Two sample comparisons: Spread and Distribution

ST551 Lecture 26

Charlotte Wickham

2017-11-22

Announcements

Practice Final (from Fall 2016) on canvas

Final study guide posted on class webpage

Office hours for Charlotte (no change):

- **Week 10:** Mon Nov 27th 2-2:50pm, Wed Nov 29th 2-2:50pm
- **Final's week:** Mon Dec 4th 2-2:50pm, Wed Dec 6th 2-2:50pm

Comparisons of scale

Comparisons of variance

Y_i i.i.d sample of size *n* from population with mean μ_Y and variance σ^2_Y
X_i i.i.d sample of size *m* from population with mean μ_Y and variance σ^2_Y

Comparison of interest: Is $\sigma^2_Y = \sigma^2_X$? I.e. do the two populations have the same variance?

Recall from the one sample case

If the population is Normal,

The sample variance,

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

has a the sampling distribution that is a scaled Chi-square distribution:

$$(n-1)\frac{s_Y^2}{\sigma_Y^2} \sim \chi^2_{(n-1)}$$

Two sample analog

Since our samples are independent of each other, if both populations are Normal, under the null hypothesis $\sigma_Y^2 = \sigma_X^2 = \sigma$,

$$\frac{s_Y^2}{s_X^2} \sim \frac{Z/(n-1)}{V/(m-1)}$$

where $Z \sim \chi^2_{(n-1)}$ and $V \sim \chi^2_{m-1}$.

The ratio of two independent Chi-square random variables scaled by their d.f. is an F-distribution.

I.e. Under null

$$\frac{s_Y^2}{s_X^2} \sim F_{(n-1),(m-1)}$$

But if the two populations aren't Normal

The performance of this *F-test for variances* isn't great for even quite large samples:

- the test is far from exact for even quite large samples (it's not even asymptotically exact, only approximately exact)
- the test is consistent, but the non-exactness makes it hard to interpret

Like in the one sample case, there is a slightly better performing alternative.

Levene's test

1. Construct new variables (deviations from center): **Option 1**

$$U_i = |Y_i - \text{median}(Y)|$$
$$V_i = |X_i - \text{median}(X)|$$

2. Perform a two sample t-test (usually Welch's) to test the null hypothesis the mean of U is the same as the mean of V.

A few common variations

Option 1 Absolute deviations from median

$$U_i = |Y_i - \text{median}(Y)|$$

$$V_i = |X_i - \text{median}(X)|$$

Option 2 Squared deviations from median

$$U_i = (Y_i - \text{median}(Y))^2$$

$$V_i = (X_i - \text{median}(X))^2$$

Option 3 Absolute deviations from mean

$$U_i = |Y_i - \overline{Y}|$$

$$V_i = |X_i - \overline{X}|$$

Option 4 Squared deviations from mean

$$U_i = \left(Y_i - \overline{Y}\right)^2$$

$$V_i = \left(X_i - \overline{X}\right)^2$$

Levene's test Comments

The correct interpretation of the results of Levene's test depends on which version is used.

Only **Option 4** can be strictly interpreted as a question about the population variances.

In R: package `car` has a function `leveneTest()`. By default, uses option 1 (absolute differences from sample median), argument `center = mean` for option 3.

If you want squared deviations rather than absolute deviation, do it by hand.

Use it when the question of interest is about variance/spread. Not recommended for choosing which t-test to use.

Comparisons of distribution

Two sample K-S test setting

Setting: two independent samples

Y_i i.i.d sample of size n from population with c.d.f F_Y
X_i i.i.d sample of size m from population with c.d.f F_X

Comparison of interest: Is F_Y the same as F_X?

Two sample K-S test procedure

Null hypothesis: $H_0 : F_Y = F_X$

Test statistic

$$D = sup_y \left| \hat{F}_Y(y) - \hat{F}_X(y) \right|$$

where

$$\hat{F}_Y(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\} \quad \text{and} \quad \hat{F}_X(y) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{X_i \leq y\}$$

i.e. $DS is the largest vertical distance between the empirical cumulative density functions.

Reference Distribution

If the null is true,

$$\sqrt{\frac{mn}{m+n}} \to_d K$$

where K is the Kolmogorov distribution.

Reject for large values of D.

Your turn

Find the KS Test Statistic for the following two samples:

- **Y**: *1.5*, *2*, *2.3*, *2.4*, *2.9*, *5.9*, *6.1*, *6.2*, *6.5* and *9.4*
- **X**: *4*, *6*, *6.6*, *6.8*, *7*, *9.5*, *9.7* and *9.9*

Some critical values for the Kolmogorov distribution are given below. What is the result of your test procedure at level 0.05?

0.1	0.05	0.025	0.01	0.005	0.001
1.22	1.36	1.48	1.63	1.73	1.95

Discrete Distributions?

The KS test applies only to continuous distributions (that is, the underlying population distributions must be continuous).

What if we want to test that equality in the setting where the underlying population distributions are discrete?

We have already seen methods for doing this: Pearson's chi-squared test for $r \times c$ contingency tables.

Next week...

Delta method

Bootstrap

Randomization distribution / Permutation tests