Serik Sagitov, Chalmers Tekniska Högskola, August 16, 2004

5. Hypotheses testing

5.1 Significance test

Def 1: null and alternative 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

Def 2: type I and type II errors

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

State of nature	Accept H_0 negative result	Reject H_0 positive result
H_0 is true	True negative specificity $= 1 - \alpha$	False positive $\alpha = \mathrm{P}(\text{reject } H_0 \mid H_0)$
H_1 is true	False negative $\beta = \mathrm{P}(\text{accept } H_0 \mid H_1)$	True positive sensitivity $= 1 - \beta$

Def 3: test statistic and rejection region

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

Def 4: significance test

fix an appropriate significance level α

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

using the null distribution of the test statistic

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

Type I error is believed to have graver consequences

therefore we control α and not β

5.2 Large-sample test for proportion

Sample count $Y \sim \mathrm{Bin}(n, p)$, $p =$ population proportion

simple null hypothesis H_0: $p = p_0$

Test statistic $Z = \frac{Y - np_0}{\sqrt{np_0 q_0}}$

approximate null distribution: $Z \approx \mathrm{N}(0,1)$

RRs for three composite alternative hypotheses

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

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

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

Ex 1: postponed deaths

H_0: death can not be postponed

H_1: death can be postponed untill after important date

Jewish data, California: $n = 1919$ deaths including

$Y = 922$ deaths during the week before Passover

$Y \sim \mathrm{Bin}(n, p)$, H_0: $p = 0.5$, H_1: $p < 0.5$

observed test statistic $Z = -1.712$

If $\alpha = 0.05$, then $z_\alpha = 1.645$ and $Z < -z_\alpha$

reject H_0 in favor of H_1 at 5% level

Seasonal effect? Chinese and Japanese data (California):

$n = 852$, $Y = 418$, $n - Y = 434$, observed $Z = -0.548$

can not reject H_0 in favor of H_1 at 5% level

Def 5: P-value of a test

P = the smallest significance level at which the test rejects H_0 for the observed data

shows how significantly observed data contradicts H_0

If $\mathrm{P} \le \alpha$, reject H_0 at the significance level α
If $\mathrm{P} > \alpha$, do not reject H_0 at level α

Ex 1: postponed deaths

Jewish data: one-sided $\mathrm{P} = \mathrm{P}(Z < -1.712) = 0.043$

Chi-Jap data: one-sided $\mathrm{P} = \mathrm{P}(Z < -0.548) = 0.292$

Ex 2: aspirin treatment

placebo group: 11034 individuals, 189 heart attacks

aspirin group: 11037 individuals, 104 heart attacks

Test $H_0 = \{\text{no aspirin effect}\}$ against

$H_1 = \{\text{aspirin reduces the risk of heart attack}\}$

Test statistic $X = \#\{\text{heart attacks in the placebo group}\}$

$\mathrm{P} = \mathrm{P}(X \ge 189 | H_0) = 0.0000003$

5.3 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 $Y \leq y'_{\alpha/2}\}$

Def 6: test power

sensitivity of the test $1 - \beta = \text{P}(\text{reject } H_0 | H_1)$

Powerless decision rule at 0% level: never reject H_0

Ex 3: extrasensory perception (ESP)

ESP test: guess the suits of 20 cards

chosen at random with replacement from a deck

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

$H_0 : p = 0.25$ (pure guessing)

$H_1 : p > 0.25$ (ESP ability)

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

p	0.27	0.3	0.4	0.5	0.6	0.7
$\text{Pw}(p)$	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 $(k\alpha)$

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: postponed deaths

Planning of sample size for H_0: $p = 0.50$, H_1: $p = 0.48$

$\alpha = 0.05$, $\beta = 0.10$ requires $n = 5351$ observations

$\alpha = \beta = 0.05$ requires $n = 6773$ observations

Larger power requires larger sample

5.4 Large-sample test for mean

PD is not necessarily normal, test H_0: $\mu = \mu_0$

Test statistic $T = \frac{\bar{X}-\mu_0}{s_{\bar{X}}}$

approximate null distribution $T \approx$ N(0,1)

Ex 4: 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 $\alpha = 5\%$

5.5 One-sample t-test

used for small n, PD must be normal

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

Ex 5: measuring fat content

Two methods of measuring in % the fat content of meat

pairwise diff. for 16 hotdogs: $\bar{X} = 0.53\%$, $s = 1.06\%$

Test H_0: $\mu = 0$ against H_1: $\mu \neq 0$

$s_{\bar{X}} = 0.265$, observed test statistic $T = 2.0$

one-sided P $= \mathrm{P}(T < 2 | T \sim t_{15}) = 0.032$

Two-sided P $= 2 \cdot 0.032 = 0.064$

do not reject H_0: $\mu = 0$ in favor of H_1: $\mu \neq 0$

Choose H_1 before seeing the data

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

CI is more informative than a test result
wider CI indicates less power of the test

Ex 5: measuring fat content

Exact 95% CI for the mean difference μ is $(-0.03, 1.08)$

do not reject H_0: $\mu = 0$ in favor of H_1: $\mu \neq 0$

note that approximate 95% CI is $(0.01, 1.05)$

5.6 Likelihood ratio test (LRT)

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 $\Lambda \leq \lambda_\alpha$ 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 $\Lambda = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$

$\hat{\theta}_0 =$ 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 distr: $2\Delta \approx \chi^2_{df}$, df = dim($\Omega$) – dim($\Omega_0$)

Properties of χ^2_k-distribution: mean and variance $\mu = k$, $\sigma^2 = 2k$

different pdf shapes $f_1(0) = \infty$, $f_2(0) = 0.5$, $f_3(0) = 0$