Ch.10 Hypothesis Testing: Categorical Data

Categorical Data: Usually discrete data and falls into categories.

Ex: blood group / gender / color etc...

In Ch.7 & 8 we used to test for continous data. In this chapter we will learn how to test for categorical data.

Two-sample test of hypothesis for binomial proportion, $\hat{p}_1 - \hat{p}_2$:

↳ like Ch.8: $z = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n}+\frac{1}{m}}}$

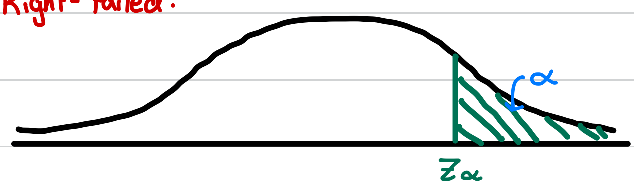
There is two methods for testing two-sample $\hat{p}$.

① Normal Theory Method, Z ② Contingency Table

Both ways give same answer, just like p-value & rejection region methods.

Ch. 7/8 idea recap:

Right-tailed:



$H_0: P_1 - P_2 = 0$ Vs $H_1: P_1 - P_2 > 0$ → right tailed

Left-tailed:

$H_0: P_1 - P_2 = 0$ Vs $H_1: P_1 - P_2 < 0$ → Left tailed

Two-tailed:

$H_0: P_1 - P_2 = 0$ Vs $H_1: P_1 - P_2 \neq 0$ → Two-tailed

1 → Normal method:

Conditions: * $n p^* q^* > 5$ & $m p^* q^* > 5$

* Sample size is large, samples discrete and independant

① $H_0: P_1 - P_2 = 0$ Vs $H_1: P_1 - P_2 \neq 0$ → two - tailed

② Test stat$_1$:
↳ no correction

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sqrt{p^* q^*} * \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$(P_1 - P_2)$ Is always Zero.

↳ $q^* = 1 - p^*$

Test Stat$_2$:
↳ with correction

$$Z_{corr} = \frac{|\hat{P}_1 - \hat{P}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)}{\sqrt{p^* q^*} * \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

pooled proportion, $p^* = \dfrac{X + Y}{n + m}$

$X: \hat{P}_1 = \dfrac{X}{n}$

③ Either p-value method or rejection region method

Let's compare between Ch.8 two test & Ch.10 two test

Ch.8: Given samples are independent & continous and σ_1 & σ_2 unknown. Test for $\bar{X}$:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} \quad \text{where} \quad S_p = \sqrt{\frac{S_1^2(n-1) + S_2^2(m-1)}{n+m-2}}$$

Ch.10: Given samples are discrete, large, and independent. Test for $\hat{p}$:

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p^* q^*}\sqrt{\frac{1}{n} + \frac{1}{m}}} \quad \text{OR} \quad Z_{corr} = \frac{|\hat{P}_1 - \hat{P}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)}{\sqrt{p^* q^*}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$p^* = \frac{X + Y}{n + m}$$

$\therefore$ Ch.8 vs Ch.10

$\Rightarrow$ $\bar{X}_1 - \bar{X}_2$ Vs $\hat{P}_1 - \hat{P}_2$ or $|\hat{P}_1 - \hat{P}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)$

$\Rightarrow$ S_p Vs $\sqrt{p^* q^*}$

$\Rightarrow$ $\sqrt{\frac{1}{n} + \frac{1}{m}}$ Vs $\sqrt{\frac{1}{n} + \frac{1}{m}}$

Standard Error $= \sqrt{\frac{P_1 q_1}{n} + \frac{P_2 q_2}{m}}$

Example 1:

$H_1: p_1 - p_2 \neq 0$

Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult parent reactions. 20 out of a random sample of 200 adults given medication A still had hives 30 minutes after taking the medication. Twelve out of another random sample of 200 adults given medication B still had hives 30 mins after taking the medication. Test using 1% significance level when:

$\alpha = 0.01$ $\qquad \frac{\alpha}{2} = 0.005$

1- No continuity correction applied

2- Apply Zcorrected to your answer

two tailed

1 → ① $H_0: P_1 - P_2 = 0$ Vs $H_1: P_1 - P_2 \neq 0$

② Test stat: $Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{p^* q^*} * \sqrt{\frac{1}{n} + \frac{1}{m}}}$ $\qquad p^* = \dfrac{x+y}{n+m}$

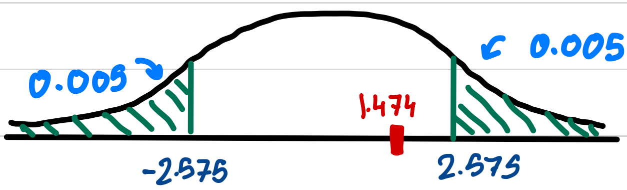
Sample n: $x = 20$ / $n = 200$ | Sample m: $y = 12$ / $m = 200$

$\hat{p}_1 = \dfrac{20}{200}$ $\qquad \hat{p}_1 = 0.1$ $\qquad \hat{p}_2 = \dfrac{12}{200}$ $\qquad \hat{p}_2 = 0.06$

$p^* = \dfrac{20 + 12}{400}$ $\qquad p^* = 0.08$

$q^* = 0.92$

$Z = \dfrac{0.1 - 0.06}{\sqrt{0.08 \times 0.92} * \sqrt{\frac{1}{200} + \frac{1}{200}}} = 1.474$



③ Rejection region method:

$P(Z > a) = 0.005$

$a = 2.575$

z-value outside rejection region

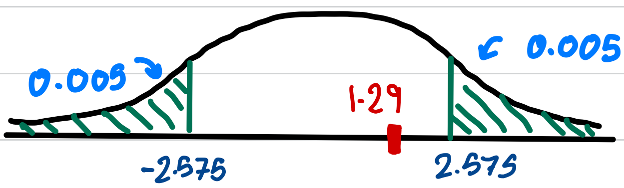
∴ Accept H_0 & Reject H_1

$2 \rightarrow \quad \hat{p}_1 = 0.1 \qquad p^* = 0.08 \qquad n = 200$

$\hat{p}_2 = 0.06 \qquad q^* = 0.92 \qquad m = 200$

$$Z_{corr} = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)}{\sqrt{p^* \, q^*} * \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$Z_{corr} = \frac{|0.1 - 0.06| - \left(\frac{1}{2 \times 200} + \frac{1}{2 \times 200}\right)}{\sqrt{0.08 \times 0.92} * \sqrt{\frac{1}{200} + \frac{1}{200}}} = 1.29$$



$\therefore$ Accept H_0
Reject H_1

Example 2:

A sample of 50 (n) males had 35 (x) smokers. A sample of 100 (m) females had 55 (y) smokers, we want to test if the males proportion is more than the females proportion. The value of the test statistic is? Use Z corrected

① $H_0: P_1 - P_2 = 0$ Vs $H_1: p_1 - p_2 > 0$ right tailed

② Test statistic:

$$Z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)}{\sqrt{p^* q^*} * \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$\hat{p}_1 = \frac{35}{50}$ $\hat{p}_2 = \frac{55}{100}$

$p^* = \frac{35 + 55}{50 + 100} = 0.6$

$\hat{p}_1 = 0.7$ $\hat{p}_2 = 0.55$

$p^* = 0.6$ $q^* = 0.4$

$$Z = \frac{|0.7 - 0.55| - \left(\frac{1}{2 \times 50} + \frac{1}{2 \times 100}\right)}{\sqrt{0.6 \times 0.4} * \sqrt{\frac{1}{50} + \frac{1}{100}}} = 1.59$$

Example 3:

The production of two items A and B is to be evaluated. A sample of 1200 items of type A showed 84 of them were defective. Another sample of 1500 of type B showed that 90 of them were defective. Testing the 1% significance level, can you conclude that the proportions of defective on two types are different? Use Z corrected

p

$P_1 \neq P_2$

Two tailed test

① $H_0: P_1 - P_2 = 0$ $\quad\quad H_1: P_1 - P_2 \neq 0$

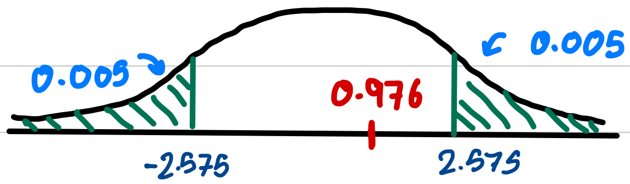
② Test statistic: $\hat{p}_1 = \dfrac{84}{1200}$ $\quad\quad \hat{p}_2 = \dfrac{90}{1500}$

$$p^* = \frac{84 + 90}{1200 + 1500}$$

$\hat{p}_1 = 0.07$ $\quad\quad \hat{p}_2 = 0.06$ $\quad\quad p^* = 0.064$

$q^* = 0.936$

$$Z = \frac{|0.07 - 0.06| - \left(\frac{1}{2 \times 1200} + \frac{1}{2 \times 1500}\right)}{\sqrt{0.064 \times 0.936} * \sqrt{\frac{1}{1200} + \frac{1}{1500}}} = 0.976$$



∴ Accept H_0

Reject H_1

1% :

$1 \to 1\%\ \alpha = 99\%$ C.I $\Rightarrow$ 0.005 B $\quad$ or $\quad$ 0.99 A $\Rightarrow Z_{\frac{\alpha}{2}} = 2.575$

$P(Z > a) = 0.005 \Rightarrow a = 2.575$

Example 4:

Police officers in New York City can stop a driver who is not wearing their seat belt. In Boston, police officers can issue citations to drivers for not wearing their seat belts only if the driver has been stopped for another violation. Data from random samples of females in 2002 is summarized as the following:

City	Drivers	Wearing Seatbelts
Boston	117 n	68 x
New york	220 m	183 y

Is there compelling evidence to conclude a difference in rate of drivers who wear their seat belts in Boston as compared to New York? $\to P_1 - P_2 \neq 0$

Assume continuity correction is applied and use significance level of 0.05

$$\hat{p}_1 = \frac{68}{117} = 0.581 \qquad \hat{p}_2 = \frac{183}{220} = 0.832$$

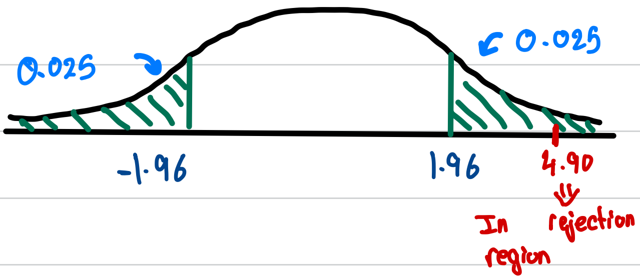
$$p^* = \frac{68 + 183}{117 + 220} = 0.745 \qquad q^* = 0.255$$

① $H_0: P_1 - P_2 = 0$ Vs $H_1: P_1 - P_2 \neq 0$

$$Z_{corr} = \frac{|0.581 - 0.832| - \left(\frac{1}{2 \times 117} + \frac{1}{2 \times 220}\right)}{\sqrt{0.745 \times 0.255} * \sqrt{\frac{1}{117} + \frac{1}{220}}} = 4.90$$

$\alpha = 0.05$ $\quad \frac{\alpha}{2} = 0.025 \to B$

95% C.I $\to$ D



$\Rightarrow Z_{\frac{\alpha}{2}} = 1.96$

or $P(Z > a) = 0.025$ $\qquad a = 1.96$

$\therefore$ Reject H_0 & Accept H_1

Example 5:

A study looked at the effects of OC use on heart disease in women 40 to 44 years of age. The researcher found that among 5000 current OC users at baseline, 13 women developed a myocardial infarction (MI) over a 3 year period, whereas among 10,000 never-OC users, 7 developed an MI over a 3 year period. Assess the statistical significance of the results

$$\hat{p}_1 = \frac{13}{5000} = 0.0026 \qquad \hat{p}_2 = \frac{7}{10\ 000} = 0.0007$$

$$p^* = \frac{13+7}{15000} = 0.0013 \qquad q^* = 0.9987$$

Statistical significance => p-value

① $H_0: p_1 - p_2 = 0$ Vs $H_1: p_1 - p_2 \neq 0$

② $$Z = \frac{|0.0026 - 0.0007| - \left(\frac{1}{2\times 5000} + \frac{1}{2\times 10000}\right)}{\sqrt{0.0013 \times 0.9987} \times \sqrt{\frac{1}{5000} + \frac{1}{10\ 000}}}$$

$Z = 2.77$

③ P-value method $\qquad P(Z > 2.77) = 0.0028$

P-value $= 0.0028 \times 2 = 0.0056$

∴ Results are highly significant

0.0056

0.001 0.01 0.05 0.10

V.H sig. | H.Sig | Sig. | not sig.

Example 6:

The production of two items A and B is to be evaluated. A sample of 1200 items of type A showed 84 of them were defective. Another sample of 1500 of type B showed that 90 of them were defective. Testing the 1% significance level, can you conclude that the proportions of defective on two types are different? Use Z corrected

$$\hat{p}_1 = \frac{84}{1200} \qquad \hat{p}_2 = \frac{90}{1500} \qquad p^* = \frac{84+90}{1200+1500}$$

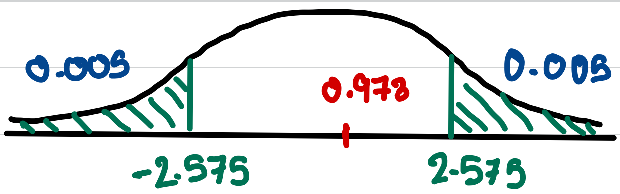
$$\hat{p}_1 = 0.07 \qquad \hat{p}_2 = 0.06 \qquad p^* = 0.0644 \qquad q^* = 0.9356$$

① $H_0: p_1 - p_2 = 0$ Vs $H_1: p_1 - p_2 \neq 0$

two-tailed

②
$$Z = \frac{|0.07 - 0.06| - \left(\frac{1}{2\times 1200} + \frac{1}{2\times 1500}\right)}{\sqrt{0.0644\times 0.9356}\sqrt{\frac{1}{1200}+\frac{1}{1500}}} = 0.973$$

③



∴ Accept H_0
Reject H_1

Propotions of defective are not different.

1→ 1% α => 0.5% $\frac{\alpha}{2}$ => $Z_{\frac{\alpha}{2}} = 2.575$
↳ B

2→ 99% C.I => $Z_{\frac{\alpha}{2}} = 2.575$
↳ D

3→ $P(Z > a) = 0.005$
$a = 2.575$

2 → Contingency - Table method (2×2) :

Ex:

Observed table (2 × 2)

	Right-hand	Left-hand	Total
Males	43 O_{11}	9 O_{12}	52 row margin$_1$
Females	44 O_{21}	4 O_{22}	48 row margin$_2$
Total	87 column margin$_1$	13 column margin$_2$	100 Grand total

⊕ ⇒ 100

⊕ ⇓ 100

Expected table: (2×2)

	Right-hand	Left-hand	Total
Males	$\frac{87\times52}{100} = 45.24$ E_{11}	$\frac{13\times52}{100} = 6.76$ E_{12}	52
Females	$\frac{87\times48}{100} = 41.76$ E_{21}	$\frac{13\times48}{100} = 6.24$ E_{22}	48
Total	87	13	100

Totals always the same for O and E.

Expected :

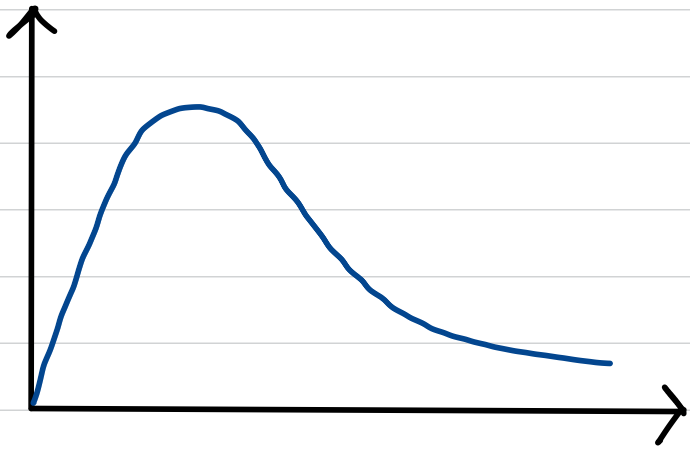
$$\boxed{E = \frac{\text{row sum} \times \text{column sum}}{\text{grand total}}}$$

Testing the hypothesis:

① $H_0: P_1 - P_2 = 0$ Vs $H_1: P_1 - P_2 \neq 0$

② Test stat = Chi-squared test, χ^2.

Chi-squared test, χ^2:



* Skewed to the right.
* All values are positive
* degree of freedom:

$$\boxed{d.f = (\underset{\text{rows}}{R} - 1) * (\underset{\text{columns}}{C} - 1)}$$

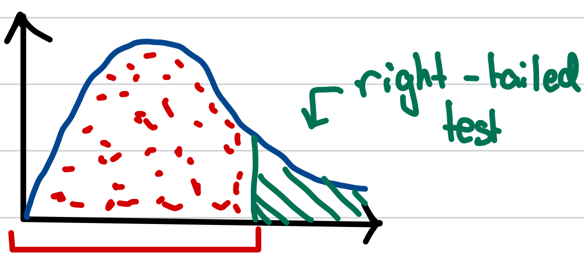
note: d.f for 2×2 contingency table is always 1.

$R = 2$ $C = 2$ $\Rightarrow$ $d.f = (2-1) \times (2-1) = 1 \times 1 = 1$

Chi-squared values is the d.f with the area to the left of a specific point. Like t-distribution, but chi-squared is not symmetrical.

Important note:

* The test is always a right-tailed test. Even if $P_1 \neq P_2$ (H₁), take right tailed always.



↳ use table to find this, then 1 - red = green

How to calculate Chi-squared, χ^2 ?

$$\boxed{* \ \chi^2 = \sum \frac{(O-E)^2}{E}}$$

$$\chi^2 = \frac{(O_{11}-E_{11})^2}{E} + \frac{(O_{12}-E_{12})^2}{E} + \frac{(O_{21}-E_{21})^2}{E} + \frac{(O_{22}+E_{22})^2}{E}$$

$$\boxed{* \ \chi^2_{corr} = \sum \frac{(|O-E| - 0.5)^2}{E}}$$

↳ note that we mainly use χ^2_{corr} for 2×2.

Contingency table & χ^2 test notes:

* Always the expected value is greater than 5.

* Your test is always right-tailed test, the table gives the area to the left.

* The purpose of the contingency table is to summarize a large set of data.

* χ^2_{corr} is called Yates-corrected chi-squared. Usually used for 2x2 table.

* The contingency table is used to determine if the two variables are associated or not.

H_0 : independant variables. Two variables NOT associated

H_1 : dependent variables. Two variables ARE associated.

Example 1:

The following table lists results from an experiment designed to test the ability of dogs to use their extraordinary sense of smell to detect malaria in sample of children's socks. The accompanying information shows the following:

Use both Z & Z_{corr}

Identify the test statistics and the p-value, and then state the conclusion about the null hypothesis.

	Malaria present	Malaria not present	Total
Dog correct	123 O_{11}	131 O_{12}	254
Dog wrong	52 O_{21}	14 O_{22}	66
Total	175	145	320

$$E = \frac{r \times c}{total}$$

138.9 E_{11}	115.09 E_{12}	253.99
36.09 E_{21}	29.9 E_{22}	65.99
174.99	144.9	

expected table

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(123-138.9)^2}{138.9} + \frac{(131-115.09)^2}{115.09} + \frac{(52-36.09)^2}{36.09} + \frac{(14-29.9)^2}{29.9}$$

$$\chi^2 = 1.82 + 2.199 + 7.01 + 8.455 = 19.49$$

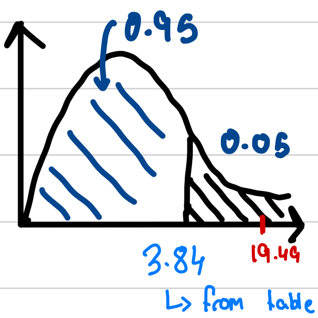
① $H_0: P_1 = P_2$ Vs $H_1: P_1 \neq P_2$] take right tailed

② $\chi^2 = 19.49$

③ Rejection region:

$\alpha = 0.05$

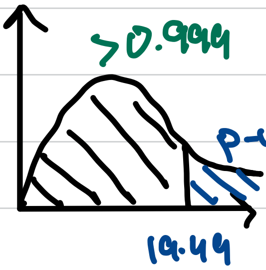
assume $d.f = 1$



∴ Reject H_0 & Accept H_1

↳ Continuation of question.

P-value ? Test stat => your critical point



d.f = 1

∴ p-value < 0.001

$$Z_{corr} = \sum \frac{(|O-E| - 0.5)^2}{E} = \frac{(|123-138.91|-0.5)^2}{138.9} + \frac{(|131-115.09|-0.5)^2}{115.09}$$

$$+ \frac{(|52-36.09|-0.5)^2}{36.09} + \frac{(|14-29.91|-0.5)^2}{29.91}$$

$$Z_{corr} = 1.71 + 2.06 + 6.58 + 7.939 = 18.29$$

Example 2: ↳ Use Z_{corr}

$E = \frac{r \times c}{total}$

$\alpha = 0.001$

Suppose we want to know if the rate of smoking in males is different from females in a sample of 203 Jordanians. The observed values set as the following:

	Smoker	Non-smoker	Total
Males	72 O_{11}	44 O_{12}	116
Females	34 O_{21}	53 O_{22}	87
Total	106	97	203

E:

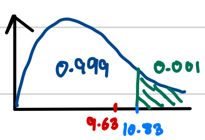
60.57 E_{11}	55.43 E_{12}
45.43 E_{21}	41.57 E_{22}

① $H_0: P_1 = P_2$ Vs $H_1: P_1 \neq P_2$ ↳ take right tailed test

② Test stat = $\chi^2 = \sum \frac{(|O-E|-0.5)^2}{E} = \frac{(|72-60.57|-0.5)^2}{60.57} + \frac{(|44-55.43|-0.5)^2}{55.43}$

$$+ \frac{(|34-45.43|-0.5)^2}{45.43} + \frac{(|53-41.57|-0.5)^2}{41.57} = 1.97 + 2.155 + 2.63 + 2.874 = \boxed{9.63}$$

③ Rejection region:



∴ Accept H_0. Reject H_1

Example 3: $E = \frac{r \times c}{total}$

The following table lists the number of females taken in a study to see whether there's an association between breast cancer and having first child after age 30. Assess the following data for statistical significance, using a contingency table approach.

↳ find p-value

O:

			Total
	683	2537	3220
	1498	8747	10245
Total	2181	11284	13466

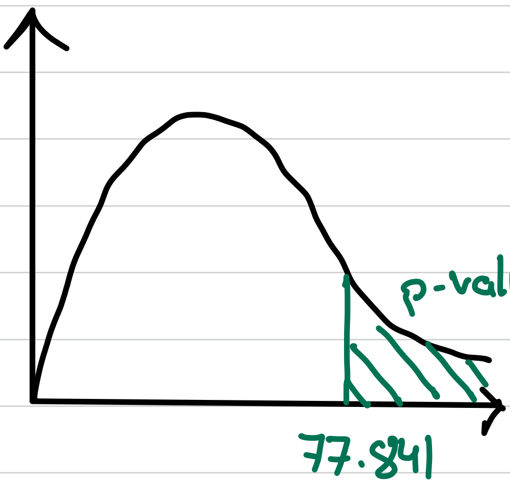
E:

521.56	2698.4
1659.4	8585.56

$$\chi^2_{corr} = \sum \frac{(|O-E| - 0.5)^2}{E} = \frac{(|683-521.56|-0.5)^2}{521.56} + \frac{(|2537-2698.4|-0.5)^2}{2698.4}$$

$$+ \frac{(|1498-1659.4|-0.5)^2}{1659.4} + \frac{(|8747-8585.56|-0.5)^2}{8585.56}$$

$$\chi^2_{corr} = 49.66 + 9.594 + 15.60 + 3.017 = 77.841$$



d.f = 1

p-value = 1 - 0.999 = 0.001

∴ Very highly statistically significant

0.001	0.01	0.05	0.1
V.H sig	H. sig	sig.	Not sig.

Example 4:

$\alpha = 0.05$

Assess statistical significance for the following? Test the hypothesis?

O:

13 O_{11}	4987 O_{12}	5000
7 O_{21}	9993 O_{22}	10000
20	14980	15000

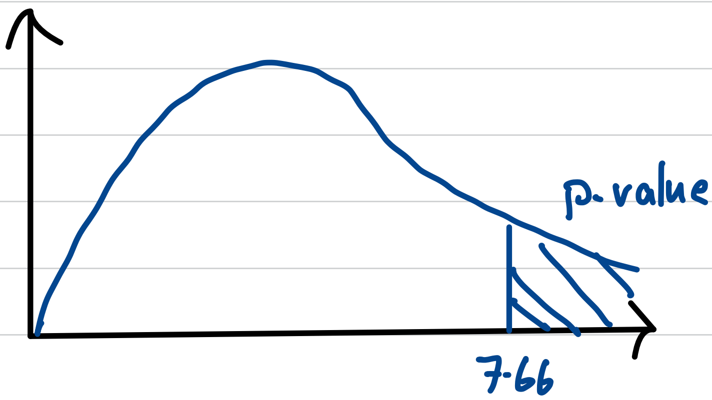
E:

6.67	4993.33
13.33	9986.67

$$\chi^2_{corr} = \sum \frac{(|O-E| - 0.5)^2}{E} = 5.10 + 0.0068 + 0.0034 + 2.55$$

$$\chi^2_{corr} = 7.66$$

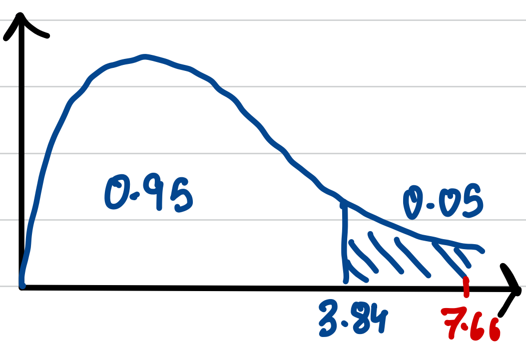
p-value = 0.005



∴ Highly statistically significant.

$H_0: P_1 = P_2$ Vs $H_1: P_1 \neq P_2$

↳ right tailed



$\chi^2 = 7.66$

∴ Reject H_0

Accept H_1

Important note for χ^2 test :

$\chi^2 = \frac{(O-E)^2}{E}$. ∴ The larger (O−E), the larger the χ^2.

The larger the χ^2. The less area to the right, so less p-value. Less p-value, the more the statistical significance.

∴ the bigger the difference b/w O & E, the more the statistical significance

H_0: no statistical significance / H_1 : statistical significance

R×C contingency table: (for 2×2 we use Yates-corrected. Here χ^2 only)

How to find error? $E = \frac{R \times C}{E}$

What Chi-square do I use? ALWAYS χ^2. NOT χ^2_{corr}

Degree of freedom? $\underset{\text{no of rows}}{(R-1)} * \underset{\text{no of columns}}{(C-1)}$

Conditions for R × C :

1→ Always the expected value is greater than 1.

2→ For every 5 cells, only one value of E less than 5 is allowed.

Ex:

5	9	121
1	17	2

Doesn't Satisfy

5	9	121
7	17	2

Satisfy

Example 1: Assess the statistical significance in 300 people. Given the following information: Assuming $\alpha = 0.05$. Test the hypothesis?

Table of Observed Values

Qualification / Marital Status	Middle School	High School	Bachelor's	Master's	Ph.D	Total
Never married	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
Total	39	90	84	54	33	300

Qualification / Marital Status	Middle School	High School	Bachelor's	Master's	Ph.D
Never Married	$\frac{90 \times 39}{300} = 11.7$	$\frac{90 \times 90}{300} = 27$	25.2	16.2	9.9
Married	19.5	45	42	27	16.5
Divorced	3.9	9	8.4	5.4	3.3
Widowed	3.9	9	8.4	5.4	3.3

$$x^2 = \frac{(O-E)^2}{E} = \frac{(18-11.7)^2}{11.7} + \cdots\cdots + \frac{(3-3.3)^2}{3.3}$$

$x^2 : 23.57$

$\therefore$ p-value = 0.025

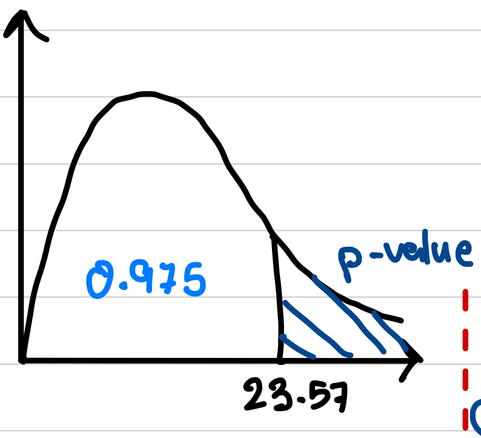
0.025

0.001 0.01 0.05 0.1

v.H sig | H.sig | sig | not sig

$\therefore$ Results are significant

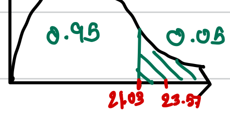
$\therefore$



H_0: Independent (not associated) Vs H_1: dependent (associated)

(2) $x^2 = 23.57$

(3)



d.f = $(4-1) \times (5-1) = 12$

$\therefore$ Reject H_0. Accept H_1

↳ makes sense. p-value < alpha. Reject H_0

Also significant Reject H_0

Also significant so is H_1 so reject H_0.

Variables are associated

Example 2:

Assume $\alpha = 0.05$

Assess the following statistical significance. And find wether the variables are associated or not:

O:

Case–control status	Age at first birth: <20	20–24	25–29	30–34	≥35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10,245
Total	1742	5638	3904	1555	626	13,465

E:

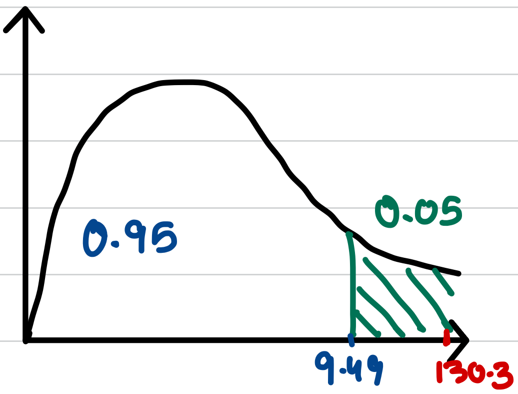
416.6	1348.3	933.6	371.9	149.7
1325.4	4289.7	2970.4	1183.1	476.3

① H_0: independent / Not associated Vs H_1: dependent / associated

② $\chi^2 = \frac{(O-E)^2}{E} = 130.3$

∴ Reject H_0
Accept H_1
They are associated

③



$d.f = (2-1) \times (5-1) = 4$

P-value:
$1 - 0.999 = < 0.001$

0.001 | 0.01 | 0.05 | 0.1

v.H sig | H.sig | sig | not sig

∴ Results are very highly significant

Example 3:

Determine to the 5 % significance level whether school and grade are dependent

O:

		Grade			Totals
		A	B	C	
School	X	18	12	20	50
	Y	26	12	32	70
Totals		44	24	52	120

E:

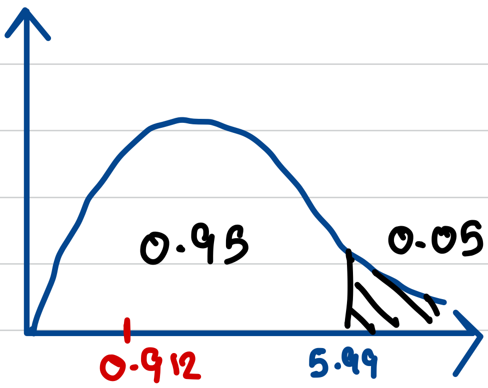
18.33	10	21.67
25.67	14	30.33

① H_0: independent / not associated   Vs H_1: dependent / associated

② $\chi^2 = \dfrac{(O-E)^2}{E} = 0.005941 + 0.4 + 0.129 + 0.004242 + 0.286 + 0.0872$

$$\boxed{\chi^2 = 0.912}$$

③



$d.f = (2-1) \times (3-1) = 2$

∴ Accept H_0
Reject H_1

Variables are independent & not associated

Chi-Square Goodness-of-Fit test:

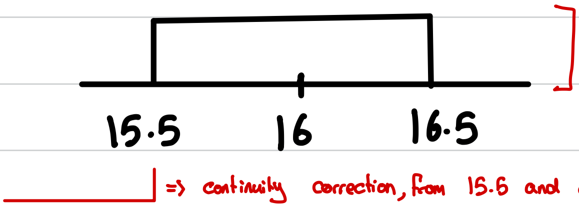
=> Approximation of discrete random variable to continuous random variable.

Discrete (Binomial) $\xrightarrow{\text{Continuity correction}}$ Continuous (normal)

How to carry out continuity correction?

① Ex: $P(X < 16) \Rightarrow P(X \leq 15.5)$

↳ after continuity correction

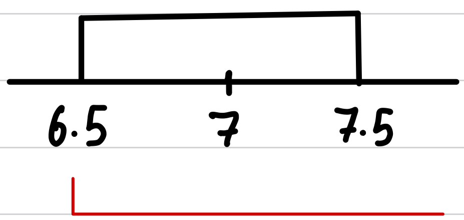


=> continuity correction, from 15.5 and going down, so ≤ 15.5

Less than 16 → outside the box => take values outside box to the left

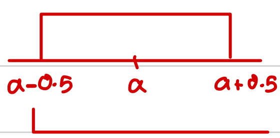
↳ cuz less than

② Ex: $P(x \geq 7) \Rightarrow P(x \geq 6.5)$



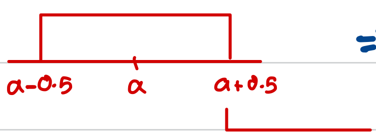
Greater than or equal → since equal take all box => Since greater than so take to the right.

① $P(x \geq a)$



$\Rightarrow P(x \geq a - 0.5)$

② $P(x > a)$



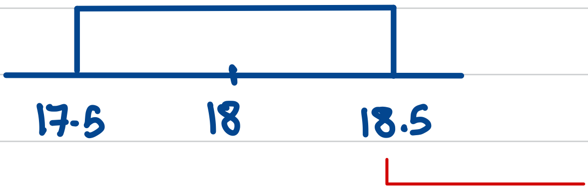
$\Rightarrow P(x \geq a + 0.5)$

Some quick examples of continuity correction:

① $P(X > 18)$ [discrete] $\rightarrow$ $P(X \geq 18.5)$ [continuous]

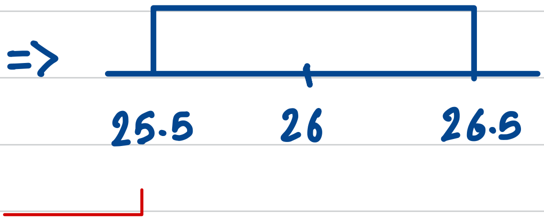
normal

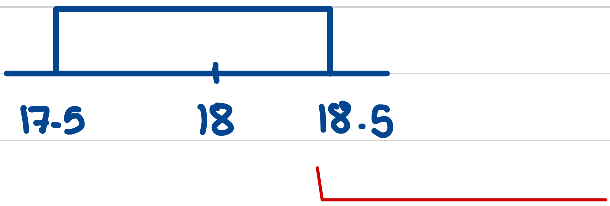
$\Rightarrow$



② $P(18 < X < 26) \Rightarrow X < 26$ & $X > 18$

$\Rightarrow$

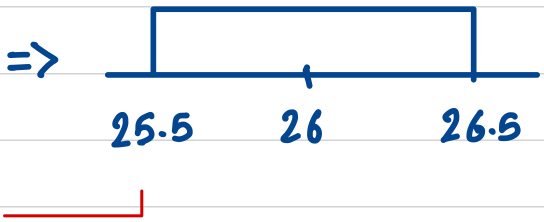


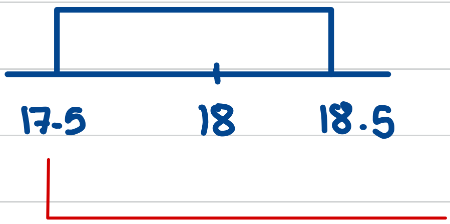


$\Rightarrow P(18.5 \leq X \leq 25.5)$

③ $P(18 \leq X < 26) \Rightarrow X \geq 18$ & $X < 26$

$\Rightarrow$





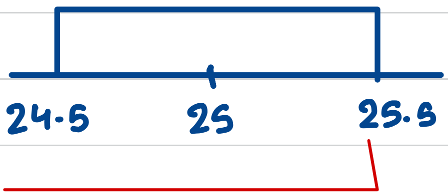
$\Rightarrow P(17.5 \leq X \leq 25.5)$

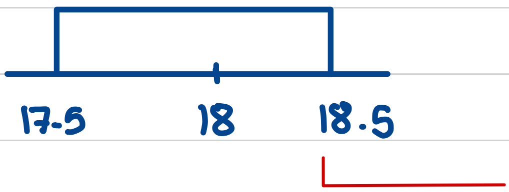
④ $P(18 < X \leq 25) \Rightarrow X > 18$ & $X \leq 25$

outside box

inside box

$\Rightarrow$





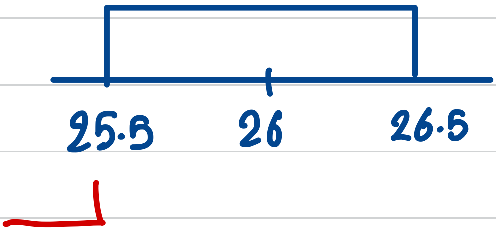
$\Rightarrow P(18.5 \leq X \leq 25.5)$

Example 1: If the $\mu = 20$ & $\sigma^2 = 16$

① Given that $P(x < 26)$ is discrete, find $P(x < 26)$

Discrete $\rightarrow$ continuous $\Rightarrow$ continuity correction.

$X \sim N(x, \sigma^2)$



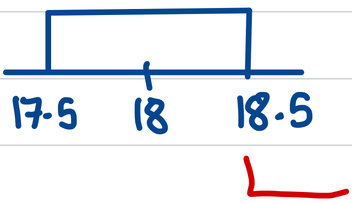
$\therefore P(x \leq 25.5)$

$Z = \dfrac{X - \mu}{\sigma}$

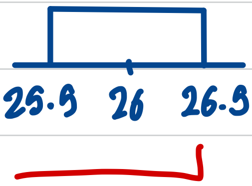
$$P\left(Z < \frac{25.5 - 20}{4}\right) = P(Z < 1.38) = \boxed{0.9162}$$

② Given that $P(18 < x \leq 26)$ is discrete, find x ?

$X > 18$ & $x \leq 26$



&



$\therefore P(18.5 \leq X \leq 26.5)$

$$\Rightarrow P\left(\frac{18.5 - 20}{4} < Z < \frac{26.5 - 20}{4}\right)$$

$$\Rightarrow P(-0.375 < Z < 1.625) \Rightarrow P(Z < 1.63) - P(Z > 0.38)$$

$$\Rightarrow 0.9484 - 0.3520 = \boxed{0.5964}$$

χ^2 Goodness-of-fit example: (7)

EXAMPLE 10.46 **Hypertension** Diastolic blood-pressure measurements were collected at home in a community-wide screening program of 14,736 adults ages 30–69 in East Boston, Massachusetts, as part of a nationwide study to detect and treat hypertensive people [6]. The people in the study were each screened in the home, with two measurements taken during one visit. A frequency distribution of the mean diastolic blood pressure is given in Table 10.20 in 10-mm Hg intervals.

We would like to assume these measurements came from an underlying normal distribution because standard methods of statistical inference could then be applied on these data as presented in this text. How can the validity of this assumption be tested?

H_0: normal adequate Vs H_1: normal not adequate

TABLE 10.20 Frequency distribution of mean diastolic blood pressure for adults 30–69 years old in a community-wide screening program in East Boston, Massachusetts

Group (mm Hg)	Observed frequency	Expected frequency	Group	Observed frequency	Expected frequency
<50	57	69.0	≥80, <90	4604	4538.6
≥50, <60	330	502.5	≥90, <100	2119	2545.9
≥60, <70	2132	2018.4	≥100, <110	659	740.4
≥70, <80	4584	4200.9	≥110	251	120.2
			Total	14,736	14,736

We want to test the assumption that the data came from a normal-distribution.

How to find expected?

$E = P(X < 50) \times \underline{\text{Total}}$ Observed frequency] for first row

↳ discrete

$\therefore E = P(X \leq 49.5) \times 14736$ ∴

$\Rightarrow E = P(Z > 2.60) \times 14736$

$\Rightarrow E = 0.0047 \times 14736 = \boxed{69}$

$\bar{X} = \dfrac{\sum fx}{n \to \sum f}$

∴

$S^2 = \dfrac{\sum x^2}{n-1} - \dfrac{(\sum x)^2}{n(n-1)}$

$\Rightarrow S^2 = \dfrac{\sum fx^2}{n-1} - \dfrac{(\sum fx)^2}{n(n-1)}$

↳ μ point estimate

$\bar{X} = 80.68$

↳ σ point estimate

$S = 12$

Lets find second row:

$E = P(50 \leq X < 60) \times O_t = P(49.5 \leq X \leq 59.5) \times O_t$

$= [P(Z > 1.765) - P(Z > 2.60)] \times O_t = (0.0388 - 0.0047) \times 14736$

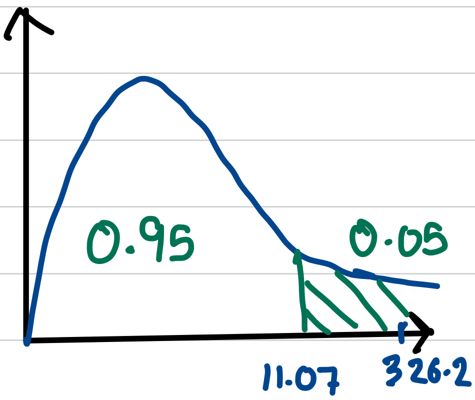
$= 502.5$

Continuation:

Test stat: $\chi^2 = \sum \frac{(O-E)^2}{E}$

$$\chi^2 = \frac{(57-69)^2}{69} + \cdots\cdots + \frac{(251-120.2)^2}{120.2}$$

$\chi^2 = 326.2$ Assume $\alpha = 0.05$



$d.f = \underset{\text{groups}}{8} - \underset{\bar{x},\, s}{2} - 1 = 5$

$\therefore$ Reject H_0 & Accept H_1

Meaning, normal method does not provide an adequate fit to the data.

* Important:

Degree of freedom for Chi-squared goodness-of-fit:

$$\boxed{d.f = g - k - 1}$$

g: number of groups / categories
↳ If $E < 5$ then for this category we join it with another category and it counts as one.

k: number of estimated parameters
↳ point estimates => $\bar{x}$, s, $\hat{p}$

Chi-squared goodness-of-fit notes:

* We study the fit of the test to the data, being H_0.

* Expected calculation:

Expected
- continuity correction
- Find probability
- probability * grand total

* Test Stat: $\chi^2 = \frac{(O-E)^2}{E}$

* Degree of freedom: $g - k - 1$ / group - point estimators - 1

* The sum of the expected should equal the sum of observed.

Example 2: The mean weights of a sample of 200 patients is 52 kg and the standard deviation is 3 kg.
↳ $\bar{x}$ ↳ s

Weight	$w < 45$	$45 \leq w < 50$	$50 \leq w < 55$	$55 \leq w < 60$	$w \geq 60$
frequency ↳ observed frequency	12	44	82	53	9
] E	1.24	39.4	118.7	39.42	1.24

Given that $\bar{x} = 52$ and $s = 3$. We would like to assume that these measurements came from the normal distribution. How can the validity of this assumption be tested?

H_0: normal distribution valid Vs H_1: normal distribution not valid.

$E = P(x < a) \times O_t$ ① $P(x < 45) \times O_t$

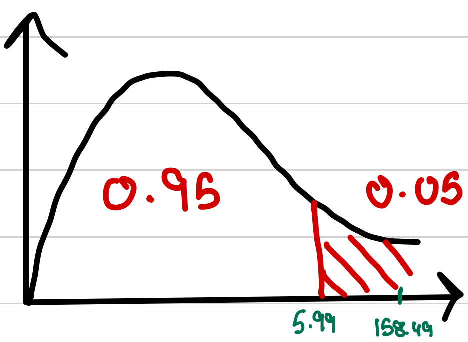
$O_t = 200 \Rightarrow P(X \leq 44.5) \times O_t \Rightarrow P(z > 2.5) \times O_t$

$E = 0.0062 \times 200 = 1.24$

Last $E \Rightarrow 200 - (1.24 + 118.7 + 39.42 + 1.24) = 39.4$

Test stat $\Rightarrow \chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(12-1.24)^2}{1.24} + \dots\dots + \frac{(9-1.24)^2}{1.24}$

$\therefore \chi^2 = 158.49$ d.f $= 5 - 2 - 1 = 2$



∴ Reject H_0 & Accept H_1
normal distribution not adequate.