Serik Sagitov, Chalmers and GU, February 1, 2016

Chapter 9. Testing hypotheses and assessing goodness of fit

1 Statistical significance

Often we need 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.

	Negative decision: do not reject H_0	Positive decision: reject H_0 in favor of H_1
If H_0 is true	True negative outcome	False positive outcome, type I error
If H_1 is true	False negative outcome, type II error	True positive outcome

A decision rule for hypotheses testing is based a test statistic T, a function of the data with distinct typical values under H_0 and H_1. For an appropriately chosen rejection region $\mathcal{R}$:
reject H_0 in favor of H_1 if and only if $T \in \mathcal{R}$.
Conditional probabilities:

$\alpha = \mathrm{P}_{H_0}(T \in \mathcal{R})$ significance level of the test, conditional probability of type I error,
$1-\alpha = \mathrm{P}_{H_0}(T \notin \mathcal{R})$ specificity of the test,
$\beta = \mathrm{P}_{H_1}(T \notin \mathcal{R})$ conditional probability of type II error,
$1-\beta = \mathrm{P}_{H_1}(T \in \mathcal{R})$ sensitivity of the test or power.

If test statistic and sample size are fixed, then either α or β gets larger when $\mathcal{R}$ is changed.

A significance test tries to control the type I error:
fix an appropriate significance level α, commonly used significance levels are 5%, 1%, 0.1%,
find $\mathcal{R}$ from $\alpha = \mathrm{P}(T \in \mathcal{R}|H_0)$ using the null distribution of the test statistic T.

2 Large-sample test for the proportion

Binomial model $X \sim \mathrm{Bin}(n,p)$. The corresponding sample proportion $\hat{p} = \frac{X}{n}$.

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

Three different composite alternative hypotheses:
one-sided H_1: $p > p_0$, one-sided H_1: $p < p_0$, two-sided H_1: $p \neq p_0$.
By the central limit theorem, the null distribution of the Z-score is approximately normal: $Z \overset{a}{\sim} \mathrm{N}(0,1)$
find z_α from $\Phi(z_\alpha) = 1-\alpha$ using the normal distribution table.

Alternative H_1	Rejection rule	P-value
$p > p_0$	$Z \geq z_\alpha$	$\mathrm{P}(Z \geq Z_{\text{obs}})$
$p < p_0$	$Z \leq -z_\alpha$	$\mathrm{P}(Z \leq Z_{\text{obs}})$
$p \neq p_0$	$Z \leq -z_{\alpha/2}$ or $Z \geq z_{\alpha/2}$	$2 \cdot \mathrm{P}(Z \geq \lvert Z_{\text{obs}}\rvert)$

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 a given significance level α,

reject H_0, if $\mathrm{P} \leq \alpha$, and do not reject H_0, if $\mathrm{P} > \alpha$.

Power function
Consider two simple hypotheses H_0: $p = p_0$ and H_1: $p = p_1$, assuming $p_1 > p_0$. The power function of the one-sided test can be computed using the normal approximation for $Z_1 = \frac{Y - np_1}{\sqrt{np_1 q_1}}$ under H_1:

$$\begin{aligned}\mathrm{Pw}(p_1) &= \mathrm{P}_{H_1}\Big(\frac{Y - np_0}{\sqrt{np_0 q_0}} \geq z_\alpha\Big) \\ &= \mathrm{P}_{H_1}\Big(\frac{Y - np_1}{\sqrt{np_1 q_1}} \geq \frac{z_\alpha \sqrt{p_0 q_0} + \sqrt{n}(p_0 - p_1)}{\sqrt{p_1 q_1}}\Big) \approx 1 - \Phi\big(\frac{z_\alpha \sqrt{p_0 q_0} + \sqrt{n}(p_0 - p_1)}{\sqrt{p_1 q_1}}\big).\end{aligned}$$

Planning of sample size: given α and β, choose sample size n such that $\sqrt{n} = \frac{z_\alpha \sqrt{p_0 q_0} + z_\beta \sqrt{p_1 q_1}}{|p_1 - p_0|}$.

Example (extrasensory perception, ESP)
An experiment: guess the suits of $n = 100$ cards chosen at random with replacement from a deck of cards with four suits. Binomial model: the number of cards guessed correctly $Y \sim \mathrm{Bin}(100, p)$. Hypotheses of interest

$H_0 : p = 0.25$ (pure guessing), $H_1 : p > 0.25$ (ESP ability).

Rejection rule 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 $1 - \Phi(\frac{1.645 \cdot 0.433 - 0.5}{0.458}) = 32\%$.
The sample size required for the 90% power is $n = (\frac{1.645 \cdot 0.433 + 1.28 \cdot 0.458}{0.05})^2 = 675$.

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

3 Small-sample test for the proportion

Binomial model $X \sim \mathrm{Bin}(n, p)$ with H_0: $p = p_0$. For small n, use exact null distribution $X \sim \mathrm{Bin}(n, p_0)$.

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

Bin(20,0.25) table

x	8	9	10	11
$\mathrm{P}(X \geq x)$	.101	.041	.014	0.004

For the one-sided alternative $H_1 : p > 0.25$ and $\alpha = 5\%$, the rejection rule is $\{X \geq 9\}$. Notice that the exact significance level = 4.1%. Warning for "fishing expeditions".

Power function

p_1	0.27	0.30	0.40	0.5	0.60	0.70
$\mathrm{P}(X \geq 9 \mid p = p_1)$	0.064	0.113	0.404	0.748	0.934	0.995

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.

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

One-sample t-test is used for small n, under the assumption that the population distribution is normal.

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

CI method of hypotheses testing

at 5% significance level the rejection rule is $\{\mu_0 \notin 95\%$ confidence interval for the mean$\}$.

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 a small Λ indicates that H_1 explains the data set better. Likelihood ratio test rejects H_0 for small values of Λ.

Neyman-Pearson lemma: the likelihood ratio 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: $\theta \in \Omega \setminus \Omega_0$. Under two nested hypotheses H_0: $\theta \in \Omega_0$, H: $\theta \in \Omega$, we get two maximum likelihood estimates

$\hat{\theta}_0$ = maximises the likelihood function $L(\theta)$ over $\theta \in \Omega_0$,

$\hat{\theta}$ = maximises the likelihood function $L(\theta)$ over $\theta \in \Omega$.

Generalised likelihood ratio test: reject H_0 for small values of $\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$ or equivalently

$$\text{Reject } H_0\colon \theta \in \Omega_0 \text{ for large values of } \Delta = \log L(\hat{\theta}) - \log L(\hat{\theta}_0).$$

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

6 Pearson's chi-square test

Data: each of n IID observations belongs to one of J classes with probabilities $(p_1, \ldots, p_J)$. Data is summarised as the vector of observed counts

$$(O_1, \ldots, O_J) \sim \mathrm{Mn}(n; p_1, \ldots, p_J), \qquad \mathrm{P}(O_1 = k_1, \ldots, O_J = k_J) = \tfrac{n!}{k_1! \cdots k_J!} p_1^{k_1} \cdots p_J^{k_J}.$$

Consider a parametric model for the data

H_0: $(p_1, \ldots, p_J) = (v_1(\lambda), \ldots, v_J(\lambda))$ with unknown parameters $\lambda = (\lambda_1, \ldots, \lambda_r)$.

To see if the proposed model fits the data, compute $\hat{\lambda}$, the maximum likelihood estimate of λ, and then the expected cell counts $E_j = n \cdot v_j(\hat{\lambda})$.

Chi-square test statistic: $X^2 = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j}$ is derived from the likelihood ratio test $2\Delta \approx X^2$.

The approximate null distribution of X^2 is χ^2_{J-1-r}, since $\dim(\Omega_0) = r$ and $\dim(\Omega) = J - 1$.

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 the number of degrees of freedom df.

Example (geometric model)
H_0 : number of hops that a bird does between flights has a geometric distribution Geom(p).
Using $\hat{p} = 0.358$ 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.

7 Gender ratio example

A 1889 study in Germany recorded the numbers of boys $Y_1, \ldots, Y_n$ for $n = 6115$ families with 12 children each. Consider three nested models for the distribution of the number of boys Y

Model 1, $Y \sim \text{Bin}(12, 0.5) \subset$ Model 2, $Y \sim \text{Bin}(12, p) \subset$ General model, $p_j = \text{E}(Y = j)$.

Model 1 leads to a simple null hypothesis H_0: $p_j = \binom{12}{j} \cdot 2^{-12}, j = 0, 1, \ldots, 12$.
Expected cell counts $E_j = 6115 \cdot \binom{12}{j} \cdot 2^{-12}$. Observed $X^2 = 249.2$, df = 12. Since $\chi^2_{12}(0.005) = 28.3$, we reject H_0 at 0.5% level.

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 2 is more flexible and leads to a composite null hypothesis
H_0: $p_j = \binom{12}{j} \cdot p^j (1-p)^{12-j}, j = 0, \ldots, 12, 0 \le p \le 1$. The expected cell counts

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

Model 2 is also rejected at 0.5% level: observed $X^2 = 110.5$, $r = 1$, df = 11, $\chi^2_{11}(0.005) = 26.76$.

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