Introduction to mathematical Statistics

Inference on two population proportions with two large independent samples.

Data:

Sample 1 (population proportion -- π_1): n_1, X_1

Sample 2 (population proportion -- π_2): n_2, X_2

1. Derivation of the Pivotal Quantity:

Parameter of Interest: $\pi_1 - \pi_2$

$$\hat{\pi}_1 = \frac{X_1}{n_1} \dot{\sim} N(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1})$$

$$\hat{\pi}_2 = \frac{X_2}{n_2} \dot{\sim} N(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2})$$

Point Estimator of $(\pi_1 - \pi_2)$ is: $(\hat{\pi}_1 - \hat{\pi}_2)$

Pivotal Quantity (P.Q.) for $(\pi_1 - \pi_2)$:

$$\hat{\pi}_1 - \hat{\pi}_2 \dot{\sim} N(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2})$$

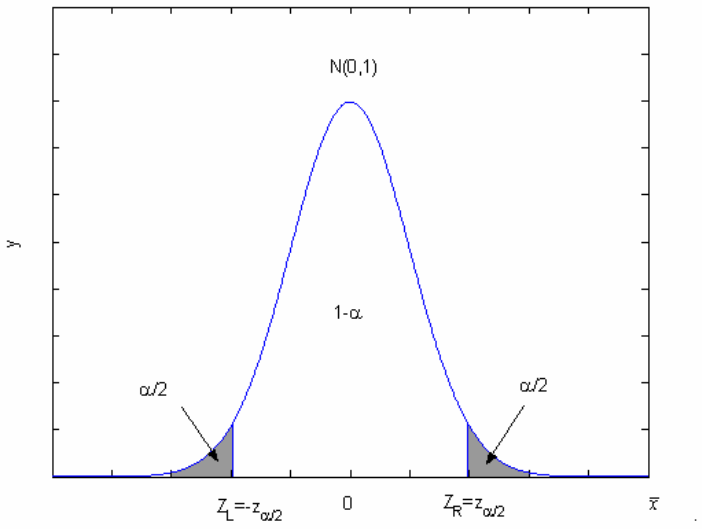
$$\frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \dot{\sim} N(0,1)$$

The above is not a P.Q. because π_1, π_2 are used in $\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$

However the following is a P.Q.:

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} \dot{\sim} N(0,1)$$

2. The 100(1-α)% CI for $\pi_1 - \pi_2$



$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) \approx 1-\alpha$$

… We can follow the same procedure as before to arrive at the following formula for the **100(1-α)% CI for** $\pi_1 - \pi_2$:

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

3. Hypothesis Test

$$H_0 : \pi_1 - \pi_2 = 0$$
$$H_a : \pi_1 - \pi_2 > 0,\ \pi_1 - \pi_2 < 0,\ \pi_1 - \pi_2 \neq 0$$

$$\hat{\pi}_1 - \hat{\pi}_2 \sim N(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2})$$

$$\frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$$

If H_0 is true : $\pi_1 = \pi_2 = \pi$

$$\frac{\hat{\pi}_1 - \hat{\pi}_2 - 0}{\sqrt{\pi(1-\pi)(\frac{1}{n_1} + \frac{1}{n_2})}} \overset{H_0}{\sim} N(0,1)$$

$$\Rightarrow Z_0 = \frac{\hat{\pi}_1 - \hat{\pi}_2 - 0}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1}+\frac{1}{n_2})}} \overset{H_0}{\sim} N(0,1)$$

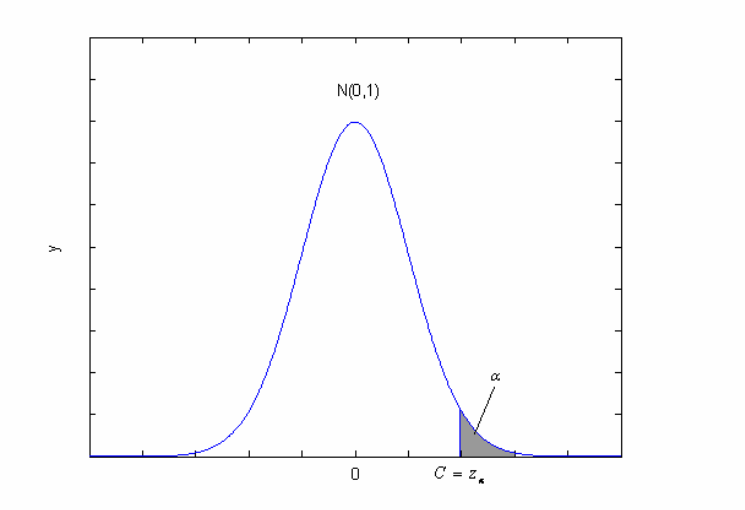
(Note: the above test statistic is only applicable to the null hypothesis $H_0 : \pi_1 - \pi_2 = 0$)

Given our samples n_1, n_2, X_1, X_2 $\Rightarrow$ $\hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2}$

For $H_0 : \pi_1 - \pi_2 = 0$ *versus* $H_a : \pi_1 - \pi_2 > 0$

α =P(reject $H_0 \mid H_0$)

$\alpha = P(Z_0 \geq c \mid H_0 : \pi_1 - \pi_2 = 0)$



$\therefore c = Z_\alpha$ $\therefore$ We reject H_0 at the significance level α if $Z_0 \geq Z_\alpha$ #

Note: We can also use the same pivotal quantity for the CI to derive a second test statistic for the hypothesis test with the null hypothesis $H_0 : \pi_1 - \pi_2 = 0$ as follows:

$$\therefore Z_0 = \frac{\hat{\pi}_1 - \hat{\pi}_2 - 0}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} \overset{H_0}{\sim} N(0,1)$$

Note: the above test statistic is also applicable to the null hypothesis $H_0 : \pi_1 - \pi_2 = c$ where c is any constant between -1 and 1 as follows:

$$Z_0 = \frac{\hat{\pi}_1 - \hat{\pi}_2 - c}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} \overset{H_0}{\sim} N(0,1)$$

*Please see our text book for related examples using the first pivotal test statistic.

*Please see the following example for related examples using the 2nd test statistic.

Example 1: We are interested in testing whether taking vitamin C would reduce the incidence of cold. A sample of 139 subjects was asked to take the VC pills continuously for 2 winter months and the incidence of cold is found to be 17 during this period. A sample of 140 subjects comparable in age, gender and health conditions was asked to take the placebo (pills that look just like the VC pills) for the same period of time and 31 subjects in this group caught cold during the same 2-month period. Please conduct a test at the significance level of 0.05.

Solution 1: The hypotheses are

$$H_0 : p_1 = p_2 \text{ vs } H_1 : p_1 < p_2$$

(Note: sometimes we use π sometimes we use p to represent the population proportion. Either is OK for our class. Just be consistent with your notation in the exams.)

For the vitamin C group, the sample proportion catching cold is $\hat{p}_1 = 17/139 = 0.122$. For the placebo group, the sample proportion catching cold is $\hat{p}_2 = 31/140 = 0.221$.

The pooled proportion under H_0 is:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{17 + 31}{139 + 140} \approx 0.172$$

The test statistic is

$$Z_0 = \frac{\hat{p}_1 - p_2 - 0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0.122 - 0.221}{\sqrt{0.172(1-0.172)(\frac{1}{139} + \frac{1}{140})}} \approx -2.191$$

Since $z_0 = -2.191 < -Z_{0.05} = -1.645$, we reject H_0 and conclude that taking vitamin C reduces the incidence rate of colds compared to a placebo at the significance level of 0.05.

Alternatively, we can calculate the p-value and make our decision based on the p-value as follows.

The P-value is $P = P(Z_0 \le -2.191 \mid H_0) = 0.0142$

Since $P = 0.0142 < \alpha = 0.05$, we reject H_0 and conclude that taking vitamin C reduces the incidence rate of colds compared to a placebo at the significance level of 0.05

Solution 2: The hypotheses are

$$H_0 : p_1 = p_2 \text{ vs } H_1 : p_1 < p_2$$

(Note: sometimes we use π sometimes we use p to represent the population proportion. Either is OK for our class. Just be consistent with your notation in the exams.)

For the vitamin C group, the sample proportion catching cold is $\hat{p}_1 = 17/139 = 0.122$. For the placebo group, the sample proportion catching cold is $\hat{p}_2 = 31/140 = 0.221$. The test statistic is

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} = \frac{0.122 - 0..221}{\sqrt{\frac{(0.122)(0.878)}{139} + \frac{(0.221)(0.779)}{140}}} = -2.212$$

Since $z_0 = -2.212 < -Z_{0.05} = -1.645$, we reject H_0 and conclude that taking vitamin C reduces the incidence rate of colds compared to a placebo at the significance level of 0.05.

Alternatively, we can calculate the p-value and make our decision based on the p-value as follows.

The P-value is $P = P(Z_0 \le -2.212 \,|\, H_0) = 0.0135$

Since $P = 0.0135 < \alpha = 0.05$, we reject H_0 and conclude that taking vitamin C reduces the incidence rate of colds compared to a placebo at the significance level of 0.05

Inference on several proportions—the Chi-square test (large sample)

Def. Multinomial Experiment.

We have a total of n trials (sample size=n)

① For each trial, it will result in 1 of k possible outcomes.

② The probability of getting outcome i is p_i, and $\sum_{i=1}^{k} p_i = 1$

③ These trials are independent.

Example 2. Previous experience indicates that the probability of obtaining 1 healthy calf from a mating is 0.83. Similarly, the probabilities of obtaining 0 and 2 healthy calves are 0.15 and 0.02 respectively. If the farmer breeds 3 dams from the herd, find the probability of getting exact 3 health calves.

Def. Multinomial Distribution

Let X_i be the number of trials resulted in i-th category out of a total of n trials and p_i be the probability of getting i-th category outcome, then

$$P(X_1 = x_1, X_2 = x_2, ..., X_k = x_k) = \frac{n!}{x_1!x_2!...x_k!} p_1^{x_1} ... p_k^{x_k}$$

Solution: n = 3, k = 3
Category 1: 0 healthy calf
Category 2: 1 healthy calf
Category 3: 2 healthy calves
Note: K! = K*(K-1)*…3*2*1; 0! = 1; 1! = 1
P(exact 3 health calves)
= P(one gave birth to 0 healthy calf, one to 1, and 1 to 2) + P(each gave birth to 1 healthy calf)

$$= P(X_1 = 1, X_2 = 1, X_3 = 1) + P(X_1 = 0, X_2 = 3, X_3 = 0)$$

$$= \frac{3!}{1!1!1!}(0.15)^1(0.83)^1(0.02)^1 + \frac{3!}{0!3!0!}(0.15)^0(0.83)^3(0.02)^0$$

= 0.015+0.572 = 0.59

*Relations to the Binomial Distribution (k=2)

Category	1	2
Probability	p_1=p	p_2=1-p
# trials	X_1=x	X_2=n-x

$$\Rightarrow\ P(X_1 = x, X_2 = n-x) = \frac{n!}{x!(n-x)!} p_1^{x_1} p_2^{x_2} = \binom{n}{x} p^x (1-p)^{n-x}$$

Chi-square goodness of fit test

Example 3. Gregor Mendel (1822-1884) was an Austrian monk whose genetic theory is one of the greatest scientific discovery of all time. In his famous experiment with garden peas, he proposed a genetic model that would explain inheritance. In particular, he studied how the shape (smooth or wrinkled) and color (yellow or green) of pea seeds are transmitted through generations. His model shows that the second generation of peas from a certain ancestry should have the following distribution.

	wrinkled-green	wrinkled-yellow	smooth-green	smooth-yellow
Theoretical probabilities	$p_1 = \frac{1}{16}$	$p_2 = \frac{3}{16}$	$p_3 = \frac{3}{16}$	$p_4 = \frac{9}{16}$

n=556

General test:

Test whether the theoretical probability is correct

$$\begin{cases} H_0 : p_1 = p_1^{\,0}, p_2 = p_2^{\,0}, \ldots, p_k = p_k^{\,0} \\ H_a : H_0 \text{ is not true} \end{cases}$$

$$\text{T.S} \quad W_0 = \sum_{i=1}^{k} \frac{(x_i - e_i)^2}{e_i} \overset{H_0}{\sim} \chi^2_{k-1}$$

where x_i is the observed number of observations in category i

e_i is the expected count of the i-th category , $e_i = n \cdot p_i^{\,0}$

At the significance level α, reject H_0 iff $W_0 \geq \chi^2_{k-1,upper,\alpha}$

Note: since this is a large sample test, it is valid when $e_i \geq 5,\ i = 1, \cdots, k$

Solution:

	wrinkled-green	wrinkled-yellow	smooth-green	smooth-yellow

Theoretical probabilities	$p_1=\frac{1}{16}$	$p_2=\frac{3}{16}$	$p_3=\frac{3}{16}$	$p_4=\frac{9}{16}$
Observed count out of 556	X_1=31	X_2=102	X_3=108	X_4=315
Expected counts	$e_1=556\cdot\frac{1}{16}$=34.75	e_2=104.25	e_3=104.25	e_4=312.75

$$\begin{cases} H_0: p_1=\frac{1}{16}, p_2=\frac{3}{16}, p_3=\frac{3}{16}, p_4=\frac{9}{16} \\ H_a: H_0 \text{ is not true} \end{cases}$$

$$\text{T.S} \quad W_0=\sum_{i=1}^{k}\frac{(x_i-e_i)^2}{e_i} \overset{H_0}{\sim} \chi^2_{k-1}$$

$$\approx 0.604 < \chi^2_{3,0.05,upper}=7.815$$

$\therefore$ At significance level 0.05, we cannot reject H_0

Example 4. A classic tale involves four car-pooling students who missed a test and gave as an excuse of a flat tire. On the make-up test, the professor asked the students to identify the particular tire that went flat. If they really did not have a flat tire, would they be able to identify the same tire?

To mimic this situation, 40 other students were asked to identify the tire they would select.
The data are:

Tire	Left front	Right front	Left rear	Right rear
Frequency	11	15	8	6

At α=0.05, please test whether each tire has the same chance to be selected?

Solution:

$$\begin{cases} H_0: p_1=p_2=p_3=p_4=\frac{1}{4} \\ H_a: H_0 \text{ is not true} \end{cases}$$

n= 40, $e_i = n * p_i^0$=10

$$W_0=\sum_{i=1}^{k}\frac{(x_i-e_i)^2}{e_i}=4.6<\chi^2_{3,0.05,upper}=7.815$$

$\therefore$ Fail to reject H_0.

The chi-square goodness of fit test is an extension of the Z-test for one population proportion.

Data: sample size n,
x: successes with probability p
n-x: failures with probability 1-p

1. Two-sided Test on one-population proportion:

$H_0 : p = p_0$

$H_a : p \neq p_0$

TS. $Z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \overset{H_0}{\sim} N(0,1)$

At α, reject H_0 iff $|Z_0| \geq Z_{\alpha/2}$

2. The Chi-square goodness-of-fit test:

$$\begin{cases} H_0 : p_1 = p_0;\ \ p_2 = 1 - p_0 \\ H_a : H_0 \text{ is not true} \end{cases}$$

	Success	Failure
Expected	$p_1 = p_0 \qquad e_1 = np_0$	$p_2 = 1 - p_0 \qquad e_2 = n(1-p_0)$
Observed	$X_1 = x$	$X_2 = n - x$

$$W_0 = \frac{(x - np_0)^2}{np_0} + \frac{[n - x - n(1-p_0)]^2}{n(1-p_0)} = \frac{(x - np_0)^2 (p_0 + 1 - p_0)]^2}{np_0(1-p_0)}$$

$$= \frac{(x - np_0)^2}{np_0(1-p_0)} = \frac{(\frac{x}{n} - p_0)^2}{p_0(1-p_0)/n} = Z_0^{\,2}$$

Recall: If $Z_1, Z_2, \ldots Z_n \overset{iid}{\sim} N(0,1)$, then $W = \sum_1^k Z_i^{\,2} \sim \chi_k^{\,2}$.

When $k = 1$, $W = Z^2 \sim \chi_1^{\,2}$

At α, reject H_0 iff $W_0 \geq \chi^2_{1,\alpha,upper}$

3. Their equivalence:

Let $Z \sim N(0,1)$, then $W = Z^2 \sim \chi_1^2$

$P(|Z| \geq Z_{\alpha/2}) = P(Z^2 \geq Z^2_{\alpha/2}) = \alpha = P(W \geq \chi^2_{1,\alpha,upper})$

$\therefore$ The two tests are identical.