Serik Sagitov, Chalmers Tekniska Högskola, August 3, 2004

# Chapter 11. Comparing two samples

Data: two IID samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$
two populations with $(\mu_x, \sigma_x)$ and $(\mu_y, \sigma_y)$
Unbiased estimate $(\bar{X} - \bar{Y})$ of $(\mu_x - \mu_y)$
interval estimate of $(\mu_x - \mu_y)$, test $H_0$: $\mu_x = \mu_y$

## 1. Two independent samples

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

## Large sample test for the difference

If $n$ and $m$ are large use
$\bar{X} - \bar{Y} \overset{a}{\sim} \mathrm{N}(\mu_x - \mu_y, s_{\bar{x}}^2 + s_{\bar{y}}^2)$
Approximate CI for $(\mu_x - \mu_y)$
$\bar{X} - \bar{Y} \pm z_{\alpha/2} \cdot \sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}$
Dichotomous data: $X \sim \mathrm{Bin}(n, p_1)$, $Y \sim \mathrm{Bin}(m, p_2)$
$\hat{p}_1 - \hat{p}_2 \overset{a}{\sim} \mathrm{N}(p_1 - p_2, \frac{\hat{p}_1\hat{q}_1}{n-1} + \frac{\hat{p}_2\hat{q}_2}{m-1})$
approximate CI for $(p_1 - p_2)$: $\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1\hat{q}_1}{n-1} + \frac{\hat{p}_2\hat{q}_2}{m-1}}$

## Ex 1: Swedish polls

Two 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 $|p_1 - p_2| > 1.96 \cdot \sqrt{2 \cdot \frac{0.4 \cdot 0.6}{5000}} \approx 1.9\%$

**Two-sample t-test**

Assumption: $X \sim \mathrm{N}(\mu_x, \sigma^2)$, $Y \sim \mathrm{N}(\mu_y, \sigma^2)$

$\mathrm{Var}(\bar{X} - \bar{Y}) = \sigma^2 \cdot \frac{n+m}{nm}$

Pooled sample variance

$s_p^2 = \frac{n-1}{n+m-2} \cdot s_x^2 + \frac{m-1}{n+m-2} \cdot s_y^2$ with $\mathrm{E}(s_p^2) = \sigma^2$

$$\boxed{\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{s_p} \cdot \sqrt{\frac{nm}{n+m}} \sim t_{m+n-2}}$$

Exact CI for $(\mu_x - \mu_y)$

$\bar{X} - \bar{Y} \pm t_{m+n-2}(\frac{\alpha}{2}) \cdot s_p \cdot \sqrt{\frac{n+m}{nm}}$

Two sample $t$-test, equal population variances

$$\boxed{H_0\text{: } \mu_x = \mu_y, \text{ null distribution } \frac{\bar{X}-\bar{Y}}{s_p} \cdot \sqrt{\frac{nm}{n+m}} \sim t_{m+n-2}}$$

Different variances: $X \sim \mathrm{N}(\mu_x, \sigma_x^2)$, $Y \sim \mathrm{N}(\mu_y, \sigma_y^2)$

$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{s_{\bar{x}}^2+s_{\bar{y}}^2}} \overset{a}{\sim} t_{\mathrm{df}}$, $\mathrm{df} = \frac{(s_{\bar{x}}^2+s_{\bar{y}}^2)^2}{s_{\bar{x}}^4/n+s_{\bar{y}}^4/m} - 2$

**Ex 2: iron retention study**

Percentage of $Fe^{2+}$ and $Fe^{3+}$ retained by mice

data for the concentration 1.2 millimolar: p. 396

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

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

Boxplots and normal probability plot: p. 397

distributions are not normal

Test $H_0$: $\mu_x = \mu_y$ using observed $\frac{\bar{X}-\bar{Y}}{\sqrt{s_{\bar{x}}^2+s_{\bar{y}}^2}} = 0.7$

approximate two-sided $P-$value $= 0.48$

After the log transformation of the data

boxplots and normal probability plot: p. 398-399

$n = 18$, $\bar{X} = 2.09$, $s_x = 0.659$, $s_{\bar{x}} = 0.155$

$m = 18$, $\bar{Y} = 1.90$, $s_y = 0.574$, $s_{\bar{y}} = 0.135$

Two sample $t$-test

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

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

**Wilcoxon rank sum test**

Nonparametric test

general population distributions $F$ and $G$

$H_0$: $F = G$ against $H_1$: $F \neq G$

| Pool the samples and replace the data by ranks |
|---|

Test statistics

either $R_x =$ sum of the ranks of $X$ observations

or $R_y = \binom{n+m+1}{2} - R_x$ the sum of $Y$ ranks

Null distributions of $R_x$ and $R_y$ depend only on

sample sizes $n$ and $m$: table 8, p. A21-23

$\mathrm{E}(R_x) = \frac{n(m+n+1)}{2}$, $\mathrm{E}(R_y) = \frac{m(m+n+1)}{2}$

$\mathrm{Var}(R_x) = \mathrm{Var}(R_y) = \frac{mn(m+n+1)}{12}$

For $n \geq 10$, $m \geq 10$ apply

the normal approximation of null distributions

**Ex 3: student heights**

In class: X = females, Y = males, $R_x =$?, one-sided $P =$?

## 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

Paired sample: IID vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$

use $D_i = X_i - Y_i$ to estimate $\mu_x - \mu_y$ with $\bar{D} = \bar{X} - \bar{Y}$

Correlation coefficient $\rho = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}$

$\rho > 0$ for paired observations

$\rho = 0$ for independent observations

Smaller standard error if $\rho > 0$

$\text{Var}(\bar{D}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) - 2\sigma_{\bar{x}}\sigma_{\bar{y}}\rho$

## Ex 4: platelet aggregation

$n = 11$ individuals before $Y_i$ and after $X_i$ smoking

| $Y_i$ | $X_i$ | $D_i$ | 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

$H_0$: $\mu_x = \mu_y$ against $H_1$: $\mu_x \neq \mu_y$

Observed test statistic $\frac{\bar{D}}{s_{\bar{D}}} = \frac{10.27}{2.40} = 4.28$

$\rho \approx 0.90$

two-sided P-value = 2*(1 – tcdf(4.28,10)) = 0.0016

**The sign test**

Non-parametric test of

$H_0$: $M_D = 0$ against $H_1$: $M_D \neq 0$

no assumption except IID sampling

Test statistics

either $Y_+ = \Sigma\, I(D_i > 0)$ or $Y_- = \Sigma\, I(D_i < 0)$

null distributions $Y_+ \sim \mathrm{Bin}(n, 0.5)$, $Y_- \sim \mathrm{Bin}(n, 0.5)$

| |
|---|
| Ties $D_i = 0$: discard tied observations reduce $n$<br>or dissolve the ties by randomization |

**Ex 4: platelet aggregation**

Observed test statistic $Y_- = 1$

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 $W_+ = \Sigma\, \text{rank}(|D_i|) \cdot I(D_i > 0)$

or $W_- = \Sigma\, \text{rank}(|D_i|) \cdot I(D_i < 0)$

assuming no ties $W_+ + W_- = \frac{n(n+1)}{2}$

Null distributions of $W_+$ and $W_-$ are equal

Table 9, p. A24: whatever is the PD of $D$

Normal approximation of the null distribution

with $\mu_W = \frac{n(n+1)}{4}$, $\sigma_W^2 = \frac{n(n+1)(2n+1)}{24}$, $n \geq 20$

| Signed rank test uses more data information than sign test but requires symmetric distribution of differences |
|---|

**Ex 4: platelet aggregation**

observed value $W_- = 1$

two-sided P-value = 0.002 (check symmetry)

**3. External factors**

Double-blind, randomized controlled experiments

to balance out external factors like placebo effect

Other examples of external factors

time and background variables like temperature

locations of test animals or test plots in a field

## Ex 5: portocaval shunt

Portocaval shunt to lower blood pressure in the liver

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

## Ex 4: platelet aggregation

controll group 1 smoked lettuce cigarettes
controll group 2 “smoked” unlit cigarettes

## Ex 6: Simpson’s paradox

The factor of interest, death rate, is confounded with the patient condition distribution: + good, – bad

| 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 |