increases as the size of the samples increases
7. Consider testing $H_0 : \mu = 0$ vs. $H_a : \mu \neq 0$.
   a) Suppose that under H_0 the test statistic has a standard-normal distribution and that the observed value of the test statistic in the sample is -3. Then:
      i. the pvalue is $2 \times P(Z >= 3)$, where Z is a standard-normal random variable.
      ii. the pvalue is $2 \times P(Z <= -3)$, where Z is a standard-normal random variable.
   b) Suppose that under H_0 the test statistic has a t distribution with 19 degrees of freedom and that its observed value in the sample is -3. Then:
      i. the pvalue is $2 \times P(T >= -3)$, where the random variable T has a t distribution with 19 degrees of freedom.
      ii. the pvalue is $1 - P(-3 <= T <= 3)$, where the random variable T has a t distribution with 19 degrees of freedom.

Question 2 (Serum-zinc levels in young men)

From the output below, calculate the following CI's for the mean of serum-zinc levels in the population of young men aged 15-17 years.

(a) the 95% CI

(b) the 90% CI.

```
uu<-url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_18/Data/serzinc.txt")
serzinc <- read.table(uu,header=TRUE)
library(dplyr)
summarize(serzinc, mean=mean(zinc), sd=sd(zinc), n=n())
```

```
##       mean       sd   n
## 1 87.93723 16.00469 462
```

```
tstar1<-qt(p=(1-.95)/2, df=461, lower.tail=FALSE)
tstar1
```

```
## [1] 1.965123
```

```
tstar2<-qt(p=(1-.90)/2, df=461, lower.tail=FALSE)
tstar2
```

```
## [1] 1.648166
```

(c) Repeat the calculation of 95% and 90% CIs using the `t.test()` function in R.

Question 3 (Babies with congenital heart disease)

```
uu<-url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_18/Data/heart.txt")
heart <- read.table(uu,header=TRUE)
summarize(heart, mean=mean(mdi), sd=sd(mdi), n=n())
```

```
##       mean      sd   n
## 1 104.7063 15.6551 143
```

(a) With the output above, calculate the value of the t-statistic for testing the null hypothesis that the mean score for the mental development index (MDI) in the population of babies born with congenital heart disease is 100 versus the alternative hypothesis that it isn't 100.

(b) The value of the test statistic observed in part (a) leads to a p-value of 0.0013. Conduct the test of the hypotheses in part (a) at level $\alpha = 0.05$. What do you conclude?

(c) Repeat the test of part (b) using the `t.test()` function in R.

Question 4 (Low-birthweight babies)

```
uu<-url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_18/Data/lowbwt.txt")
lowbwt <- read.table(uu,header=TRUE)
lowbwt %>% mutate(gender=recode(sex,`0`="female", `1`="male")) %>%
  group_by(gender) %>%
  summarize(mean=mean(sbp), sd=sd(sbp), n=n())
```

```
## # A tibble: 2 x 4
##   gender  mean    sd     n
##   <chr>  <dbl> <dbl> <int>
## 1 female  46.5  11.1    56
## 2 male    47.9  11.8    44
```

(a) With the output above, calculate the value of the t-statistic for testing the null hypothesis that, among low-birthweight infants, the mean systolic blood pressure (sbp) for boys is equal to the mean for girls vs. the alternative hypothesis that these means are different.

(b) The value of the test statistic observed in part (a) leads to a p-value of 0.5480. Conduct the test of the hypotheses in part (a) at level $\alpha = 0.05$. What do you conclude?

(c) Repeat the test of part (b) using the `t.test()` function in R.