Serik Sagitov, Chalmers Tekniska Högskola, May 20, 2004

Chapter 9. Testing hypotheses and assessing goodness of fit

1. Hypotheses testing

Choose between two mutually exclusive hypotheses

null hypothesis H_0: the effect of interest is zero

alternative H_1: the effect of interest is not zero

H_0 represents an established theory that must be discredited in order to demonstrate some effect H_1

Two types of error

type I error = false positive: reject H_0 when it's true

type II error = false negative: accept H_0 when it's false

Test result	Negative: accept H_0	Positive: reject H_0
If H_0 is true	True negative specificity $= 1 - \alpha$	False positive $\alpha = \mathrm{P}(\text{reject } H_0 \vert H_0)$
If H_1 is true	False negative $\beta = \mathrm{P}(\text{accept } H_0 \vert H_1)$	True positive sensitivity $= 1 - \beta$

Significance test

Test statistic = a function of the data

with distinct typical values under H_0 and H_1

Rejection region (RR) of a test

a set of values for the test statistic where H_0 is rejected

> If test statistic and sample size are fixed, then either
> ($\alpha \nearrow \beta \searrow$) or ($\alpha \searrow \beta \nearrow$), when RR is changed

Significance test approach to choose RR

fix an appropriate significance level α

find RR from $\alpha = \text{P}(\text{test statistic} \in \text{RR}|H_0)$

using the null distribution of the test statistic

Common significance levels: 5%, 1%, 0.1%

2. Large-sample test for proportion

Sample count $Y \sim \text{Bin}(n, p)$, sample proportion $p = \frac{Y}{n}$

For H_0: $p = p_0$ use test statistic $Z = \frac{Y - np_0}{\sqrt{np_0 q_0}} = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$

approximate null distribution: $Z \overset{a}{\sim} \text{N}(0,1)$

RRs for three composite alternative hypotheses

one-sided H_1: $p > p_0$, $\text{RR} = \{Z \geq z_\alpha\}$

one-sided H_1: $p < p_0$, $\text{RR} = \{Z \leq -z_\alpha\}$

two-sided H_1: $p \neq p_0$, $\text{RR} = \{Z \geq z_{\alpha/2} \text{ or } Z \leq -z_{\alpha/2}\}$

Power function

power of the test (sensitivity): $\text{Pw} = \text{P}(\text{reject} H_0 | H_1)$

Power function of the one-sided test

$$\text{Pw}(p_1) = \text{P}\left(\frac{Y - np_0}{\sqrt{np_0 q_0}} \geq z_\alpha \,|\, p = p_1\right)$$
$$\approx 1 - \Phi\left(\frac{z_\alpha \sqrt{p_0 q_0} + \sqrt{n}(p_0 - p_1)}{\sqrt{p_1 q_1}}\right), \quad p_1 > p_0$$

Planning of sample size

given α and β for H_0: $p = p_0$, H_1: $p = p_1$

choose sample size n such that $\sqrt{n} = \frac{z_\alpha \sqrt{p_0 q_0} + z_\beta \sqrt{p_1 q_1}}{|p_1 - p_0|}$

Ex 1: extrasensory perception

ESP test: guess the suits of $n = 100$ cards
chosen at random with replacement from a deck

Number of cards guessed correctly $Y \sim \text{Bin}(100, p)$
$H_0 : p = 0.25$ (guessing), $H_1 : p > 0.25$ (ESP ability)

Rejection region at 5% significance level
$\text{RR} = \{\frac{\hat{p}-0.25}{0.0433} \geq 1.645\} = \{\hat{p} \geq 0.32\} = \{Y \geq 32\}$

Simple alternative $H_1 : p = 0.30$
power of the test $1 - \Phi(\frac{1.645 \cdot 0.0433 - 0.5}{0.0458}) = 32\%$

Sample size required for the 90% power
$n = (\frac{1.645 \cdot 0.0433 + 1.28 \cdot 0.0458}{0.05})^2 = 675$

P-value of the test

P-value is the probability of obtaining data as extreme or
more extreme than the current data given H_0 is true

If P $\leq \alpha$, reject H_0 at the significance level α
if P $> \alpha$, do not reject H_0 at level α

Difference between P-value and significance level α
α can be chosen before the data is observed

Two-sided P-value = 2× one-sided P-value

Ex 1: extrasensory perception

If observed $Y_{\text{obs}} = 30$, then $Z_{\text{obs}} = \frac{0.3-0.25}{0.0433} = 1.15$ and
one-sided P = P$(Z \geq 1.15) = 12.5\%$
the result is not significant, do not reject H_0

3. Small-sample test for the proportion

Test statistic $Y \sim \text{Bin}(n, p)$, H_0: $p = p_0$

exact null distibution $Y \sim \text{Bin}(n, p_0)$

if n is small, we can not use normal approximation

Significance tests

one-sided H_1: $p > p_0$, RR $= \{Y \geq y_\alpha\}$

one-sided H_1: $p < p_0$, RR $= \{Y \leq y'_\alpha\}$

two-sided H_1: $p \neq p_0$, RR $= \{Y \geq y_{\alpha/2}$ or $Y \leq y'_{\alpha/2}\}$

Ex 1: extrasensory perception

ESP test: guess the suits of $n = 20$ cards

number of cards guessed correctly $Y \sim \text{Bin}(20, p)$

$H_0 : p = 0.25$ against $H_1 : p > 0.25$

Null distribution

Bin(20,0.25) table:

y	8	9	10	11
$\text{P}(Y \geq y)$	.101	.041	.014	0.004

Rejection region at 5% significance level $= \{Y \geq 9\}$

exact significance level $= 4.1\%$

Power function: $\text{Pw}(p_1) = \text{P}[Y \geq 9 | Y \sim \text{Bin}(20, p_1)]$

p_1	0.27	0.3	0.4	0.5	0.6	0.7
$\text{Pw}(p_1)$	0.064	0.113	0.404	0.748	0.934	0.995

Warning for "fishing expeditions": the number of false positives in k tests at level α is Pois $(k\alpha)$

4. Tests for mean

Test H_0: $\mu = \mu_0$ for continuous or discrete data

Large-sample test for mean

PD is not necessarily normal

Test statistic $T = \frac{\bar{X}-\mu_0}{s_{\bar{X}}}$
approximate null distribution $T \overset{a}{\sim} \mathrm{N}(0,1)$

Ex 2: radon level in home

Swedish official limit of the radon level in home:

year average = 400 disintegrations per second and m^3

Data: 36 measurements in your home: $\bar{X} = 450$, $s = 180$

PD is non-normal, test H_0: $\mu = 400$ vs H_1: $\mu \geq 400$

Observed test statistic $T = \frac{450-400}{30} = 1.67$

one-sided P = 0.048, reject H_0 at $\alpha = 5\%$

One-sample t-test

Use for small n, PD must be normal

H_0: $\mu = \mu_0$, test statistic: $T = \frac{\bar{X}-\mu_0}{s_{\bar{X}}}$
exact null distribution: $T \sim t_{n-1}$

CI method of hypotheses testing

accept H_0: $\mu{=}\mu_0$ at 5% level if a 95% CI covers μ_0

reject H_0 at 5% level if a 95% CI does not cover μ_0

5. Likelihood ratio test

A general method of finding asymptotically optimal tests with the largest power for a given level α

Two simple hypotheses

H_0: $\theta = \theta_0$, H_1: $\theta = \theta_1$, likelihood ratio: $\Lambda = \frac{L(\theta_0)}{L(\theta_1)}$

large Λ: H_0 explains the data set better than H_1

small Λ: H_1 explains the data set better

LRT: reject H_0 for $\Lambda \le \lambda_\alpha$ Neyman-Pearson lemma: LRT is optimal

Nested hypotheses

H_0: $\theta \in \Omega_0$, H: $\theta \in \Omega$, nested parameter sets $\Omega_0 \subset \Omega$

alternative hypothesis H_1: $\theta \in \Omega \setminus \Omega_0$

Generalized LRT: reject H_0 for small values of $\frac{L(\hat\theta_0)}{L(\hat\theta)}$

$\hat\theta_0$ = maximizes likelihood over $\theta \in \Omega_0$

$\hat\theta$ = maximizes likelihood over $\theta \in \Omega$

GLRT: reject H_0 for large $\Delta = \log L(\hat\theta) - \log L(\hat\theta_0)$

Approximate null distribution:

$2\Delta \overset{a}{\sim} \chi^2_{\mathrm{df}}$, df = dim($\Omega$) – dim($\Omega_0$)

6. Pearson's chi-square test

Test how well a model fits the data

H_0: $(p_1, \ldots, p_J) = (p_1(\lambda), \ldots, p_J(\lambda))$

unknown parameter $\lambda = (\lambda_1, \ldots, \lambda_r)$, $\dim(\Omega_0) = r$

MLE $\hat{\lambda}$ assuming H_0

expected cell counts $E_j = n \cdot p_j(\hat{\lambda})$

Chi-square test statistic: $X^2 = \Sigma_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j}$

Reject H_0 for large values of $2\Delta \approx X^2$

GLRT: approximate null distribution of X^2 is χ^2_{J-1-r}

df = (number of cells) – 1 – (number of independent parameters estimated from the data)

All expected counts are recommended to be at least 5

combine small cells and recalculate df

Ex 3: red mites

H_0 : number of red mites on a leaf $\sim$ Pois(λ)

$\hat{\lambda} = 1.147$, $X^2 = 52.8$

5 cells, df = 3, $\chi^2_3(0.001) = 16.3$, reject H_0

Ex 4: bird hops

H_0 : no. hops that a bird does between flights $\sim$ Geom(p)

$\hat{p} = 0.358$, $X^2 = 1.86$, number of cells = 7

df = 5, P-value = 0.87, accept H_0

Ex 5: gender ratio

Germany 1889: $n = 6115$ families with 12 children

data: $Y_1, \ldots, Y_n$ numbers of boys in each family

Simple model: $Y \sim$ Bin(12, 0.5)

simple H_0: $p_j = \binom{12}{j-1} \cdot 2^{-12}, j = 1, \ldots, 13$

Expected cell counts $E_j = 6115 \cdot \binom{12}{j-1} \cdot 2^{-12}$

$X^2 = 249.2$, df $= 12$, $\chi^2_{12}(0.005) = 28.3$, reject H_0

y	cell j	O_j	E_j	$\frac{(O_j-E_j)^2}{E_j}$	E_j	$\frac{(O_j-E_j)^2}{E_j}$
0	1	7	1.5	20.2	2.3	9.6
1	2	45	17.9	41.0	26.1	13.7
2	3	181	98.5	69.1	132.8	17.5
3	4	478	328.4	68.1	410.0	11.3
4	5	829	739.0	11.0	854.2	0.7
5	6	1112	1182.4	4.2	1265.6	18.6
6	7	1343	1379.5	1.0	1367.3	0.4
7	8	1033	1182.4	18.9	1085.2	2.5
8	9	670	739.0	6.4	628.1	2.8
9	10	286	328.4	5.5	258.5	2.9
10	11	104	98.5	0.3	71.8	14.4
11	12	24	17.9	2.1	12.1	11.7
12	13	3	1.5	1.5	0.9	4.9
Total		6115	6115	249.2	6115	110.5

More flexible model: $Y \sim \text{Bin}(12, p)$ unspecified p

composite H_0: $p_j = \binom{12}{j-1} \cdot p^{j-1} \cdot q^{13-j}, j = 1, \ldots, 13$

$\hat{p} = \frac{\text{number of boys}}{\text{number of children}} = \frac{1 \cdot 45 + 2 \cdot 181 + \ldots + 12 \cdot 3}{6115 \cdot 12} = 0.4808$

Expected cell counts $E_j = 6115 \cdot \binom{12}{j-1} \cdot \hat{p}^{j-1} \cdot \hat{q}^{13-j}$

observed test statistic $X^2 = 110.5$

$r = 1$, df $= 11$, $\chi^2_{11}(0.005) = 26.76$

reject H_0 at 0.5% level

Possible explanation : probability of a male child

p differs from family to family, larger variation