Serik Sagitov, Chalmers and GU, February 22, 2016

Chapter 11. Comparing two samples

Suppose we wish to compare two population distributions with means and standard deviations (μ_1, σ_1) and (μ_2, σ_2). Given two IID samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ from these two populations, we can compute two sample means and their standard errors

$\mathrm{E}\,\bar{X} = \mu_1, \quad \mathrm{Var}\,\bar{X} = \frac{\sigma_1^2}{n}, \quad s_{\bar{X}} = \frac{s_1}{\sqrt{n}}, \quad s_1^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2,$

$\mathrm{E}\,\bar{Y} = \mu_2, \quad \mathrm{Var}\,\bar{Y} = \frac{\sigma_2^2}{m}, \quad s_{\bar{Y}} = \frac{s_2}{\sqrt{m}}, \quad s_2^2 = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \bar{Y})^2.$

The difference $(\bar{X} - \bar{Y})$ is an unbiased estimate of $(\mu_1 - \mu_2)$. We are interested in finding the standard error of $\bar{X} - \bar{Y}$ and an interval estimate for $(\mu_1 - \mu_2)$, as well as testing the null hypothesis of equality H_0: $\mu_1 = \mu_2$.

Two main settings: independent samples and paired samples.

1 Two independent samples

If $(X_1, \ldots, X_n)$ is independent from $(Y_1, \ldots, Y_m)$, then $\mathrm{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$.

Therefore, $s_{\bar{X}-\bar{Y}}^2 = s_{\bar{X}}^2 + s_{\bar{Y}}^2 = \frac{s_1^2}{n} + \frac{s_2^2}{m}$ gives an unbiased estimate of $\mathrm{Var}(\bar{X} - \bar{Y})$.

Large sample test for the difference

If n and m are large, we can use a normal approximation $\bar{X} - \bar{Y} \overset{a}{\sim} \mathrm{N}(\mu_1 - \mu_2, s_{\bar{X}}^2 + s_{\bar{Y}}^2)$. The hypothesis H_0: $\mu_1 = \mu_2$ is tested using the test statistic $T = \frac{\bar{X}-\bar{Y}}{\sqrt{s_{\bar{X}}^2 + s_{\bar{Y}}^2}}$.

Approximate CI for $(\mu_1 - \mu_2)$ is given by $\bar{X} - \bar{Y} \pm z_{\alpha/2} \cdot \sqrt{s_{\bar{X}}^2 + s_{\bar{Y}}^2}$.

For the binomial model $X \sim \mathrm{Bin}(n, p_1)$, $Y \sim \mathrm{Bin}(m, p_2)$, the sample proportions $\hat{p}_1 = \frac{X}{n}, \hat{p}_2 = \frac{Y}{m}$ have standard errors $s_{\hat{p}_1}^2 = \frac{\hat{p}_1\hat{q}_1}{n-1}, s_{\hat{p}_2}^2 = \frac{\hat{p}_2\hat{q}_2}{m-1}$, then a 95 % CI for $(p_1 - p_2)$ is given by $\hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n-1} + \frac{\hat{p}_2\hat{q}_2}{m-1}}$.

Example (swedish polls)

Consider two consecutive poll results $\hat{p}_1$ and $\hat{p}_2$ with $n \approx m \approx 5000$ interviews. A change in support to Social Democrats at $\hat{p}_1 \approx 0.4$ is significant if

$$|\hat{p}_1 - \hat{p}_2| > 1.96 \cdot \sqrt{2 \cdot \frac{0.4 \cdot 0.6}{5000}} \approx 1.9\%.$$

This should be compared with the one-sample hypothesis testing $H_0 : p = 0.4$ vs $H_0 : p \neq 0.4$. The approximate 95% CI for p is $\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}\hat{q}}{n-1}}$ and if $\hat{p} \approx 0.4$, then the difference is significant if

$$|\hat{p} - p_0| > 1.96 \cdot \sqrt{\frac{0.4 \cdot 0.6}{5000}} \approx 1.3\%.$$

Two-sample t-test

The key assumption of the two-sample t-test:

two normal population distributions $X \sim \mathrm{N}(\mu_1, \sigma^2)$, $Y \sim \mathrm{N}(\mu_2, \sigma^2)$ have equal variances.

Given $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the pooled sample variance

$$s_p^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2}{n + m - 2} = \frac{n-1}{n+m-2} \cdot s_1^2 + \frac{m-1}{n+m-2} \cdot s_2^2$$

is an unbiased estimate of the variance with $\mathrm{E}(s_p^2) = \sigma^2$.

In view of $\mathrm{Var}(\bar{X} - \bar{Y}) = \sigma^2 \cdot \frac{n+m}{nm}$, we arrive at an alternative unbiased estimate $s_{\bar{X}-\bar{Y}}^2 = s_p^2 \cdot \frac{n+m}{nm}$ for the variance $\mathrm{Var}(\bar{X} - \bar{Y})$ of the sampling distribution.

Exact distribution $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{s_p} \cdot \sqrt{\frac{nm}{n+m}} \sim t_{n+m-2}$

Exact CI for $(\mu_1 - \mu_2)$ is given by $\bar{X} - \bar{Y} \pm t_{n+m-2}(\frac{\alpha}{2}) \cdot s_p \cdot \sqrt{\frac{n+m}{nm}}$.

Two sample t-test, equal population variances

For H_0: $\mu_1 = \mu_2$, the null distribution of $T = \frac{\bar{X}-\bar{Y}}{s_p} \cdot \sqrt{\frac{nm}{n+m}}$ is $T \sim t_{n+m-2}$.

Welch's t-test. If variances are not assumed to be equal so that $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$, then the t-test can be modified using the fact that $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{s_{\bar{X}}^2+s_{\bar{Y}}^2}}$ has an approximate t_{df}-distribution with

$$\mathrm{df} = \frac{(s_{\bar{X}}^2 + s_{\bar{Y}}^2)^2}{s_{\bar{X}}^4/(n-1) + s_{\bar{Y}}^4/(m-1)}$$

Example (iron retention)

Percentage of Fe^{2+} and Fe^{3+} retained by mice data at concentration 1.2 millimolar.

Fe^{2+}: $n = 18$, $\bar{X} = 9.63$, $s_1 = 6.69$, $s_{\bar{X}} = 1.58$

Fe^{3+}: $m = 18$, $\bar{Y} = 8.20$, $s_2 = 5.45$, $s_{\bar{Y}} = 1.28$

Boxplots and normal probability plots show that the population distributions are not normal.

We test H_0: $\mu_1 = \mu_2$ the large sample test. Using the observed value $T_{\mathrm{obs}} = \frac{\bar{X}-\bar{Y}}{\sqrt{s_{\bar{X}}^2+s_{\bar{Y}}^2}} = 0.7$, the approximate two-sided $P-$value $= 0.48$.

After the log transformation the data look more like normally distributed, as seen from the boxplots and normal probability plots. For the transformed data, we have

$n = 18$, $\bar{X}' = 2.09$, $s_1' = 0.659$, $s_{\bar{X}'} = 0.155$,

$m = 18$, $\bar{Y}' = 1.90$, $s_2' = 0.574$, $s_{\bar{Y}'} = 0.135$.

Two sample t-test for the transformed data

equal variances: $T = 0.917$, df $= 34$, $P = 0.3656$,

unequal variances: $T = 0.917$, df $= 33$, $P = 0.3658$.

Wilcoxon rank sum test

Assume general nonparametric population distributions F_1 and F_2, and consider H_0: $F_1 = F_2$ against H_1: $F_1 \neq F_2$. The rank sum test procedure:

- pool the samples and replace the data values by their ranks $1, 2, \ldots, n + m$,
- compute test statistics R_1 = sum of the ranks of X observations, and R_2 = sum of Y ranks,
- use the null distribution table for R_1 and R_2, which depend only on sample sizes n and m.

Example (in class experiment)
Height distributions for females F_1, and males F_2. For $n = m = 3$, compute R_1 and one-sided P-value.

For $n \geq 10$, $m \geq 10$ apply the normal approximation for the null distributions of R_1 and R_2.

$$R_1 + R_2 = \binom{n+m+1}{2},\ \mathrm{E}(R_1) = \frac{n(n+m+1)}{2},\ \mathrm{E}(R_2) = \frac{m(n+m+1)}{2},\ \mathrm{Var}(R_1) = \mathrm{Var}(R_2) = \frac{mn(n+m+1)}{12}.$$

2 Paired samples

Examples of paired observations:
different drugs for two patients matched by age, sex,
a fruit weighed before and after shipment,
two types of tires tested on the same car.
A paired sample is a vector of IID pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. This should be treated a one-dimensional IID sample $(D_1, \ldots, D_n)$ of the sample differences $D_i = X_i - Y_i$. Again, estimate the population difference $\mu_1 - \mu_2$ using the sample mean $\bar{D} = \bar{X} - \bar{Y}$.
Correlation coefficient $\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_1 \sigma_2}$ is a unit-free measure of dependence.
We have $\rho = 0$ for independent pairs. Smaller standard error if $\rho > 0$:

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) - 2\sigma_{\bar{X}}\sigma_{\bar{Y}}\rho < \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y})$$

Example (platelet aggregation)
Paired measurements of $n = 11$ individuals before smoking, Y_i, and after smoking, X_i. Using the data we estimate correlation as $\rho \approx 0.90$.

Y_i	X_i	D_i	Signed rank
25	27	2	+2
25	29	4	+3.5
27	37	10	+6
44	56	12	+7
30	46	16	+10
67	82	15	+8.5
53	57	4	+3.5
53	80	27	+11
52	61	9	+5
60	59	–1	–1
28	43	15	+8.5

Assuming $D \sim \mathrm{N}(\mu, \sigma^2)$ apply the one-sample t-test to H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$.
Observed test statistic $\frac{\bar{D}}{s_{\bar{D}}} = \frac{10.27}{2.40} = 4.28$. Two-sided P-value = 2*(1 – tcdf(4.28,10)) = 0.0016.

Sign test
No assumption except IID sampling. Non-parametric test of H_0: $M_D = 0$ against H_1: $M_D \neq 0$.
Test statistics: either $Y_+ = \sum 1_{\{D_i > 0\}}$ or $Y_- = \sum 1_{\{D_i < 0\}}$. Both have null distribution $\mathrm{Bin}(n, 0.5)$.

Ties $D_i = 0$: discard the tied observations and reduce n or dissolve the ties by randomisation.

Example (platelet aggregation)
Observed test statistic $Y_- = 1$. A two-sided P-value $= 2[(0.5)^{11} + 11(0.5)^{11}] = 0.012$.

Wilcoxon signed rank test
Non-parametric test of H_0: distribution of D is symmetric about $M_D = 0$. Test statistics:
either $W_+ = \sum \text{rank}(|D_i|) \cdot I(D_i > 0)$ or $W_- = \sum \text{rank}(|D_i|) \cdot I(D_i < 0)$.
Assuming no ties we get $W_+ + W_- = \frac{n(n+1)}{2}$. The null distributions of W_+ and W_- are the same and tabulated for smaller values of n. For $n \geq 20$, use the normal approximation of the null distribution with $\mu_W = \frac{n(n+1)}{4}$ and $\sigma_W^2 = \frac{n(n+1)(2n+1)}{24}$.

> The signed rank test uses more data information than the sign test but requires symmetric distribution of differences.

Example (platelet aggregation)
Observed value of the test statistic $W_- = 1$. It gives a two-sided P-value = 0.002 (check symmetry).

3 Influence of external factors

Double-blind, randomised controlled experiments are used to balance out such external factors as placebo effect, time factor, background variables like temperature, location factor.

Example (portocaval shunt)
Portocaval shunt is an operation used to lower blood pressure in the liver. People believed in its high efficiency until the controlled experiments were performed.

Enthusiasm level	Marked	Moderate	None
No controls	24	7	1
Nonrandomized controls	10	3	2
Randomized controls	0	1	3

Example (platelet aggregation)
Further parts of the experimental design: control group 1 smoked lettuce cigarettes, control group 2 "smoked" unlit cigarettes.

Simpson's paradox
Hospital A has higher overall death rate than hospital B. However, if we split the data in two parts, patients in good (+) and bad (−) conditions, for both parts A performs better.

Hospital:	A	B	A+	B+	A–	B–
Died	63	16	6	8	57	8
Survived	2037	784	594	592	1443	192
Total	2100	800	600	600	1500	200
Death Rate	.030	.020	.010	.013	.038	.040

Here, the external factor, patient condition, is an example of a confounding factor:

$$\text{Hospital performance} \leftarrow \text{Patient condition} \rightarrow \text{Death rate}$$