Serik Sagitov, Chalmers Tekniska Högskola, February 4, 2013

Chapter 9. Testing hypotheses and assessing goodness of fit

1 Hypotheses testing

Find a rule based on data for choosing 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: do not reject H_0	Positive: reject H_0
If H_0 is true	True negative. Specificity $= 1-\alpha$	False positive. Significance level α
If H_1 is true	False negative $\beta = \mathrm{P}(\text{accept } H_0 \mid H_1)$	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 when H_0 is rejected.

If test statistic and sample size are fixed, then either α or β gets larger when RR is changed.

Significance test approach to choose a rejection region:
fix an appropriate significance level α,
find a RR from $\alpha = \mathrm{P}(\text{test statistic} \in \mathrm{RR} \mid H_0)$ using the null distribution of the test statistic.

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

2 Large-sample test for the proportion

Data is modeled by a sample count $Y \sim \mathrm{Bin}(n,p)$. An unbiased point estimate for the population proportion p is the sample proportion $p = \frac{Y}{n}$.

For H_0: $p = p_0$ use the test statistic $Z = \frac{Y-np_0}{\sqrt{np_0q_0}} = \frac{\hat{p}-p_0}{\sqrt{p_0q_0/n}}$.

Approximate null distribution: $Z \overset{a}{\sim} \mathrm{N}(0,1)$. Let $\Phi(z_\alpha) = 1-\alpha$. Three different rejection regions 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: $p \neq p_0$, $\mathrm{RR} = \{Z \geq z_{\alpha/2} \text{ or } Z \leq -z_{\alpha/2}\}$.

Power function
The power of the test (sensitivity): Pw = P(reject $H_0|H_1$ is true).
Let H_0: $p = p_0$, H_1: $p = p_1$, and $p_1 > p_0$. The power function of the one-sided test

$$\text{Pw}(p_1) = \text{P}(\tfrac{Y-np_0}{\sqrt{np_0q_0}} \geq z_\alpha \,|\, p = p_1) \approx 1 - \Phi(\tfrac{z_\alpha\sqrt{p_0q_0}+\sqrt{n}(p_0-p_1)}{\sqrt{p_1q_1}}), \quad p_1 > p_0.$$

Planning of sample size: given α and β, choose sample size n such that $\sqrt{n} = \frac{z_\alpha\sqrt{p_0q_0}+z_\beta\sqrt{p_1q_1}}{|p_1-p_0|}$.

Example: extrasensory perception.
ESP test: guess the suits of $n = 100$ cards chosen at random with replacement from a deck of cards with four suits. Number of cards guessed correctly $Y \sim \text{Bin}(100, p)$

$H_0 : p = 0.25$ (pure guessing), $H_1 : p > 0.25$ (ESP ability).
Rejection region at 5% significance level $= \{\frac{\hat{p}-0.25}{0.0433} \geq 1.645\} = \{\hat{p} \geq 0.32\} = \{Y \geq 32\}$.
With a simple alternative $H_1 : p = 0.30$ the power of the test is $1 - \Phi(\frac{1.645\cdot 0.0433-0.5}{0.0458}) = 32\%$.
The sample size required for the 90% power is $n = (\frac{1.645\cdot 0.0433+1.28\cdot 0.0458}{0.05})^2 = 675$.

P-value of the test
P-value is the probability of obtaining a test statistic value as extreme or more extreme than the observed one, given that H_0 is true.
For the significance level α, reject H_0, if $\text{P} \leq \alpha$, and do not reject H_0, if $\text{P} > \alpha$.

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

Example: extrasensory perception.
If the observed sample count is $Y_{\text{obs}} = 30$, then $Z_{\text{obs}} = \frac{0.3-0.25}{0.0433} = 1.15$ and a one-sided P-value is $\text{P}(Z \geq 1.15) = 12.5\%$. The result is not significant, do not reject H_0.

3 Small-sample test for the proportion

With H_0: $p = p_0$ the test statistic $Y \sim \text{Bin}(n, p)$ for small n we have to rely on the exact null distribution $Y \sim \text{Bin}(n, p_0)$. Three rejection regions

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

Example: extrasensory perception.
ESP test: guess the suits of $n = 20$ cards. Model: the number of cards guessed correctly is $Y \sim \text{Bin}(20, p)$. For $H_0 : p = 0.25$ the null distribution is

Bin(20,0.25) table:

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

One-sided alternative $H_1 : p > 0.25$. Rejection region at 5% significance level $= \{Y \geq 9\}$. Notice that the 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 the mean

Test H_0: $\mu = \mu_0$ for continuous or discrete data. Large-sample test for mean is used when the population distribution is not necessarily normal but the sample size n is sufficiently large.

H_0: $\mu = \mu_0$, test statistic $T = \frac{\bar{X}-\mu_0}{s_{\bar{X}}}$ with an approximate null distribution $T \overset{a}{\sim}$ N(0,1).

The one-sample t-test is used for small n, assuming that the population distribution is normal.

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

CI method of hypotheses testing:
reject H_0: μ=μ_0 at 5% level if and only if a 95% confidence interval for the mean does not cover μ_0.

5 Likelihood ratio test

A general method of finding asymptotically optimal tests (having the largest power for a given α).

Two simple hypotheses
For testing H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$ use the likelihood ratio $\Lambda = \frac{L(\theta_0)}{L(\theta_1)}$ as a test statistic. Large values of Λ suggest that H_0 explains the data set better than H_1, while small Λ indicate that H_1 explains the data set better.

Likelihood raio rejection rule: reject H_0 for $\Lambda \leq \lambda_\alpha$.

Neyman-Pearson lemma: the likelihood ration test is optimal in the case of two simple hypothesis.

Nested hypotheses
With a pair of nested parameter sets $\Omega_0 \subset \Omega$ we get two composite alternatives, H_0: $\theta \in \Omega_0$ and H_1: $\theta \in \Omega \setminus \Omega_0$. Two nested hypotheses H_0: $\theta \in \Omega_0$, H: $\theta \in \Omega$, and two maximum likelihood estimates

$\hat{\theta}_0$ = maximizes likelihood over $\theta \in \Omega_0$,
$\hat{\theta}$ = maximizes likelihood over $\theta \in \Omega$.

Generalized LRT: reject H_0 for small values of $\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$ or equivalently

GLRT: reject H_0 for large values of $\Delta = \log L(\hat{\theta}) - \log L(\hat{\theta}_0)$.

Approximate null distribution: $2\Delta \overset{a}{\sim} \chi^2_{\text{df}}$, where df = dim($\Omega$) − dim($\Omega_0$).

6 Pearson's chi-square test

Data: each observation belongs to one of J classes. A null hypothesis proposing a model for the data

H_0: $(p_1, \ldots, p_J) = (p_1(\lambda), \ldots, p_J(\lambda))$ with unknown parameter $\lambda = (\lambda_1, \ldots, \lambda_r)$, $\dim(\Omega_0) = r$.

Test how well a model fits the data using the MLE $\hat{\lambda}$ of λ describing H_0. Data is summarized as the vector of observed counts $(O_1, \ldots, O_J)$.

Chi-square test statistic: $X^2 = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j}$, expected cell counts $E_j = n \cdot p_j(\hat{\lambda})$.

Generalized likelihood ratio test approach: reject H_0 for large values of $2\Delta \approx X^2$ having an approximate null distribution χ^2_{J-1-r}.

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

Since the chi-square test is approximate, all expected counts are recommended to be at least 5. If not, combine small cells and recalculate df.

Example: bird hops.
H_0 : number of hops that a bird does between flights has a Geom(p) distribution. Using a MLE $\hat{p} = 0.358$ and $J = 7$ we obtain $X^2 = 1.86$. With df = 5 and P-value = 0.87 we do not reject the geometric distribution model for number of bird hops.

Example: gender ratio.
In a study made in Germany in 1889 the gender ratios for $n = 6115$ families with 12 children were recorded. The data give $Y_1, \ldots, Y_n$ numbers of boys in each family. Each Y_i has $J = 13$ possible values. Here we discuss two models for the gender ratio.
Model 1. A symmetric binomial model: $Y \sim$ Bin(12, 0.5) corresponds to a simple null hypothesis H_0: $p_j = \binom{12}{j} \cdot 2^{-12}, j = 0, 1, \ldots, 12$. Expected cell counts $E_j = 6115 \cdot \binom{12}{j} \cdot 2^{-12}$.

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

Model 1 results: $X^2 = 249.2$, df = 12, $\chi^2_{12}(0.005) = 28.3$, reject H_0 at 0.5% level.
Model 2. More flexible model: $Y \sim$ Bin(12, p) with an unspecified p. It leads to a composite null hypothesis H_0: $p_j = \binom{12}{j} \cdot p^j (1-p)^{12-j}, j = 0, \ldots, 12,\ 0 \leq p \leq 1$. The MLE and expected cell counts

$$\hat{p} = \frac{\text{number of boys}}{\text{number of children}} = \frac{1\cdot 45 + 2\cdot 181 + \ldots + 12\cdot 3}{6115\cdot 12} = 0.4808, \quad E_j = 6115 \cdot \binom{12}{j} \cdot \hat{p}^j \cdot (1-\hat{p})^{12-j}\ .$$

Model 2 results: 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.
Conclusion: even more flexible model is needed to address large variation in the observed cell counts.
Suggestion: let the probability of a male child p to differ from family to family.