FISHER EXACT TEST

- When we interested in whether two sample came from the same population, we may examine if they have the same population mean or variance?

-When examining the whether the variances are the same:

F-test=$\frac{S_1^2}{S_2^2}$~F_{n_1,n_2}, the F distribution with degree of freedom n_1 and n_2

-When examining the whether the means are the same:

T test

FISHER EXACT TEST

- If there are extreme values or violations of the normality assumption for the data, then the Fisher's exact test may be more suitable for determining whether two samples come from the same population.

FISHER EXACT TEST

- Consider there a 2X2 table with n<20.

	Sample		
treatment	1	2	Sum
1	A	B	$A+B$
2	C	D	$C+D$
	$A+C = n_1$	$B+D = n_2$	$N=n_1+n_2$

FISHER EXACT TEST

Therefore the exact probability is

$$P = \frac{\binom{A+C}{A}\binom{B+D}{D}}{\binom{N}{A+B}}$$

$$= \frac{(A+B)!\,(C+D)!\,(A+C)!\,(B+D)!}{N!\,A!\,B!\,C!\,D!}$$

FISHER EXACT TEST

P- value :

Sample

Treatment	1	2	
1	18	2	20
2	11	9	20
	29	11	

		P
20	0	0.00007
9	11	
19	1	0.0016
10	10	
18	2	0.0138
11	9	
17	3	0.06212
12	8	
16	4	0.16246
13	7	
15	5	0.25994
14	6	

		P
14	6	0.25994
15	5	
13	7	0.16246
16	4	
12	8	0.06212
17	3	
11	9	0.0138
18	2	
10	10	0.0016
19	1	
9	11	0.00007
20	0	

Therefore : p-value

$$= \begin{cases} (0.00007 + 0.0016 + 0.0138) \times 2 & \text{双尾} \\ 0.00007 + 0.0016 + 0.0138 & \text{單尾} \end{cases}$$

FISHER EXACT TEST

- Two products were evaluated by a total of 19 individuals, and the results are recorded as follows:

Product

	A	B	total
like	5	4	9
dislike	1	9	10
total	6	13	19

p-value

$$\text{Probability}\left(\begin{array}{|cc|}\hline 5 & 4 \\ 1 & 9 \\ \hline\end{array}\right) = 0.0464$$

$$\begin{array}{|cc|}\hline 6 & 3 \\ 0 & 10 \\ \hline\end{array} =$$

- H_0: Two products are favored by the same preferences

FISHER EXACT TEST

Drug manufacturers claim they have developed a new pesticide that is more powerful than the old ones

H_0: New and old pesticide have the same effect

Pesticide

Number of	New	old	Sum
live insects	350	100	450
Dead insects	50	20	70
Sum	400	120	520 Total

FISHER EXACT TEST

When $n \geq 20$

$\hat{p}_1 = A/n_1$

$\hat{p}_2 = B/n_2$

$\hat{p} = A+B/N$

Under $H_0: p_1 = p_2$ v.s. $H_a: \hat{p}_1 \neq \hat{p}_2$

Hence

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p(1-p)\frac{n_1+n_2}{n_1 n_2}}}$$

CHI-SQUARE OF HOMOGENEITY

- If two independent categorical datasets cannot be organized into a 2×2 table, the Chi-square test for homogeneity can be used to examine whether they come from the same population.

H_0: Two population came from the same population

(or there is no treatment effect)

CHI-SQUARE OF HOMOGENEITY

		Treatment 1	Treatment 2	Sum
category	1	O_{11}	O_{21}	R_1
	2	O_{12}	O_{22}	R_2
	$\vdots$			$\vdots$
	k	O_{1k}	O_{2k}	R_k
Sum		C_1	C_2	n

O_{ij} : observation for ith. category and jth. treatment

$R_i = \sum_{j=1}^{2} O_{ij}$, $i \in \{1, 2, \cdots, k\}$

$C_j = \sum_{i=1}^{k} O_{ij}$, $j \in \{1, 2\}$

CHI-SQUARE OF HOMOGENEITY

$$\chi = \sum_{i=1}^{2} \sum_{j=1}^{K} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{k-1}, \text{ reject } H_0$$

Correctness of continuous

$$\chi_c = \sum_{i=1}^{2} \sum_{j=1}^{K} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

CHI-SQUARE OF HOMOGENEITY

- If we are wondering whether males with higher body weight have a higher heart rate than those with normal weight.

$$E_{ij} = \underbrace{\frac{R_i}{n}}_{= P_i} \cdot C_j$$

number

heart	normal	overweight	sum
<72	51 (45.6)	43 (48.4)	94
72-80	64 (62.6)	65 (66.4)	129
80>	50 (56.8)	67 (60.2)	117
Sum	165	175	340

CHI-SQUARE OF HOMOGENEITY

heart	number: normal	number: overweight	sum
<72	51 (45.6)	43 (48.4)	94
72-80	64 (62.6)	65 (66.4)	129
80>	50 (56.8)	67 (60.2)	117
sum	165	175	340

$$\frac{(51-45.6)^2}{45.6} + \frac{(64-62.6)^2}{62.6} + \frac{(50-56.8)^2}{56.8}$$

$$+ \frac{(43-48.4)^2}{48.4} + \cdots = 2.885 = \chi$$

$$\chi = 2.885 < \chi^2_{0.05,(\underset{3}{\underset{\parallel}{k}}-1)} = 5.991$$

$\frac{94}{340} \cdot 165$

$\frac{129}{340} \cdot 175$

CHI-SQUARE OF HOMOGENEITY

- Small sample correction:

Under the small sample Wilks proposed a statistics obtained from the likelihood ratio approach:

$$G = 2\left[\sum_{i}^{K}\sum_{j}^{C} O_{ij}\log O_{ij} - \sum_{i}^{k}\sum_{j}^{c} O_{ij}\log E_{ij}\right]$$

$$G \sim \chi^2_{(k-1)(c-1)}$$

CHI-SQUARE OF HOMOGENEITY

- If we are interested in determining whether the number of individuals who violate traffic laws differs between those over and under the age of 25, we can analyze the data as follows

# of violation (per year)	Age: <25	Age: 25<	
0	23 (30.8)	38 (30.2)	61
1	18 (24.7)	31 (24.3)	49
2	45 (35.4)	25 (34.6)	70
3	14 (9.1)	4 (8.9)	18
	100	98	198

$\frac{61}{198} \cdot 100$

$\frac{49}{198} \cdot 98$

CHI-SQUARE OF HOMOGENEITY

- 〈 If we are interested in determining whether the number of individuals who violate traffic laws differs between those over and under the age of 25, we can analyze the data as follows

# of violation (per year)	Age <25	Age 25<	
0	23 (30.8)	38 (30.2)	61
1	18 (24.7)	31 (24.3)	49
2	45 (35.4)	25 (34.6)	70
3	14 (9.1)	4 (8.9)	18
	100	98	198

$$23(\log 23 - \log 30.8)$$

$$+\ 38(\log 38 - \log 30.2)$$

$$\vdots \qquad \vdots$$

$$+\ 4(\log 4 - \log 8.9)$$

$$= 18.744 > \chi^2_{0.05,\,(2-1)(3-1)} = 7.815$$

KOLMOGOROV-SMIRNOV

- Kolmogorov-Smirnov also can be adopted under the two samples problem.
- Assume there are two groups with their independent outcomes.

Moreover, $x_1, x_2, \dots\ x_m$ are the data from the group 1; $y_1, y_2, \dots\ y_n$ are the data from the group 2.

$$H_0\colon\ F_X(\cdot) =\ F_Y(\cdot)$$

- Denote that $S_X(t)$ and $S_Y(t)$ are the empirical CDF for X and Y, respectively.

- The testing statistics can be obtained by $D = \max_t |\ S_X(t) - S_Y(t)|$

KOLMOGOROV-SMIRNOV

- Example

If the course is conducted by two teachers, the quality of the course is evaluated by a total of 50 students before the end of the semester. A comprehensive evaluation using 5 levels (very bad(1), bad(2), normal(3), good(4), very good(5)) is designed for sampling.

Teacher \ evaluation	1	2	3	4	5	Sum
A	1	10	16	15	8	50
B	2	4	10	24	10	50

H_0: The popularity of the two teachers is the same.

KOLMOGOROV-SMIRNOV

	Teacher (A)		Teache (B)				
evaluation (x)	counts	cumulative	counts	cumulative	$S_A(x)$	$S_B(x)$	D
1	1	1	2	2	1/50	2/50	1/50
2	10	11	4	6	11/50	6/50	5/50
3	16	27	10	16	27/50	16/50	11/50
4	15	42	24	40	42/50	40/50	2/50
5	8	50	10	50	1	1	0

Therefore

$$D = \max_x |S_A(x) - S_B(x)| = 11/50 = 0.22 < D_{0.05/2,\,50} = \frac{1.92}{\sqrt{50}} = 0.272$$

KOLMOGOROV-SMIRNOV

- If we want to examine the amount of lead in the soil between two areas, A and B.

A: 3.25,3.92,3.81,3.41,2.91,4.18,2.43,3.24,3.26

B: 3.91,4.12,4.52,4.62,4.84,4.87,5.50,4.06,4.72

H_0: The amount of lead in the soil are the same between A and B.

KOLMOGOROV-SMIRNOV

treatment	A	B	S_A	S_B	D
(2.4, 2.8)	1	0	1/10	0/10	1/10
(2.8, 3.2)	1	0	2/10	0/10	2/10
(3.2, 3.6)	3	0	5/10	0/10	5/10
(3.6, 4.0)	2	0	7/10	0/10	7/10
(4.0, 4.4)	3	3	10/10	3/10	7/10
(4.4, 4.8)	0	2	1	5/10	5/10
(4.8, 5.2)	0	3	1	8/10	2/10
(5.2, 5.6)	0	2	1	1	0

Table ST8. Critical Values of the Kolmogorov–Smirnov Test Statistic for Two Samples of Equal Size[a]

One-Sided Test: $\alpha =$	0.10	0.05	0.025	0.01	0.005	$\alpha =$	0.10	0.05	0.025	0.01	0.005
Two-Sided Test: $\alpha =$	0.20	0.10	0.05	0.02	0.01	$\alpha =$	0.20	0.10	0.05	0.02	0.01
$n = 3$	2/3	2/3				$n = 20$	6/20	7/20	8/20	9/20	10/20
4	3/4	3/4	3/4			21	6/21	7/21	8/21	9/21	10/21
5	3/5	3/5	4/5	4/5	4/5	22	7/22	8/22	8/22	10/22	10/22
6	3/6	4/6	4/6	5/6	5/6	23	7/23	8/23	9/23	10/23	10/23
7	4/7	4/7	5/7	5/7	5/7	24	7/24	8/24	9/24	10/24	11/24
8	4/8	4/8	5/8	5/8	6/8	25	7/25	8/25	9/25	10/25	11/25
9	4/9	5/9	5/9	6/9	6/9	26	7/26	8/26	9/26	10/26	11/26
10	4/10	5/10	6/10	6/10	7/10	27	7/27	8/27	9/27	11/27	11/27
11	5/11	5/11	6/11	7/11	7/11	28	8/28	9/28	10/28	11/28	12/28
12	5/12	5/12	6/12	7/12	7/12	29	8/29	9/29	10/29	11/29	12/29
13	5/13	6/13	6/13	7/13	8/13	30	8/30	9/30	10/30	11/30	12/30
14	5/14	6/14	7/14	7/14	8/14	31	8/31	9/31	10/31	11/31	12/31
15	5/15	6/15	7/15	8/15	8/15	32	8/32	9/32	10/32	12/32	12/32
16	6/16	6/16	7/16	8/16	9/16	34	8/34	10/34	11/34	12/34	13/34
17	6/17	7/17	7/17	8/17	9/17	36	9/36	10/36	11/36	12/36	13/36
18	6/18	7/18	8/18	9/18	9/18	38	9/38	10/38	11/38	13/38	14/38
19	6/19	7/19	8/19	9/19	9/19	40	9/40	10/40	12/40	13/40	14/40
						Approximation for $n > 40$:	$\frac{1.52}{\sqrt{n}}$	$\frac{1.73}{\sqrt{n}}$	$\frac{1.92}{\sqrt{n}}$	$\frac{2.15}{\sqrt{n}}$	$\frac{2.30}{\sqrt{n}}$

Source: Adapted by permission from Tables 2 and 3 of Z. W. Birnbaum and R. A. Hall, Small sample distributions for multisample statistics of the Smirnov type, *Ann. Math. Stat.* 31 (1960), 710–720.

[a] This table gives the values of $D^+_{n,n,\alpha}$ and $D_{n,n,\alpha}$ for which $\alpha \geq P\{D^+_{n,n} > D^+_{n,n,\alpha}\}$ and $\alpha \geq P\{D_{n,n} > D_{n,n,\alpha}\}$ for some selected values of n and α.

Table ST9. Critical Values of the Kolmogorov–Smirnov Test Statistic for Two Samples of Unequal Size[a]

One-Sided Test:	$\alpha =$	0.10	0.05	0.025	0.01	0.005
Two-Sided Test:	$\alpha =$	0.20	0.10	0.05	0.02	0.01
$N_1 = 1$	$N_2 = 9$	17/18				
	10	9/10				
$N_1 = 2$	$N_2 = 3$	5/6				
	4	3/4				
	5	4/5	4/5			
	6	5/6	5/6			
	7	5/7	6/7			
	8	3/4	7/8	7/8		
	9	7/9	8/9	8/9		
	10	7/10	4/5	9/10		
$N_1 = 3$	$N_2 = 4$	3/4	3/4			
	5	2/3	4/5	4/5		
	6	2/3	2/3	5/6		
	7	2/3	5/7	6/7	6/7	
	8	5/8	3/4	3/4	7/8	
	9	2/3	2/3	7/9	8/9	8/9
	10	3/5	7/10	4/5	9/10	9/10
	12	7/12	2/3	3/4	5/6	11/12
$N_1 = 4$	$N_2 = 5$	3/5	3/4	4/5	4/5	
	6	7/12	2/3	3/4	5/6	5/6
	7	17/28	5/7	3/4	6/7	6/7
	8	5/8	5/8	3/4	7/8	7/8
	9	5/9	2/3	3/4	7/9	8/9
	10	11/20	13/20	7/10	4/5	4/5
	12	7/12	2/3	2/3	3/4	5/6
	16	9/16	5/8	11/16	3/4	13/16
$N_1 = 5$	$N_2 = 6$	3/5	2/3	2/3	5/6	5/6
	7	4/7	23/35	5/7	29/35	6/7
	8	11/20	5/8	27/40	4/5	4/5
	9	5/9	3/5	31/45	7/9	4/5
	10	1/2	3/5	7/10	7/10	4/5
	15	8/15	3/5	2/3	11/15	11/15
	20	1/2	11/20	3/5	7/10	3/4

Table ST9. *(Continued)*

One-Sided Test:	$\alpha =$	0.10	0.05	0.025	0.01	0.005
Two-Sided Test:	$\alpha =$	0.20	0.10	0.05	0.02	0.01
$N_1 = 6$	$N_2 = 7$	23/42	4/7	29/42	5/7	5/6
	8	1/2	7/12	2/3	3/4	3/4
	9	1/2	5/9	2/3	13/18	7/9
	10	1/2	17/30	19/30	7/10	11/15
	12	1/2	7/12	7/12	2/3	3/4
	18	4/9	5/9	11/18	2/3	13/18
	24	11/24	1/2	7/12	5/8	2/3
$N_1 = 7$	$N_2 = 8$	27/56	33/56	5/8	41/56	3/4
	9	31/63	5/9	40/63	5/7	47/63
	10	33/70	39/70	43/70	7/10	5/7
	14	3/7	1/2	4/7	9/14	5/7
	28	3/7	13/28	15/28	17/28	9/14
$N_1 = 8$	$N_2 = 9$	4/9	13/24	5/8	2/3	3/4
	10	19/40	21/40	23/40	27/40	7/10
	12	11/24	1/2	7/12	5/8	2/3
	16	7/16	1/2	9/16	5/8	5/8
	32	13/32	7/16	1/2	9/16	19/32
$N_1 = 9$	$N_2 = 10$	7/15	1/2	26/45	2/3	31/45
	12	4/9	1/2	5/9	11/18	2/3
	15	19/45	22/45	8/15	3/5	29/45
	18	7/18	4/9	1/2	5/9	11/18
	36	13/36	5/12	17/36	19/36	5/9
$N_1 = 10$	$N_2 = 15$	2/5	7/15	1/2	17/30	19/30
	20	2/5	9/20	1/2	11/20	3/5
	40	7/20	2/5	9/20	1/2	
$N_1 = 12$	$N_2 = 15$	23/60	9/20	1/2	11/20	7/12
	16	3/8	7/16	23/48	13/24	7/12
	18	13/36	5/12	17/36	19/36	5/9
	20	11/30	5/12	7/15	31/60	17/30
$N_1 = 15$	$N_2 = 20$	7/20	2/5	13/30	29/60	31/60
$N_1 = 16$	$N_2 = 20$	27/80	31/80	17/40	19/40	41/80
Large-sample approximation		$1.07\sqrt{\frac{m+n}{mn}}$	$1.22\sqrt{\frac{m+n}{mn}}$	$1.36\sqrt{\frac{m+n}{mn}}$	$1.52\sqrt{\frac{m+n}{mn}}$	$1.63\sqrt{\frac{m+n}{mn}}$

Source: Adapted by permission from F. J. Massey, Distribution table for the deviation between two sample cumulatives, *Ann. Math. Stat.* 23 (1952), 435–441.

[a]This table gives the values of $D^+_{m,n,\alpha}$ and $D_{m,n,\alpha}$ for which $\alpha \geq P\{D^+_{m,n} > D^+_{m,n,\alpha}\}$ and $\alpha \geq P\{D_{m,n} > D_{m,n,\alpha}\}$ for some selected values of N_1 = smaller sample size, N_2 = larger sample size, and α.