Discussion 7

Review – Inference for One Population Mean

point estimator: a natural point estimator for population mean is the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, with standard error $= s/\sqrt{n}$, where

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}.$$

sampling distribution of $\bar{X}$: By Central Limit Theorem, $\bar{X}$ is approximately normal when n is large, even if we do not know the underlying distribution of population.
The following results are based on the assumption that the underlying distribution of population is *normal.*
P% confidence interval for μ:

$$\bar{x} - t^* \frac{s}{\sqrt{n}} < \mu < \bar{x} + t^* \frac{s}{\sqrt{n}},$$

where t^* is the critical value such that $P(|T_{n-1}| < t^*) = P\%$. Here, T_{n-1} denotes a random variable having t-distribution with degrees of freedom = n-1.
In R, qt(p,df) gives the quantile q such that $P(T_{df} < q) = p$.
Hypothesis Test:

- choose proper H_0 and H_A.

$$H_0 : \mu = \mu_0 \qquad H_A : \mu \neq \mu_0 \text{ or } \mu < \mu_0 \text{ or } \mu > \mu_0.$$

- test statistic

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

In R, the function t.test() carries out the t-test.
What if the underlying distribution is not normal? Use bootstrap. The general idea of bootstrap is as follows.

- Suppose we already obtained one sample of size n from the population, from which we already calculated sample mean $\bar{x}$.
- We use the existing sample as a proxy for the unknown population, take many samples of size n from the sample we already obtain, **with replacement**, and compute the sample mean of each.
- The middle 95% of this distribution is an approximate 95% confidence interval for μ.

Practice Problems

Example 1. Six healthy three-year-old female Suffolk sheep were injected with the antibiotic Gentamicin, at a dosage of 10 mg/kg body weight. Their blood serum concentration of Gentamicin 1.5 hours after injection were as follows:

33 26 34 31 23 25

(a) Construct a 90% confidence interval for the population mean.
(b) Interpret your result of (a)
(c) The interval in (a) nearly contains all of the observations; will this typically be true for a 90% confidence interval? Explain.
(d) Suppose we know that the s.d for the sheep's blood serum concentrations of Gentamicin 1.5 hours after injection is 5. How many samples do we need in order to guarantee that the error of estimation does not exceed 2?

Example 2. Madison Department of Transportation needs to analyze the average speed of cars passing downtown area, in order to see whether more traffic lights are needed for safety. Suppose, it is believed that an average speed of 15mph (or 24.14 kmph:kilometers per hour) is ideal for the balance of safety and smooth traffic. The velocity-detecting radar records 10 cars speed as shown in the following table. Use a two-sided test to analyze this dataset and represent your conclusions. (choose $\alpha = .05$)

Cars	1	2	3	4	5	6	7	8	9	10	$\bar{X}$	S
Speed (mph)	13.31	13.59	16.22	19.13	17.42	22.25	18.30	22.32	17.34	16.97	17.685	3.047
Speed (kmph)	21.42	21.87	26.10	30.78	28.03	35.80	29.44	35.91	27.90	27.30	28.455	4.9

Example 3. Refer to the salmon data in lecture, use the bootstrap method to find a 95% confidence interval for population mean and compare it to the t-test.

Solutions

Example 1. (a) $\bar{x} = 28.667$, $s = 4.59$, $n = 6$, $t^* = 2.015$. Then, the 90% confidence interval is (24.991, 32.442).
(b) Interpretation: We are 90% sure that the true mean blood serum concentration of Gentamicin is between 24.991 and 32.442.
(c) In general, this is not true for a 90% confidence interval. The meaning of a 90% confidence interval is not to say that such intervals will cover 90% data points in a sample.
(d) If the margin of error equals 2, we have

$$t^*_{n-1} \frac{s}{\sqrt{n}} \leq 2.$$

Hence,

$$\sqrt{n} \geq \frac{t^*_{n-1} s}{2}.$$

Note that, in the above formula, the quantile t^*_{n-1} also depends on n. Since there is no closed form for the quantile from t-distribution. We use the following R program to figure out the lower bound for n.

```
> s=5
> for (n in 2:50){
+   q=qt(0.95,n-1)
+   if(sqrt(n) > q*s/2) break;
+ }
> n
[1] 19
```

Therefore, $n = 19$ is already enough to reach a margin of error less than 2.

Example 2.

$$H_0 : \mu = 15, \qquad H_A : \mu \neq 15.$$

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{17.685 - 15}{3.047/\sqrt{10}} = 2.787.$$

Then, the p-value equals

$$2 \times P(T_9 > 2.287) = 0.021 < \alpha.$$

Hence, we reject H_0.

Example 3. Refer to Section 8 of R handout 3.