Brief Summary of

Analysis of Variance (ANOVA)

&

Nonparametric Statistics

for MFIN Students

Mahmut PARLAR

日 月

2014-05-05

Software: MegaStat add-in for Excel (free)

Go to my web site (google: Mahmut Parlar) for links

Contents

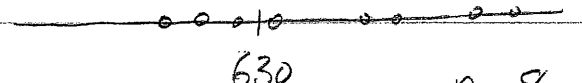
① ANOVA

Let's start with a quick review of hypothesis testing

1. Hypothesis Testing (One pop'n)

$H_0: \mu = \mu_0$ (=630)
$H_{ai}: \mu < \mu_0$

Avg GMAT score at Deboots 2014



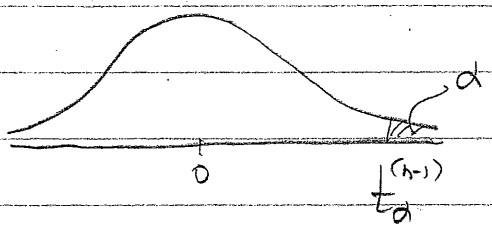
$n = 8$
$\sigma = ?$

Can you reject H_0?

Test stat (σ unknown)

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad df = n - 1$$

Assume
- pop'n normal



Reject H_0

Don't reject H_0

DeGroote

618 Hypothesis Test: Mean vs. Hypothesized Value

621

625 630.000 hypothesized value

632 632.875 mean DeGroote

636 10.723 std. dev.

640 3.791 std. error

645 8 n

646 7 df

Megastat

0.76 t

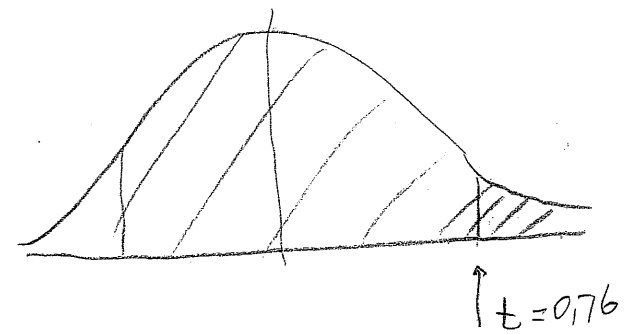
.7635 p-value (one-tailed, lower)

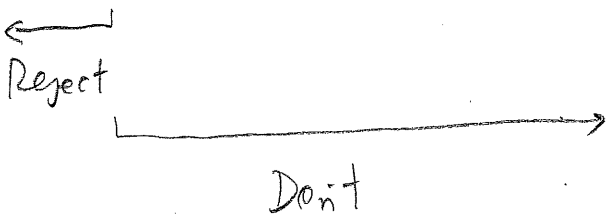
623.910 confidence interval 95.% lower

641.840 confidence interval 95.% upper

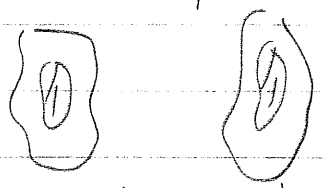
8.965 margin of error

p too large! Don't reject!





2. HT (Two popn's)



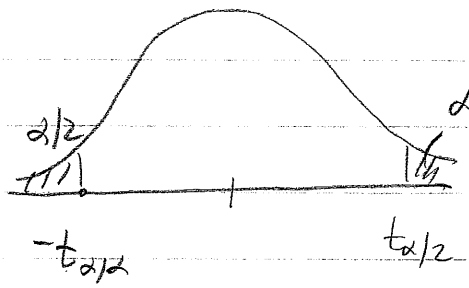
DeGroote Rotman

μ_1	μ_2	
σ_1^2	σ_2^2	: unknown but equal(!)
n_1	n_2	

H_0: $\mu_1 = \mu_2$ $\quad (\mu_1 - \mu_2 = 0)$

H_a: $\mu_1 \neq \mu_2$ $\quad (\mu_1 - \mu_2 \neq 0)$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$



Assume
- popn normal
- variances equal
- samples ind't

← Reject H_0

Don't reject

DeGroote	Rotman
618	615
621	642
625	630
632	632
636	610
640	622
645	636
646	644
	610

Hypothesis Test: Independent Groups (t-test, pooled variance)

DeGroote	Rotman	
632.88	626.78	mean
10.72	13.11	std. dev.
8	9	n

15	df
6.097	difference (DeGroote - Rotman)
145.362	pooled variance
12.057	pooled std. dev.
5.858	standard error of difference
0	hypothesized difference
1.04	t
.3145	p-value (two-tailed)
-6.390	confidence interval 95.% lower
18.584	confidence interval 95.% upper
12.487	margin of error

F-test for equality of variance

171.94	variance: Rotman
114.98	variance: DeGroote
1.50	F
.6092	p-value

(See Appendix A for preliminaries)

3. ANOVA ⟶ $\begin{cases} H_0: \mu_1 = \mu_2 = \dots = \mu_m \\ H_a: \text{at least two differ} \end{cases}$, m: # pop's

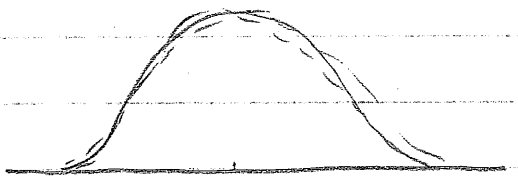
a) Motivation

D R S

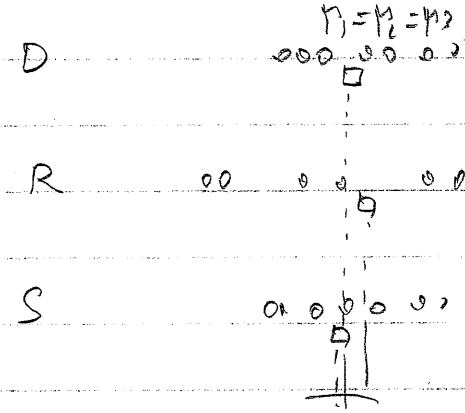
Three MBA programs DSB, Rotman, Schulich

Claim all three have equal GMAT averages

If all equal



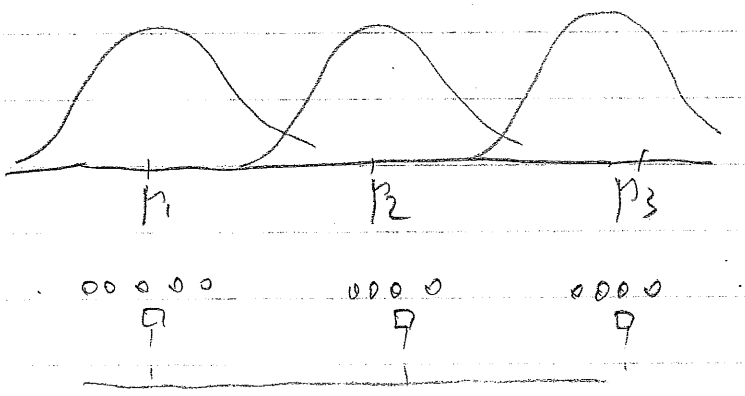
Take samples



If all equal

Between groups variation small than within group

If they're different



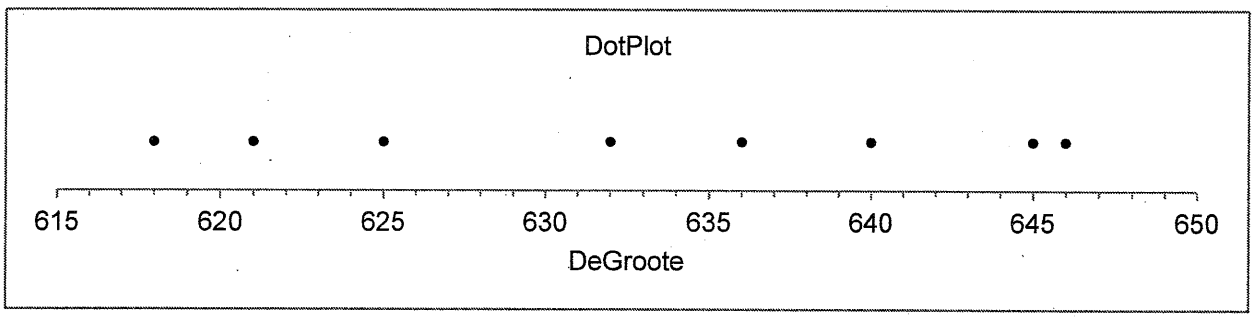
Between group > within groups

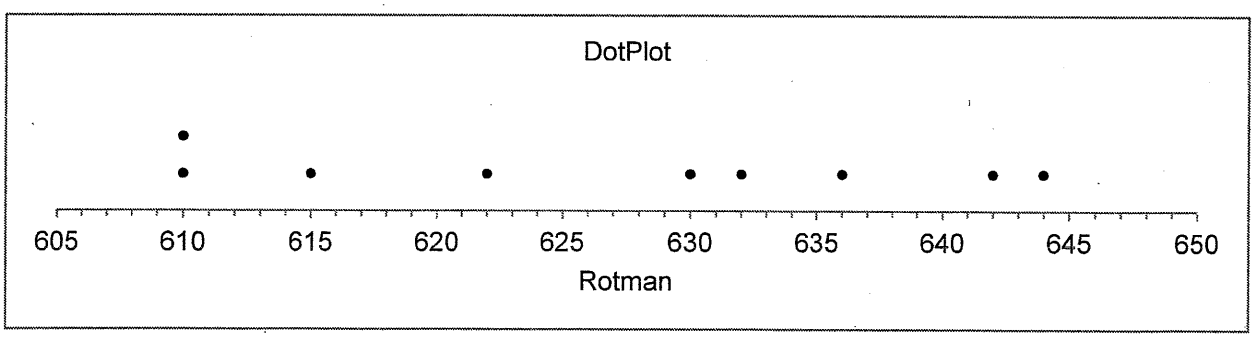
How do we test this?

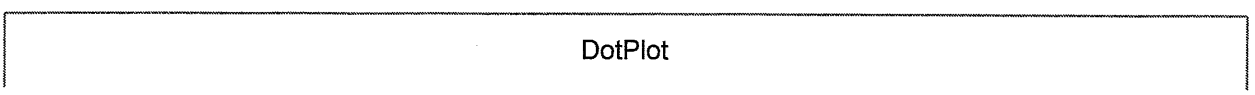
DeGroote	Rotman	Schulich
618	615	600
621	642	599
625	630	630
632	632	602
636	610	630
640	622	605
645	636	
646	644	
	610	

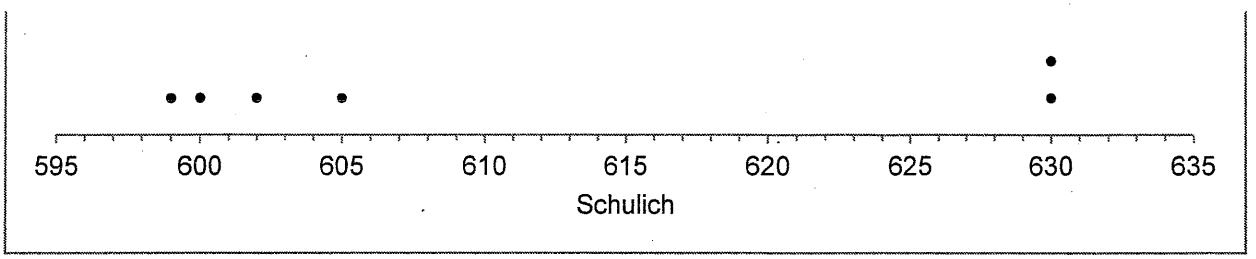
Descriptive statistics

	DeGroote	*Rotman*	*Schulich*
count	8	9	6
mean	632.88	626.78	611.00
sample standa	10.72	13.11	14.86
sample varianc	114.98	171.94	220.80
minimum	618	610	599
maximum	646	644	630
range	28	34	31









One factor ANOVA

Mean	*n*	*Std. Dev*	
632.9	8	10.72	DeGroote
626.8	9	13.11	Rotman
611.0	6	14.86	Schulich
624.8	23	15.05	Total

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Treatment	1,699.48	2	849.741	5.17	.0154
Error	3,284.43	20	164.222		
Total	4,983.91	22			

Post hoc analysis

p-values for pairwise t-tests

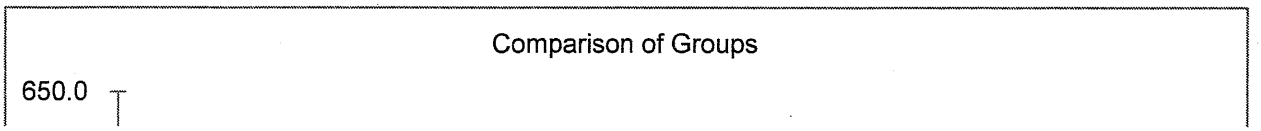
		Schulich 611.0	Rotman 626.8	DeGroote 632.9
Schulich	611.0			
Rotman	626.8	.0300		
DeGroote	632.9	.0049	.3392	

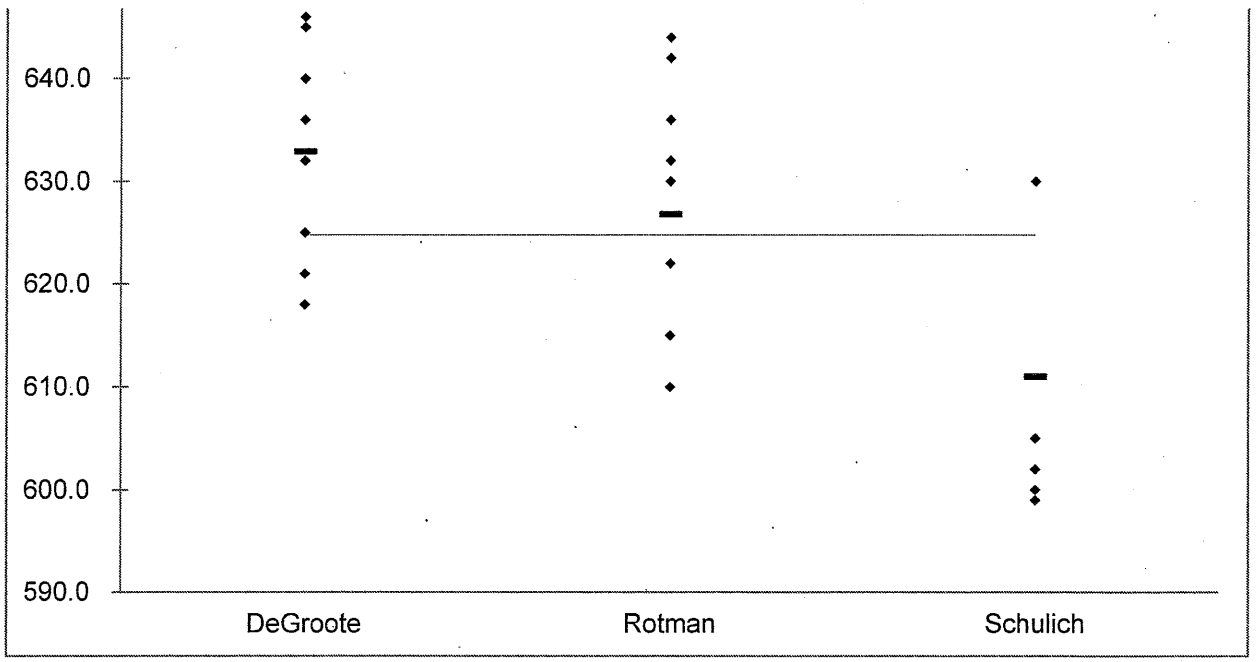
Tukey simultaneous comparison t-values (d.f. = 20)

		Schulich 611.0	Rotman 626.8	DeGroote 632.9
Schulich	611.0			
Rotman	626.8	2.34		
DeGroote	632.9	3.16	0.98	

critical values for experimentwise error rate:

0.05	2.53
0.01	3.28





(See Appendix B for the logic behind ANOVA)

'14/01/22

(2) Nonparametric Methods

Assumptions in previous chapter
- normality
- equality of variances
- independence of observations

But, what if  etc
and variances are not equal?

Violation of above assumptions distort results
So, we need nonparametrics.

One Pop'n	Two Pop'ns	Three or more
- Sign test (median)	- Wilcoxon - Mann/Whitney	- Kruskall-Wallis

1. Sign Test

If the popn is not symmetric, it makes sense to test the median (not mean)! Why?

Consider again the GMAT@NJB case

$\left.\begin{array}{l} H_0: M_d = 630 \\ H_a: M_d < 630 \end{array}\right\}$ Same data

But, can't assume normality/known σ^2, etc...

Let π: proportion of students who have < 630

$H_0: \pi = 0.5$
$H_a: \pi < 0.5$

Let X: total # who get < 630

So, $X \sim Bin(8, 0.5)$ if H_0 true

If I can reject $H_0: \pi \geq 0.5$, I can also reject $H_0: \pi = 0.5$.

In our ex. 618, 621, 625 | 632, 636, 640, 645, 640, $n = 8$

3 | 5

We expected $\mu_X = n\pi = 8(.5) = 4$ to fall below 630

$$p\text{-value} = Pr(X \leq 3) = \sum_{x=0}^{3} \frac{8!}{x!(8-x)!} (0.5)^x (0.5)^{8-x} = 0.363$$

This is an extremely high value for p, so can't reject H_0.

DeGroote

Sign Test

618
621
625
632
636
640
645
646

630 hypothesized value
634 median DeGroote
3 below
0 equal
5 above

8 n

binomial

.3633 p-value (one-tailed, lower)

normal approximation

0.35 z

.6382 p-value (one-tailed, lower)

Don't reject

Binomial distribution

8 n

0.5 p

X	P(X)	cumulative probability
0	0.00391	0.00391
1	0.03125	0.03516
2	0.10938	0.14453
3	0.21875	0.36328
4	0.27344	0.63672
5	0.21875	0.85547
6	0.10938	0.96484
7	0.03125	0.99609
8	0.00391	1.00000
	1.00000	

4.000 expected value
2.000 variance
1.414 standard deviation

2. Mann-Whitney Rank Sum Test (Two pop's)

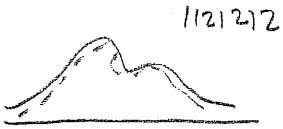
Pop. 1	Pop. 2
$f_1(x)$	$f_2(x)$
n_1	n_2

Rank n_1+n_2 in $\uparrow$ order (Ties → equal rank to avg)

T_1: Sum of ranks in n_1 $\quad$ T_2: Sum of ranks in n_2

Test stat $T = \begin{cases} T_1, & n_1 \le n_2 \\ T_2, & n_1 > n_2 \end{cases}$

H_0: $f_1(x)$ & $f_2(x)$ are identical



H_a: " " " NOT "

or

Reject H_0 if $T \le T_L$ or $T \ge T_U$ (in tables)

Our ex.

n_2 \ n_1	8	
	T_L	T_U
9	54	90

But $T = 83$ → can't reject

DeGroote	Rotman
618	615
621	642
625	630
632	632
636	610
640	622
645	636
646	644
	601

Wilcoxon - Mann/Whitney Test

n	sum of ranks	
8	83	DeGroote
9	70	Rotman
17	153	total

$E(T_1) = \frac{n_1(n_1+n_2+1)}{2}$ = 72.000 expected value

10.380 standard deviation

1.060 z corrected for ties

.2892 p-value (two-tailed)

No.	Label	Data	Rank
1	DeGroote	618	4
2	DeGroote	621	5
3	DeGroote	625	7
4	DeGroote	632	9.5
5	DeGroote	636	11.5
6	DeGroote	640	13
7	DeGroote	645	16
8	DeGroote	646	17
9	Rotman	615	3
10	Rotman	642	14
11	Rotman	630	8
12	Rotman	632	9.5
13	Rotman	610	2
14	Rotman	622	6
15	Rotman	636	11.5
16	Rotman	644	15
17	Rotman	601	1

$Var(T_2) = \frac{n_1 n_2 (n_1+n_2+1)}{12}$

3. Kruskall-Wallis Test (Nonparametric ANOVA)

m samples but no assumptions on normality or equal variances

DeGroote Kotman Schulich

$H_0: \mu_1 = \mu_2 = \mu_3$ $(m=3)$

H_a: at least two $\neq$

Procedure:

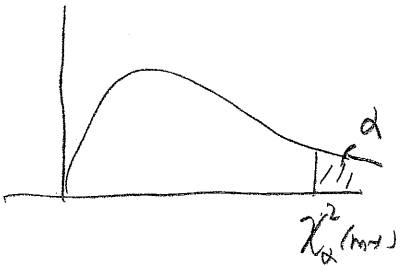
① Rank data (Ties assigned mid-value)

② R_i: sum of ranks of ith group $i = 1, \ldots, m$

③ Test stat

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{m} \frac{R_i^2}{n_i} - 3(n+1)$$

It is approx $\chi^2_{(m-1)}$



DeGroote	Rotman	Schulich
618	615	600
621	642	599
625	630	630
632	632	602
636	610	630
640	622	605
645	636	
646	644	
	610	

Kruskal-Wallis Test

Median	*n*	*Avg. Rank*	
634.00	8	15.63	DeGroote
630.00	9	12.78	Rotman
603.50	6	6.00	Schulich
630.00	23		Total

7.124	H (corrected for ties)
2	d.f.
.0284	p-value

p small. Reject H_0

multiple comparison values for avg. ranks

8.29 (.05) 10.17 (.01)

No.	Label	Data	Rank
1	DeGroote	618	8
2	DeGroote	621	9
3	DeGroote	625	11
4	DeGroote	632	15.5
5	DeGroote	636	17.5
6	DeGroote	640	19
7	DeGroote	645	22
8	DeGroote	646	23
9	Rotman	615	7
10	Rotman	642	20
11	Rotman	630	13
12	Rotman	632	15.5
13	Rotman	610	5.5
14	Rotman	622	10
15	Rotman	636	17.5
16	Rotman	644	21
17	Rotman	610	5.5
18	Schulich	600	2
19	Schulich	599	1
20	Schulich	630	13
21	Schulich	602	3
22	Schulich	630	13
23	Schulich	605	4

See Appendix C for the logic behind K-W test

Appendix A

Appendix B

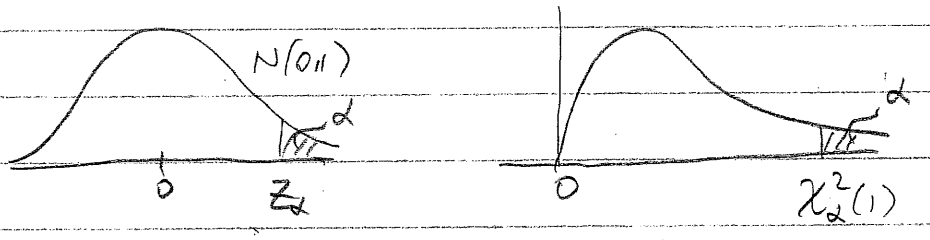
Appendix C

Appendix A: Preliminaries (Review)

We need to remember some facts

χ^2

If $Z \sim N(0,1)$, then $Z^2 = \chi^2(1)$ Chi-square df=1



If $Z_1, \ldots, Z_n \sim N(0,1)$, then $Y = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$, df=n

$E(Y) = n$

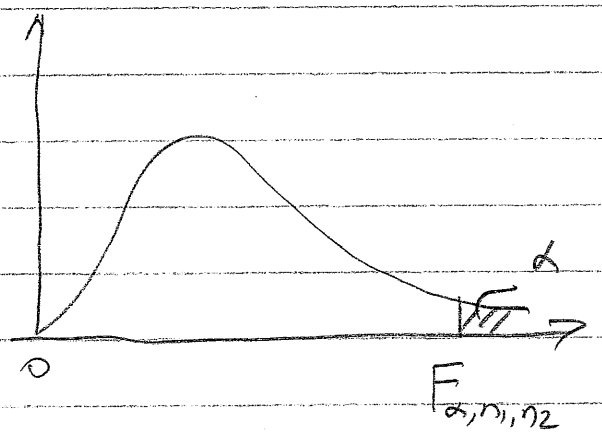
$$f(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{(n-2)/2} e^{-y/2}, \quad y > 0$$

$$M_Y(t) = \frac{1}{(1-2t)^{n/2}} \quad \text{mgf}$$

F

If U (df=n_1) & V (df=n_2) indt chi-square, then

$$F = \frac{U/n_1}{V/n_2} \sim \text{F-distribution } (df_1, df_2) = (n_1, n_2)$$



Sum of χ^2: If $\overbrace{X_1, \ldots, X_k}^{\text{indit}}$ are $\chi^2(r_1), \ldots, \chi^2(r_k)$,

then $Y = X_1 + \cdots + X_k \sim \chi^2(r_1 + \cdots + r_k)$

Theorem: If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

then $\dfrac{(n-1)S^2}{\sigma^2} = \dfrac{\sum(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2_{(n-1)}$ ∎

- Consider $SS_t = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2$: each $X_{ij} \sim N(\mu, \sigma^2)$

So, $\boxed{\dfrac{SS_t}{\sigma^2} \sim \chi^2_{(n-1)}}$, $n = \sum_{i=1}^{m} n_i$

- Consider $SS_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2$

$$= \sum_{j=1}^{n_1}(X_{1j} - \bar{X}_{1\cdot})^2 + \sum^{n_2}(X_{2j} - \bar{X}_{2\cdot})^2 + \ldots + \sum^{n_m}(X_{mj} - \bar{X}_{m\cdot})^2$$

Divide by σ^2 all terms $\Rightarrow$

$$\frac{SS_w}{\sigma^2} = \chi^2_{(n_1-1)} + \ldots + \chi^2_{(n_m-1)}$$

$n_1 - 1$, $n_2 - 1$, …, $n_m - 1$ } m

df: $n - m$

$$\boxed{\frac{SS_w}{\sigma^2} = \chi^2_{(n-m)}}$$

So, since $SS_t = SS_w + SS_b$

$\chi^2_{(n-1)} : \chi^2_{(n-m)}$ ↑ $n - 1 - n + m = m - 1$

i.e., $\boxed{SS_b \sim \chi^2_{(m-1)}}$

Useful property.

Recall that if $Y \sim \chi^2(n)$, then $E(Y) = n$

Since $\frac{SS_b}{\sigma^2} \sim \chi^2(m-1) \Rightarrow E\left(\frac{SS_b}{\sigma^2}\right) = m-1$

under H_0: $\mu_1 = \cdots = \mu_m$

$\Downarrow$

$E\left(\frac{SS_b}{m-1}\right) = \sigma^2$

$\frac{SS_w}{\sigma^2} \sim \chi^2(n-m) \Rightarrow E\left(\frac{SS_w}{n-m}\right) = \sigma^2$ also

But if H_0 false, i.e., $\mu_1, \ldots, \mu_m$ not equal, then

$$E\left(\frac{SS_b}{m-1}\right) = \sigma^2 + \sum_{i=1}^{m} n_i \frac{(\mu_i - \bar{\mu})^2}{m-1} > \sigma^2$$

$\therefore$ Estimator for σ^2 using SS_b $>$ using SS_w when H_0 false (usually (on average))

So, base test of H_0 on $\left(\frac{SS_b}{m-1}\right) \Big/ \left(\frac{SS_w}{n-m}\right)$

both unbiased under H_0

~ 1 if H_0 true

Under H_0: $F = \dfrac{\overbrace{SS_b/(m-1)}^{\chi^2}}{\underbrace{SS_w/(n-m)}_{\chi^2}}$

Appendix B

Logic behind ANOVA

$H_0: \mu_1 = \mu_2 = \dots = \mu_m$

H_a: at least two $\neq$

Assume: - Popn's normal
- variances equal
- samples indep't

m groups $i = 1, \dots, m$

n observations (total)

n_i in group i

Variation (SS)

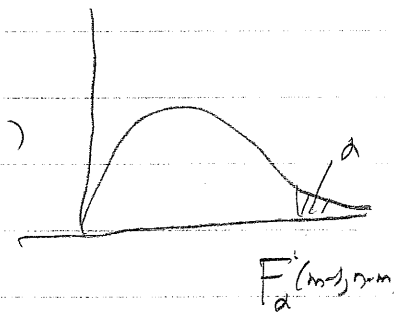
Within group | Between group

$$SS_W = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2 \qquad SS_b = \sum_{i=1}^{m} n_i (\bar{x}_{i\cdot} - \bar{\bar{x}})^2$$

$$MS_W = \frac{SS_W}{n-m} \sim \chi^2_{(n-m)} \qquad MS_b = \frac{SS_b}{m-1} \sim \chi^2_{(m-1)}$$

So, if H_0 is true, then

$$F = \frac{SS_b/(m-1)}{SS_W/(n-m)} \quad \text{is} \quad F_{(m-1, n-m)}$$



If F is large, reject H_0

Note: Total SS_t is

$$SS_t = SS_W + SS_b.$$

Appendix C: Another example & logic behind K-W Test

'14/01/24

EX. SAT Scores

	Group 1	2	3	$m=3$
	772	792	752	
$H_0: \mu_1=\mu_2=\mu_3$	764	612	680	
	600	592	624	
	564		580	
			572	

Score	792	772	764	752	680	624	612	600	592	580	572	564
Rank	1	2	3	4	5	6	7	8	9	10	11	12
Group	2	1	1	3	3	3	2	1	2	3	3	1

R_i: rank sum of ith group

$R_1 = 2+3+8+12 = 25$ $\quad n_1=4$ $\quad \frac{1}{2}(N+1)=\frac{13}{2}$

$R_2 = 1+7+9 = 17$ $\quad n_2=3$

$R_3 = 4+5+6+10+11 = 36$ $\quad n_3=5$

$\bar{\bar{x}}$: grand mean

$\bar{x}_i = R_i/n_i$ $\quad \frac{25}{4}$ $\quad \frac{17}{3}$ $\quad \frac{36}{5}$ $\quad E\left(\frac{R_i}{n_i}\right)=\frac{1}{2}(N+1)$

$\bar{\bar{x}}$ $\quad \frac{13}{2}$ $\quad \frac{13}{2}$ $\quad \frac{13}{2}$

$-\frac{1}{4}$ $\quad -\frac{5}{6}$ $\quad \frac{7}{10}$

$$(SS_b): D = 4\left(-\frac{1}{4}\right)^2 + 3\left(-\frac{5}{6}\right)^2 + 5\left(\frac{7}{10}\right)^2 = \frac{287}{60}$$

$E(X-\mu)^2 = Var(X)$ $\quad E\left(\frac{R_i}{n_i}\right) = \frac{N+1}{2}$

In general $D = \sum_{i=1}^{m} n_i \left(\frac{R_i}{n_i} - \frac{N+1}{2}\right)^2$

$$= \sum_{i=1}^{m} \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}$$

Now $E\left(\frac{R_i}{n_i} - \frac{N+1}{2}\right)^2 = Var\left(\frac{R_i}{n_i}\right)$ since $E(X-\mu)^2 = Var(X)$

But from sampling w/o replacement theory,

$$Var(\bar{x}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \quad \text{(basic stat)}$$

So, here $Var(\text{mean of a sample of } n_i) = \frac{\sigma^2}{n_i}\left(\frac{N-n_i}{N-1}\right)$

But we have a uniform distribution of N measurements

So,

$$\sigma^2 = \sum_{n=1}^{N} \frac{1}{N}\left[n - \frac{1}{2}(N+1)\right]^2 = \frac{N^2-1}{12}$$

$$\therefore E(D) = \sum_{i=1}^{m} n_i \frac{\sigma^2}{n_i}\left(\frac{N-n_i}{N-1}\right) = \frac{(m-1)N(N+1)}{12}$$

We want a statistic whose mean is indep of sample size. So, define

$$H = \frac{D}{N(N+1)/12} = \frac{12}{N(N+1)} \sum_{i=1}^{m} \frac{R_i^2}{n_i} - 3(N+1)$$

Note: Nothing similar to SS_W since we can't be sure about equality of variances.

Kruskal-Wallis Test

Median	*n*	*Avg. Rank*	
682.00	4	6.75	Group 1
612.00	3	7.33	Group 2
624.00	5	5.80	Group 3
618.00	12		Total

0.368	H
2	d.f.
.8320	p-value

multiple comparison values for avg. ranks
6.10 (.05) 7.48 (.01)

No.	Label	Data	Rank
1	Group 1	772	11
2	Group 1	764	10
3	Group 1	600	5
4	Group 1	564	1
5	Group 2	792	12
6	Group 2	612	6
7	Group 2	592	4
8	Group 3	752	9
9	Group 3	680	8
10	Group 3	624	7
11	Group 3	580	3
12	Group 3	572	2