STA286 Lecture 26

Neil Montgomery

Last edited: 2017-03-23 12:24

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The "obvious" estimator is $\overline{X}_1 - \overline{X}_2$, which will have a normal distribution with:

$$E\left(\overline{X}_1 - \overline{X}_2\right)$$

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The "obvious" estimator is $\overline{X}_1 - \overline{X}_2$, which will have a normal distribution with:

$$E\left(\overline{X}_1 - \overline{X}_2\right) = \mu_1 - \mu_2$$

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The "obvious" estimator is $\overline{X}_1 - \overline{X}_2$, which will have a normal distribution with:

$$E\left(\overline{X}_1 - \overline{X}_2\right) = \mu_1 - \mu_2$$

$$\mathrm{Var}\left(\overline{X}_1 - \overline{X}_2\right)$$

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The "obvious" estimator is $\overline{X}_1 - \overline{X}_2$, which will have a normal distribution with:

$$
\begin{aligned}
E\left(\overline{X}_1 - \overline{X}_2\right) &= \mu_1 - \mu_2 \\
\mathrm{Var}\left(\overline{X}_1 - \overline{X}_2\right) &= \mathrm{Var}\left(\overline{X}_1\right) + \mathrm{Var}\left(\overline{X}_2\right) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}
\end{aligned}
$$

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The "obvious" estimator is $\overline{X}_1 - \overline{X}_2$, which will have a normal distribution with:

$$\begin{aligned} E\left(\overline{X}_1 - \overline{X}_2\right) &= \mu_1 - \mu_2 \\ \mathrm{Var}\left(\overline{X}_1 - \overline{X}_2\right) &= \mathrm{Var}\left(\overline{X}_1\right) + \mathrm{Var}\left(\overline{X}_2\right) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \end{aligned}$$

We need to figure out what to do about σ^2.

for σ^2, use the data from both samples

The sample variances S_1^2 and S_2^2 are both unbiased estimators for σ^2, so any weighted average (with weights that add up to 1) of them will also be an unbiased estimator.

for σ^2, use the data from both samples

The sample variances S_1^2 and S_2^2 are both unbiased estimators for σ^2, so any weighted average (with weights that add up to 1) of them will also be an unbiased estimator.

We call this the pooled variance estimator:

$$S_p^2 = \frac{(n_1 - 1)S_1^2}{(n_1 - 1) + (n_2 - 1)} + \frac{(n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

for σ^2, use the data from both samples

The sample variances S_1^2 and S_2^2 are both unbiased estimators for σ^2, so any weighted average (with weights that add up to 1) of them will also be an unbiased estimator.

We call this the pooled variance estimator:

$$S_p^2 = \frac{(n_1 - 1)S_1^2}{(n_1 - 1) + (n_2 - 1)} + \frac{(n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Note that when $n_1 = n_2$ this is just the average of the two sample variances.

putting it all together

We want an interval estimator for $\mu_1 - \mu_2$. So far we have:

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

putting it all together

We want an interval estimator for $\mu_1 - \mu_2$. So far we have:

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

Who wants to guess what the distribution of this will be:

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim$$

putting it all together

We want an interval estimator for $\mu_1 - \mu_2$. So far we have:

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

Who wants to guess what the distribution of this will be:

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t$$

putting it all together

We want an interval estimator for $\mu_1 - \mu_2$. So far we have:

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

Who wants to guess what the distribution of this will be:

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Isolating $\mu_1 - \mu_2$ in the usual way gives the confidence interval formula.

two normal samples, equal variances C.I.

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm t_{n_1+n_2-2,\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

In the 95% case, another instance of my patented:

$$\text{estimator} \pm \text{“2”s.e.(estimator)}$$

example - watching two kinds of paint dry

If the world of one brand of paint drying was too fast-paced, this example is for you. (Question 9.49 from the textbook.) Two brands of paint will have their drying times compared.

example - watching two kinds of paint dry

If the world of one brand of paint drying was too fast-paced, this example is for you. (Question 9.49 from the textbook.) Two brands of paint will have their drying times compared.

The goal is to estimate the difference between the mean drying times.

example - watching two kinds of paint dry

If the world of one brand of paint drying was too fast-paced, this example is for you. (Question 9.49 from the textbook.) Two brands of paint will have their drying times compared.

The goal is to estimate the difference between the mean drying times.

Here's a glance at the "dataset" as is, organized in a way it should never be collected:

```
## # A tibble: 15 × 2
##    PaintA PaintB
##     <dbl>  <dbl>
## 1     3.5    4.7
## 2     2.7    3.9
## 3     3.9    4.5
## 4     4.2    5.5
## 5     3.6    4.0
## 6     2.7    5.3
## 7     3.3    4.3
```

watching two kinds of paint dry

A real dataset looks like this:

```
## # A tibble: 30 × 2
##     brand  time
##     <chr> <dbl>
## 1  PaintA   3.5
## 2  PaintA   2.7
## 3  PaintA   3.9
## 4  PaintA   4.2
## 5  PaintA   3.6
## 6  PaintA   2.7
## 7  PaintA   3.3
## 8  PaintA   5.2
## 9  PaintA   4.2
## 10 PaintA   2.9
## # ... with 20 more rows
```

watching two kinds of paint dry

Anyway, here is a summary of the two groups:

brand	x_bar	samp_var	n
PaintA	3.82	0.607	15
PaintB	4.94	0.568	15

The degrees of freedom is 28. The number from the t distribution is $t_{28,0.025} = 2.048$.

The 95% confidence interval is $[-1.693, -0.547]$.

when the variances cannot be assumed to be equal

Most two-sample analyses in practice don't bother with the equal variance assumption, and just use the following sequence of facts.

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

when the variances cannot be assumed to be equal

Most two-sample analyses in practice don't bother with the equal variance assumption, and just use the following sequence of facts.

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim^{approx} t_\nu$$

where ν has one of the most disgraceful formulae in the history of formulae. Don't look at its formula on the next slide.

what did I just tell you?

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$$

This formula is not for humans to use.

But there are few things to notice about it:

- if $S_1^2 \approx S_2^2$, then $\nu \approx n_1 + n_2 - 2$.

what did I just tell you?

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$$

This formula is not for humans to use.

But there are few things to notice about it:

- if $S_1^2 \approx S_2^2$, then $\nu \approx n_1 + n_2 - 2$.
- if $S_1^2 \ll S_2^2$, then $\nu \approx n_2 - 1$, and vice versa.

what did I just tell you?

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$$

This formula is not for humans to use.

But there are few things to notice about it:

- if $S_1^2 \approx S_2^2$, then $\nu \approx n_1 + n_2 - 2$.
- if $S_1^2 \ll S_2^2$, then $\nu \approx n_2 - 1$, and vice versa.

what did I just tell you?

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$$

This formula is not for humans to use.

But there are few things to notice about it:

- if $S_1^2 \approx S_2^2$, then $\nu \approx n_1 + n_2 - 2$.
- if $S_1^2 \ll S_2^2$, then $\nu \approx n_2 - 1$, and vice versa.

It won't usually be an integer, so if you need to use this method on a test (where I'd give you the value of ν), just use whatever nearby integer that is convenient.

watching two kinds of paint dry, now in a very slightly different way

The C.I. formula becomes:

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm t_{/nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

which in the 95% case is another instance of my patented formula.

The paint drying example, redux:

brand	x_bar	samp_var	n
PaintA	3.82	0.607	15
PaintB	4.94	0.568	15

The degrees of freedom is $\nu = 27.969$. The number from the t distribution is $t_{27.969,0.025} = 2.049$.

The 95% confidence interval is $[-1.694, -0.546]$ (as compared to $[-1.693, -0.547]$)

watching plants grow

Instead of watching paint dry, let's watch plants grow. (Textbook question 9.40.)

20 tree seeds are planted. 10 get a nitrogen fertilizer. After 140 days all the stem growths are measured in grams. The goal is to estimate the mean difference between the two groups.

Here is a summary of the data:

fertilizer	x_bar	samp_var	n
Nitrogen	0.565	0.035	10
NoNitrogen	0.399	0.005	10

The degrees of freedom is $\nu = 11.673$. The number from the t distribution is $t_{11.673,0.025} = 2.186$.

The 95% confidence interval is $[0.027, 0.305]$.