MATH 1070 - Chapter 20 Notes

Inference about a Population Mean: t Procedures

Recall: So far we have dealt with inference in situations when the following assumptions applied:

(1) Simple Random Sample (SRS)
(2) Normality of population or adequate sample size.
(3) Population standard deviation or σ known.

As discussed before, the population value of σ is never really known.

It may only be reasonable to assume the value in some situations.

What if we cannot make such an assumption about a population? ____________________.

The t distribution is a useful tool in inference procedures in such situations.

Recall z inference:

(1) SRS, Normality of population or large sample size, assume σ to be known.
(2) Derive sampling distribution of mean as: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.
(3) Do inference using sampling distribution of mean by converting to

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

(Note: μ is replaced by value under the null hypothesis: μ_0 in Hypothesis testing.)

At what stage do we get stuck, if we do not have an assumed value of σ?

1. Estimating σ.

If population standard deviation is unknown, then it is estimated by:____, also referred to as the ____________________________ which is given by the formula:

When we estimate σ with s, our one-sample z statistic becomes a one-sample t statistic.

The t test statistic follows a t distribution with $n-1$ degrees of freedom. In other words, the only parameter of the t distribution is the sample size minus one which in this case is referred to as the degrees of freedom.

Recall the properties of the z distribution:

(1) Parameters:

(2)

(3)

What is the relationship between the two? Consider the figure below:

The characteristics of the t distribution:

(1) Parameter:

(2)

(3)

(4)

2. One-Sample t Confidence Interval

Suppose we have an SRS of size n from a population with unknown mean μ and unknown standard deviation σ. A level C confidence interval for μ is:

where t^* is the critical value for confidence level C from the t density curve with $n - 1$ degrees of freedom. **The critical values of the t distribution are given in Table C (page 679) of your text book.**

Note: This interval is:

- Exact when the population distribution is Normal.
- Approximate for large n if population is not necessarily Normal.

Example (American Adult Heights): A study of 7 American adults from an SRS yields an average height of $\bar{x} = 67.2$ inches and a standard deviation of $s = 3.9$ inches. Find a 95% confidence interval for the average height of all American adults assuming that heights are normally distributed. (State all assumptions and interpret.)

3. One-Sample t test

Like the confidence interval, the t test is close in form to the z test learned earlier. When estimating σ with s, the test statistic becomes:

where t follows the t density curve with $n - 1$ degrees of freedom, and the p-value of t is determined from that curve.

Note: As with the confidence interval, the p-value is:

- Exact when the population distribution is Normal.
- Approximate for large n if population is not necessarily Normal.

3.1. p-value for Testing Means.

As you would expect the p-value of the test depends on the specific alternative hypothesis:

$H_A : \mu > \mu_0$:

p-value is the probability of getting a value as large or larger than the observed test statistic (t) value.

$H_A : \mu < \mu_0$:

p-value is the probability of getting a value as small or smaller than the observed test statistic (t) value.

$H_A : \mu \neq \mu_0$:

p-value is two times the probability of getting a value as large or larger than the absolute value of the observed test statistic (t) value.

Example (Sweetening Colas): Cola makers test new recipes for loss of sweetness during storage. Trained tasters rate the sweetness before and after storage. Here are the sweetness losses (sweetness before storage minus sweetness after storage) found by 10 randomly selected tasters for a new cola recipe:

2.0 0.4 0.7 2.0 -0.4 2.2 -1.3 1.2 1.1 2.3

Are these data good evidence (at the $\alpha = 0.05$ significance level) that the cola lost sweetness during storage?

What are the assumptions here? Do we need to make any more? Test the implied hypothesis and interpret your solution.

Note: The t may still be used to get approximate intervals and to do tests in moderate sample sizes where the data doesn't necessarily look Normal. See discussion on robustness of t.

4. Matched Pairs t Procedures

Goal: To compare the effect of two treatments on what is being measured or to compare the characteristics of two populations.

Subjects are matched in pairs and each treatment is given to one subject in each pair.

Before-and-after observations on the same subjects also calls for using matched pairs.

To compare the responses to the two treatments in a matched pairs design, apply the one-sample t procedures to the observed differences (one treatment observation minus the other).

The parameter μ is the mean difference in the responses to the two treatments within matched pairs of subjects in the entire population.

Example (Air Pollution): Pollution index measurements were recorded for two areas of a city on each of 8 days.

Area A	**Area B**	**A-B**
2.92	1.84	
1.88	0.95	
5.35	4.26	
3.81	3.18	
4.69	3.44	
4.86	3.69	
5.81	4.95	
5.55	4.47	

Question: Are the average pollution levels the same for the two areas of the city? Test the implied hypothesis, assumptions if any and interpret the solutions.

So far, we have not made any assumptions about Normality of the data. Let us make a stem and leaf plot.

While we cannot judge Normality from just 8 observations, the data shows no outliers, clusters, or extreme skewness. It turns out, p-values for the t test will be reasonably accurate and may be used.

5. Robustness of t procedures

- The t confidence interval and test are exactly correct when the distribution of the population is exactly normal.
  - In reality exact normalcy is not often satisfied. Some data resemble the normal in crucial ways more than others do.
  - The usefulness of the t procedures in practice therefore depends on how strongly they are affected by lack of Normality.
- A confidence interval or significance test is called **robust** if the confidence level or p-value does not change very much when the conditions for use of the procedure are violated.

5.1. Thumb-rules for using the t procedure:

(1) Except in the case of small samples, the assumption that the data are from an SRS from the population of interest is more important than the assumption that the population distribution is Normal.
(2) $n < 15$: Use t procedures if the data appear close to Normal (symmetric, single peak, no outliers). If the data are skewed or if outliers are present, do not use t.
(3) $n \geq 15$: The t procedures can be used except in the presence of outliers or strong skewness in the data.
(4) $n \geq 40$: The t procedures can be used even for clearly skewed distributions when the sample is large.

Can we use the t in the following examples?

Example: Consider the following stem plot of the force required to pull apart 20 pieces of Douglas fir.

23	0
24	0
25	
26	5
27	
28	7
29	
30	259
31	399
32	033677
33	0236

Example: This histogram shows the distribution of **all** word lengths in Shakespeare's plays.

Example: This histogram shows the heights of college students.

Chapter 20 Objectives:

- Assumptions for a t .
- Differences in assumptions between z and t based inferences.
- Properties of t distribution.
- Relationship between t and z distribution.
- t based confidence intervals and hypothesis tests.
- Robustness of t distribution and thumb-rules for using t distribution.