STAT509: Inference on Population Mean

Peijie Hou

University of South Carolina

October 28, 2014

Sampling distribution of $\overline{Y}$

- In this chapter, we will focus on the inference on population mean μ.
- Recall sample mean $\overline{Y}$ is a reasonable point estimator of the population mean μ.
- *RESULT:* Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. Then the sample mean $\overline{Y}$ has the following **sampling distribution**:

$$\overline{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

- The above result reminds us
  - $\overline{Y}$ is an **unbiased** estimator of μ.
  - $\mathrm{se}(\overline{Y}) = \sigma/\sqrt{n}$

Example

- Let Y =time (in seconds) to react to brake lights during in-traffic driving.
- We assume $Y \sim \mathcal{N}(\mu = 1.5, \sigma^2 = 0.16)$. We call this the **population distribution.**
- Suppose that we take a random sample of $n = 5$ drivers with times $Y_1, \ldots, Y_5$. What is the distribution of the sample mean $\overline{Y}$?

t distribution

- If we standardized $\overline{Y}$, we obtain

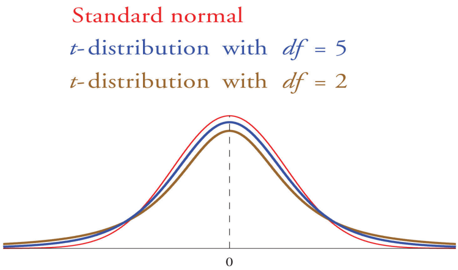
$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

- However, population standard deviation σ is usually unknown. Replacing it with the sample standard deviation S, we get a new sampling distribution:

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

  a t **distribution** with degrees of freedom $\nu = n - 1$.

t distribution



The t distribution has the following characteristics:

- It is continuous and symmetric about 0.
- It is indexed by a value ν called the degrees of freedom.
- As $\nu \longrightarrow \infty$, $t(\nu) \longrightarrow \mathcal{N}(0,1)$.
- When compared to the standard normal distribution, the t distribution, in general, is less peaked and has more probability (area) in the tails.

Example

Hollow pipes are to be used in an electrical wiring project. In testing 1-inch pipes, the data below were collected by a design engineer. The data are measurements of Y, the outside diameter of this type of pipe (measured in inches). These $n = 25$ pipes were randomly selected and measured-all in the same location.

1.296	1.320	1.311	1.298	1.315
1.305	1.278	1.294	1.311	1.290
1.284	1.287	1.289	1.292	1.301
1.298	1.287	1.302	1.304	1.301
1.313	1.315	1.306	1.289	1.291

Under this assumption (which may or may not be true), calculate the value of

$$t = \frac{\overline{Y} - \mu}{s/\sqrt{n}}$$

Example cont'd

We use R to find the sample mean $\overline{y}$ and the sample standard deviation s:

```
> mean(pipes) ## sample mean
[1] 1.29908
> sd(pipes) ## sample standard deviation
[1] 0.01108272
```

With $n = 25$, we have

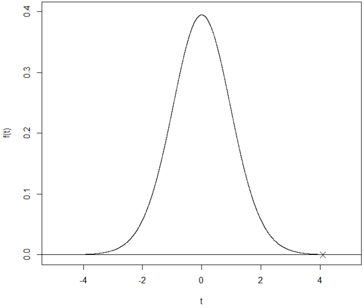
$$t =$$

Example cont'd

If the manufacturers claim is true (that is, if $\mu = 1.29$ inches), then

$$t = \frac{\overline{y} - \mu}{s/\sqrt{n}}$$

comes from a $t(24)$ distribution. The $t(24)$ pdf is displayed below:



Example cont'd

Key question: Does $t = 4.096$ seem like a value you would expect to see from this distribution? If not, what might this suggest? Recall that t was computed under the assumption that $\mu = 1.29$ inches (the manufacturers claim).

QUESTION: The value $t = 4.096$ is what percentile of the $t(24)$ distribution?

```
> pt(4.096,24)
[1] 0.9997934
```

Normal quantile-quantile (qq) plots

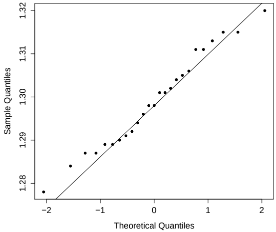
- Recall if $Y_1, \ldots, Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, then

$$t = \frac{\overline{Y} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

- An obvious question arises: "What if $Y_1, \ldots, Y_n$ are non-normal?"
- Answer: The t distribution result still approximately holds. That is, the t distribution is **robust** to the normality assumption.
- How to assess the normal distribution assumption? **Normal quantile-quantile (qq) plot**

Normal quantile-quantile (qq) plots cont'd

- The plot is constructed as follows:
  - On the vertical axis, we plot the observed data, ordered from low to high.
  - On the horizontal axis, we plot the (ordered) theoretical quantiles from the distribution assumed for the observed data (here, normal).
- For the pipe diameter data, below is the qq plot.




```
> qqnorm(pipes,pch=16,main="")
> qqline(pipes)
```

Normal quantile-quantile (qq) plots cont'd

- The ordered data do not match up perfectly with the normal quantiles, but the plot doesnt set off any serious alarms.
- *Fat pencil test:* Imagine a "fat pencil" lying along the line. If all the points are covered by this imaginary pencil, a normal distribution adequately describes the data.
- If there is no gross departure from the straight line on the plot, we should accept that the model describes the data well.

CI for μ. Assume normality and **KNOWN** σ^2

- Recall if $Y_1, \ldots, Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution and σ^2 is **known**, then

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

- Similar to CI derivation in population proportion, a $100(1-\alpha)\%$ CI for μ is given by

$$\left(\overline{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \overline{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Example: pipe diameter data

Go back to pipe diameter data. Based on past experience, the engineers assume a normal population distribution (for the pipe diameters) with known population standard deviation 0.02. We want to find a 95% CI for μ, the mean pipe diameter.

Solution: We have the sample mean $\overline{Y} = 1.30$. So a 95% C.I. interval is

$$
\begin{aligned}
& \left(\overline{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \overline{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\
= \quad & (1.30 - 1.96 \times \frac{0.02}{\sqrt{25}}, \quad 1.30 + 1.96 \times \frac{0.02}{\sqrt{25}}) \\
= \quad & (1.292, \quad 1.308)
\end{aligned}
$$

Practical Interpretation: Based on the sample data, with 95% confidence, the outside diameters for the hollow pipe is between 1.29 and 1.31 inches.

CI for μ. Assume normality and **UNKNOWN** σ^2

- Recall if $Y_1, \ldots, Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution and σ^2 is **unknown**, then

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

- A $100(1-\alpha)\%$ CI for μ is given by

$$\left(\overline{y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \quad \overline{y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

where S is the sample standard deviation.

Example: pipe diameter data

Go back to pipe diameter data. Assume the pipe diameters are normality distributed with unknown population variance. We want to find a 95% CI for μ, the mean pipe diameter.

Solution: We have the sample mean $\overline{Y} = 1.30$ and sample standard deviation $S = 0.011$ So a 95% C.I. interval is

$$\left(\overline{y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \quad \overline{y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

$$= \quad (1.30 - 2.064 \times \frac{0.011}{\sqrt{25}}, \quad 1.30 + 2.064 \times \frac{0.011}{\sqrt{25}})$$

$$= \quad (1.295, \quad 1.305)$$

Practical Interpretation: Based on the sample data, with 95% confidence, the outside diameters for the hollow pipe is between 1.295 and 1.305 inches.

Sample size determination

- In a planning stages of an experiment or investigation, it is often of interest to determine **how many individuals** are needed to write a confidence interval with a given level of precision.
- For example, we might want to construct a 95 percent confidence interval for a population mean so that the **interval length** is no more than 5 units (e.g., days, inches, dollars, etc.).
- **Sample size determination** is also associated with the practical issues like cost, time spent in data collection, personnel training, etc.

Sample size determination cont'd

- Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population with σ^2 known. Recall that a $100(1-\alpha)\%$ CI for μ is given by

$$\overline{Y} \pm \underbrace{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{=B,\ say}$$

- B is called the **margin of error**.

Sample size determination cont'd

- It is possible to determine the sample size n once we specify these three pieces of information:
  - the value of σ^2 (or an educated guess at its value; e.g., from past information, etc.)
  - the confidence level, $100(1-\alpha)$
  - the margin of error, B.
- This is true because

$$B = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \Longleftrightarrow n = \left(\frac{\sigma z_{\alpha/2}}{B}\right)^2$$

Sample size determination example

In a biomedical experiment, we would like to estimate the population mean remaining life of healthy rats that are given a certain dose of a toxic substance. Suppose that we would like to write a 95 percent confidence interval for with a margin of error equal to B = 2 days. From past studies, remaining rat lifetimes have been approximated by a normal distribution with standard deviation s = 8 days. How many rats should we use for the experiment?

Solution: With $z_{\alpha/2} = z_{0.025} \approx 1.96$, $B = 2$ and $\sigma = 8$, the desired sample size to estimate μ is

$$n = \left(\frac{\sigma z_{\alpha/2}}{B}\right)^2 = \left(\frac{1.96 \times 8}{2}\right)^2 \approx 61.46$$

We would sample $n = 62$ rats to achieve these goals.

Hypothesis Test on population mean

1. State the null (H_0) and alternative (H_a) hypotheses.
   - Null hypothesis $H_0 : \mu = \mu_0$
   - Alternative hypothesis
     - Right-tail $H_a : \mu > \mu_0$
     - Left-tail $H_a : \mu < \mu_0$
     - Two-tail $H_a : \mu \neq \mu_0$
2. Collect the data and calculate test statistic assuming H_0 is true.

$$\sigma \text{ known: } Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}} \quad \text{OR} \quad \sigma \text{ unknown: } t = \frac{\overline{Y} - \mu_0}{s/\sqrt{n}}$$

3. Assuming the null hypothesis is true, calculate the p-value.

Alternative	Type	p-value (based on the knowledge of σ)
$H_a : \mu > \mu_0$	Right-tail	$P(Z > z_0)$ or $P(t > t_0)$
$H_a : \mu < \mu_0$	Left-tail	$P(Z < z_0)$ or $P(t < t_0)$
$H_a : \mu \neq \mu_0$	Two-tail	$2P(Z < -\lvert z_0\rvert)$ or $2P(t < -\lvert t_0\rvert)$

4. Draw conclusion based on the p-value. We either reject H_0 or fail to reject H_0.

Hypothesis test example: known σ^2

Go back to pipe diameter data. Based on past experience, the engineers assume a normal population distribution (for the pipe diameters) with known population standard deviation 0.02. Researchers what to find out whether the pipe diameter is 1.31. Assume a significance level 0.05.

Hypothesis test example: known σ^2

Solution:

- **Step 1: State hypothesis**

$$H_0 : \mu = 1.31$$
$$H_a : \mu \neq 1.31$$

- **Step 2: Test statistic**

$$z_0 = \frac{1.30 - 1.31}{0.02/\sqrt{25}} = -2.5$$

- **Step 3: p-value**

$$p - value = 2P(Z < -|-2.5|) = 2P(Z < -2.5) = 0.012$$

- **Step 3: Conclusion**
  p-value=0.012$< \alpha$, we reject H_0. We have enough evidence to conclude the mean pipe diameter is not 1.31 inches.

Hypothesis test example: unknown σ^2

Go back to pipe diameter data. Assume the pipe diameters are normality distributed with unknown population variance. Researchers what to find out whether the pipe diameter is less than 1.308. Assume a significance level 0.05.

Hypothesis test example: known σ^2

Solution:

- **Step 1: State hypothesis**
- **Step 2: Test statistic**
- **Step 3: *p*-value**
- **Step 3: Conclusion**

R code to produce t-test

```
> t.test(pipes,alternative="less", mu=1.308)

        One Sample t-test

data:  pipes
t = -4.0243, df = 24, p-value = 0.0002478
alternative hypothesis: true mean is less than 1.308
95 percent confidence interval:
     -Inf 1.302872
sample estimates:
mean of x
  1.29908
```