Two sample inference

September 11, 2017

Suppose we have data from two populations.

$$X_1, X_2, \ldots, X_n \text{ iid } f(x; \theta_x)$$

and

$$Y_1, Y_2, \ldots, Y_m \text{ iid } f(y; \theta_y)$$

Interest is often in determining whether the samples are from the same population.

EDA

- Look at each sample individually:
  - check normality
  - plot histogram
  - look at sample mean and variance
- Look at the samples together
  - look at side by side boxplots
  - look at a qqplot(x,y) - if the two samples are from the sample population, this plot should be the 45 degree line

Inferences about the Mean

Assume the two samples are normal and independent of each other.

$$X_1, X_2, \ldots, X_n \ iid \ N(\mu_x, \sigma_x^2)$$

independent of

$$Y_1, Y_2, \ldots, Y_m \ iid \ N(\mu_y, \sigma_y^2)$$

Inferences about the mean

Equal variances: $\sigma_x^2 = \sigma_y^2 = \sigma^2$

If σ^2 is known:

$$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

and so

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1)$$

A $100(1-\alpha)\%$ confidence interval for the difference in the means, $\mu_x - \mu_y$ is

$$\bar{X} - \bar{Y} \pm z_{\alpha/s}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

To test $H_0 : \mu_x - \mu_y = \delta_0$,the test statistic below has a standard normal distribution under H_0

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Inferences about the mean, normal distribution, equal, unknown variance

Most often σ^2 is unknown and must be estimated. Let

$$S_x^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1), S_y^2 = \sum_{i=1}^{m}(Y_i - \bar{Y})^2/(m-1)$$

The pooled estimate of σ^2 is

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

Inferences about the mean, normal distribution, equal, unknown variance

Then

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_p\sqrt{1/n + 1/m}} \sim t_{n+m-2}$$

A $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$ is

$$\bar{X} - \bar{Y} \pm t_{n+m-2,\alpha/2} S\sqrt{1/n + 1/m}$$

Inferences about the mean, normal distribution, equal, unknown variance

Hypothesis testing

Test statistic under H_0

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{S\sqrt{1/n + 1/m}} \sim t_{n+m-2}$$

H_0	H_1	p-value
$\mu_x - \mu_y \leq \delta_0$	$\mu_x - \mu_y > \delta_0$	$P(t_{n+m-2} > t_{obs})$
$\mu_x - \mu_y \geq \delta_0$	$\mu_x - \mu_y < \delta_0$	$P(t_{n+m-2} < t_{obs})$
$\mu_x - \mu_y = \delta_0$	$\mu_x - \mu_y \neq \delta)0$	$2 * P(t_{n+m-2} > \lvert t_{obs} \rvert)$

Inferences about the mean, normal distribution, variances unequal and unknown

Case 1: n, m large

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S_x^2/n + S_y^2/m}} \approx N(0,1)$$

Case 2: n, m not large - Use Welch's method

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S_x^2/n + S_y^2/m}} \approx t_v$$

where $v = \dfrac{(w_1 + w_2)^2}{w_1^2/(n-1) + w_2^2/(m-1)}$ for $w_i = S_i^2/n_i$

Inferences about the mean, normal distribution, variances unequal and unknown

It is important that if $\sigma_x^2 >> \sigma_y^2$, $n >> m$. We need more information about the more variable population. In practice if $\sigma_x^2 >> \sigma_y^2$ and $n < m$ don't use this approach.

Matched Pair Design

The X and Y samples are not independent but are matched. so (X_i, Y_i) are measurements on the ith pair. (Sample sizes must be the same)
Assuming normality $X_i \sim N(\mu_x, \sigma_x^2)$ and $Y_i \sim N(\mu_y, \sigma_y^2)$ for $i = 1, \ldots, n$. Since the measurements are not independent, $Corr(X_i, Y_i) = \rho$.
Let $D_i = X_i - Y_i$ then $D_1, D_2, \ldots, D_n$ are iid $N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y)$.

Matched Pair Design

If X and Y are positively correlated, $\rho > 0$ and $var(D_i) < \sigma_x^2 + \sigma_y^2$ and the resulting inferences are more precise. A $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$ is

$$\bar{D} \pm t_{n-1,\alpha/2} \frac{S_d}{\sqrt{n}}$$

Matched Pair Design

To test $\mu_x - \mu_y = \delta_0$ we use a paired t-test. Under H_0

$$T = \frac{\bar{D} - \delta_0}{S_d/\sqrt{n}} \sim t_{n-1}$$

To test whether X and Y have equal means, $\delta_0 = 1$.

Non-normality

- Bootstrap Confidence Interval for $\mu_x - \mu_y$
- Randomization test
- Independent samples - Mann Whitney Test
- Paired samples - Wiloxon signed rank test as in single sample case

Randomization test

For two independent samples, to test hypotheses like $H_0 : E(X) = E(Y)$ versus $H_1 : E(X) \neq E(Y)$. If H_0 is true, we can consider the samples X and Y as a group of $n + m$ observations from a population with mean $E(X) = E(Y)$. There are $\binom{n+m}{n}$ ways of choosing a sample of size n from the $m + n$ observations. For each possible sample compute $T = \sum X_i$. Compare the observed T from the original sample with all the $\binom{m+n}{n}$ possible values of T to get the p-value. Count the number of T's more extreme than observed.

Mann Whitney Test

- To test $H_0 : \mu_x = \mu_y$
- Combine the sample into a single sample of size $n + m$
- Assign ranks form 1 to $n + m = N$ to the combined sample
- Let $R(X_i)$ be the rank assigned to X_i, $R(Y_i)$ be the rank assigned to Y_i
- Test statistic when there are no ties $T = \sum_{i=1}^{n} R(X_i)$.
- Reject H_0 if T is too big or too small.

Inferences on Variances

Normal Assumption

If we are interested in determining whether $\sigma_x^2 = \sigma_y^2$ we can use $\frac{S_x^2}{S_y^2}$ to estimate the ratio of the variances. If this ratio $= 1$, the two variances are equal.

$$\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim F_{n-1,m-1}$$

Inferences on Variances

Normal assumption

A $100(1-\alpha)\%$ confidence interval for σ_x^2/σ_y^2 is

$$\left(\frac{1}{f_{n-1,m-1,\alpha/2}} \frac{S_x^2}{S_y^2}, \frac{1}{f_{n-1,m-1,1-\alpha/2}} \frac{S_x^2}{X_y^2} \right)$$

where $f_{n-1,m-1,\alpha}$ is the value in an F distribution with $n-1$ and $m-1$ degrees of freedom such that the area to the right of that number is equal to α.

Inferences on Variances

Normal assumption

To test $H_0 : \sigma_x^2 = \sigma_y^2$ we can use the test statistic $T = S_x^2/S_y^2$ which has an F distribution with $n-1$ and $m-1$ degrees of freedom under the null hypothesis.

Null hypothesis	Alternate hypothesis	p-value
$\sigma_x^2 \geq \sigma_y^2$	$\sigma_x^2 < \sigma_y^2$	$f_{n-1,m-1} < t_{obs}$
$\sigma_x^2 \leq \sigma_y^2$	$\sigma_x^2 > \sigma_y^2$	$f_{n-1,m-1} > t_{obs}$
$\sigma_x^2 = \sigma_y^2$	$\sigma_x^2 \neq \sigma_y^2$	***

The two sided p-value in this case is found as 2*$P(f_{n-1,m-1} > t_u$ or 2*$P(f_{n-1,m-1} < t_l$ depending on where the observed T occurs.

This method can be used to test for equal variances to help decide what inferences to use on the mean.

Inferences on Variances

Non-normal populations

If the populations are not normal, do not use the F test as it can lead to misleading results

Instead you can use squared rank non-parametric test.