Class 25

Daniel B. Rowe, Ph.D.

Department of Mathematical and Statistical Sciences

Agenda:

Recap Chapter 12.1

Statistical Inference

Recap Chapter 12.1

12: Analysis of Variance

12.1 Introduction to the Analysis of Variance

Previously we learned about hypothesis testing for:
One Population: μ, p, and σ^2.
Two Populations: $\mu_d=\mu_1-\mu_2$, $\mu_1-\mu_2$, p_1-p_2, and σ_1^2/σ_2^2.

(We also learned about hypothesis testing for contingency tables.)

Now we are going to study hypothesis testing for three or more populations.

Three Populations: at least two of $\mu_1,\mu_2,\mu_3,\ldots$ different.

12: Analysis of Variance

12.1 Introduction to the Analysis of Variance

If we are testing for differences in means,
…why are we analyzing variance?

As it turns out, we calculate two variances and take the ratio.

If all the means are truly the same, the two variances will be the same and the ratio will be 1.

12: Analysis of Variance

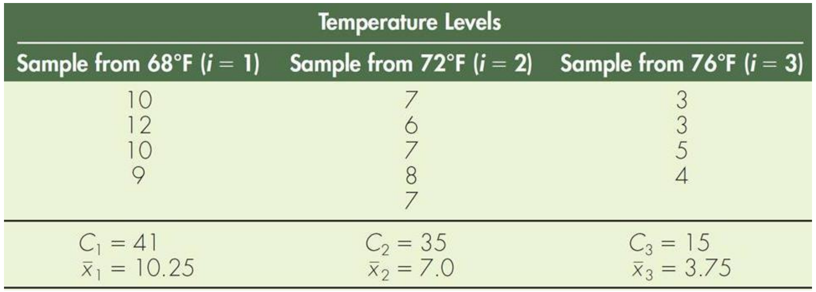
12.1 Introduction to the Analysis of Variance

Hypothesis Testing Procedure

Step 1 $H_0 : \mu_1 = \mu_2 = \mu_3$ vs.

H_a: *at least two* μ*'s different*

Step 2



Temperature Levels		
Sample from 68°F ($i = 1$)	Sample from 72°F ($i = 2$)	Sample from 76°F ($i = 3$)
10	7	3
12	6	3
10	7	5
9	8	4
	7	
$C_1 = 41$ $\bar{x}_1 = 10.25$	$C_2 = 35$ $\bar{x}_2 = 7.0$	$C_3 = 15$ $\bar{x}_3 = 3.75$

Sum of Squares Due to Factor

$$SS(\text{factor}) = \left(\frac{C_1^2}{k_1} + \frac{C_2^2}{k_2} + \frac{C_3^2}{k_3} + \cdots\right) - \frac{(\sum x)^2}{n}$$

$$\text{df(factor)} = c - 1$$

$$\text{MS(factor)} = \frac{SS(\text{factor})}{\text{df(factor)}}$$

Sum of Squares Due to Error

$$SS(\text{error}) = \sum(x^2) - \left(\frac{C_1^2}{k_1} + \frac{C_2^2}{k_2} + \frac{C_3^2}{k_3} + \cdots\right)$$

$$\text{df(error)} = n - c$$

$$\text{MS(error)} = \frac{SS(\text{error})}{\text{df(error)}}$$

$$F\star = \frac{\text{MS(factor)}}{\text{MS(error)}}$$

$$\alpha = .05$$

Shortcut for Total Sum of Squares

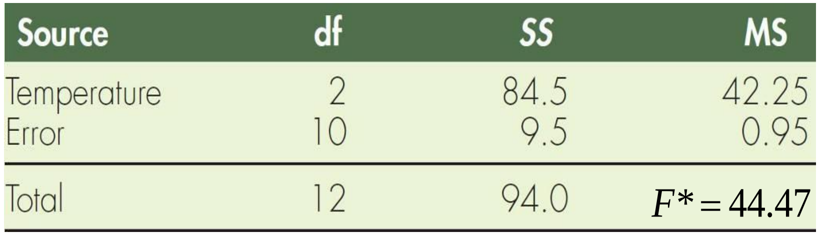
$$SS(\text{total}) = \sum(x^2) - \frac{(\sum x)^2}{n}$$

$$\text{df(total)} = n - 1$$

12: Analysis of Variance

12.1 Introduction to the Analysis of Variance

Hypothesis Testing Procedure



Source	df	SS	MS
Temperature	2	84.5	42.25
Error	10	9.5	0.95
Total	12	94.0	$F^*=44.47$

$$SS(temperature)=\left(\frac{41^2}{4}+\frac{35^2}{5}+\frac{15^2}{4}\right)-\frac{(91)^2}{13}=84.5$$

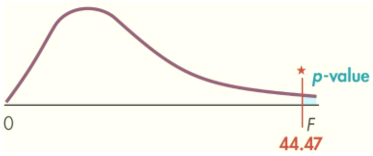
$$SS(error)=731.0-\left(\frac{41^2}{4}+\frac{35^2}{5}+\frac{15^2}{4}\right)=9.5$$

Step 3

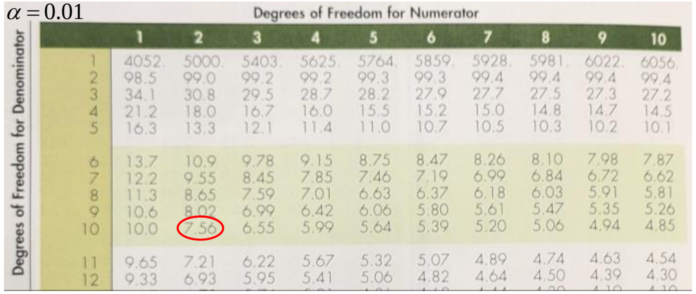
$$F^*=\frac{42.25}{0.95}=44.47$$

Step 4

p-value approach



$\alpha=0.01$

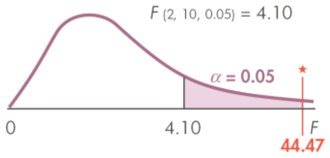


Degrees of Freedom for Numerator

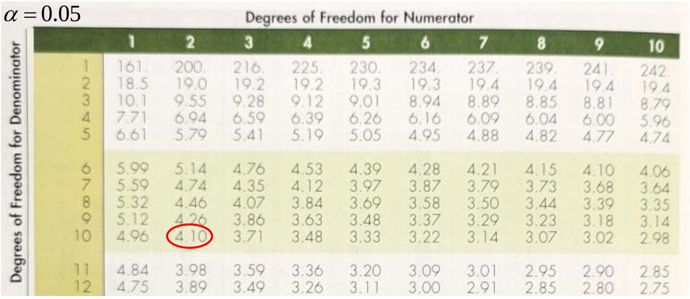
Degrees of Freedom for Denominator	1	2	3	4	5	6	7	8	9	10
1	4052.	5000.	5403.	5625.	5764.	5859.	5928.	5981.	6022.	6056.
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30

$$.00<p-value<.01$$

Classical approach



$\alpha=0.05$



Degrees of Freedom for Numerator

Degrees of Freedom for Denominator	1	2	3	4	5	6	7	8	9	10
1	161.	200.	216.	225.	230.	234.	237.	239.	241.	242.
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75

$$F(2,10,0.05)=4.10$$

Step 5

Decision: Reject H_0

$$p-value<\alpha$$

$$.00<p-value<.01$$

$$\alpha=0.05$$

$$F^*>F_{crit}$$

$$F^*=44.47$$

$$F_{crit}=4.10$$

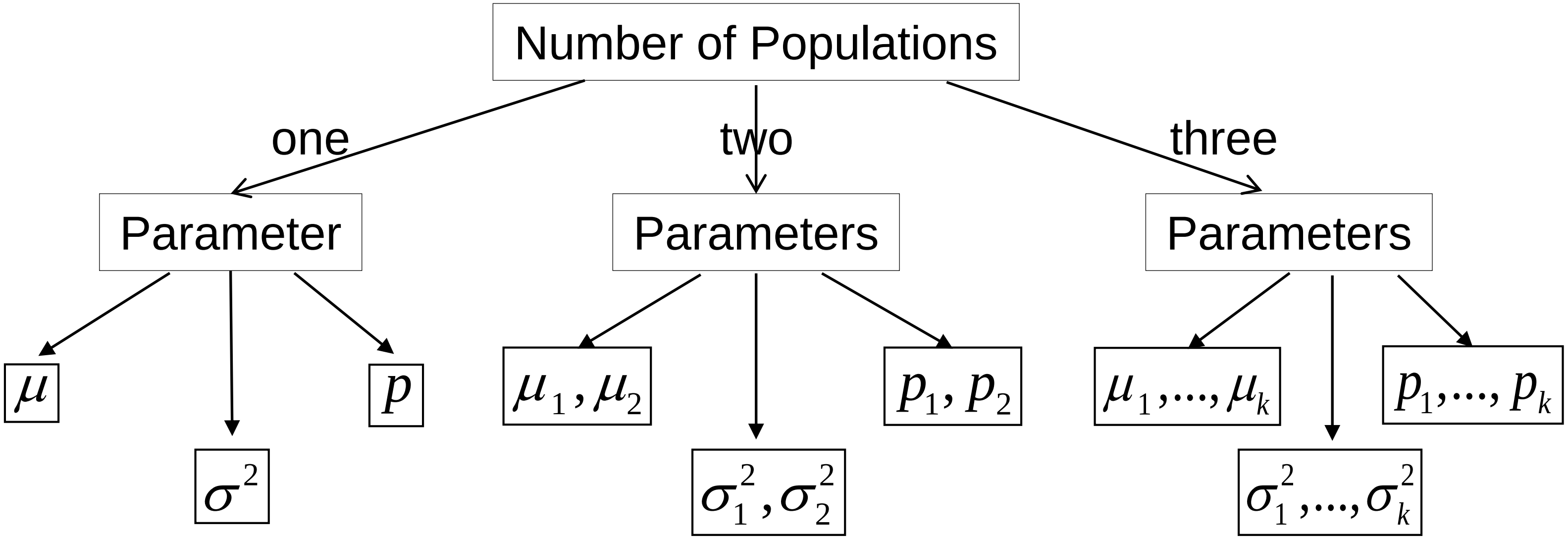
Statistical Inference

Daniel B. Rowe, Ph.D.

Department of Mathematical and Statistical

Statistical Inference

Statistical Inference:



Number of Populations

one

two

three

Parameter

Parameters

Parameters

μ

σ^2

p

μ_1, μ_2

σ_1^2, σ_2^2

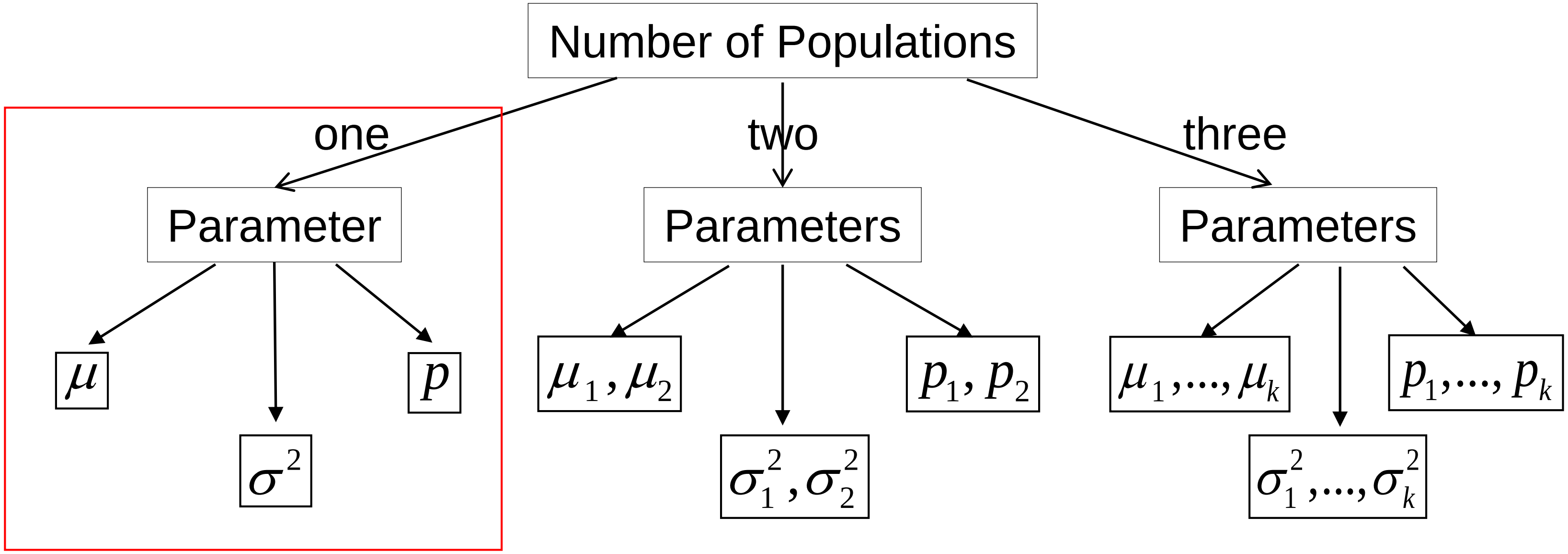
p_1, p_2

$\mu_1, \ldots, \mu_k$

$\sigma_1^2, \ldots, \sigma_k^2$

$p_1, \ldots, p_k$

Statistical Inference: 1 Population



Number of Populations

one

Parameter

μ

σ^2

p

two

Parameters

μ_1, μ_2

σ_1^2, σ_2^2

p_1, p_2

three

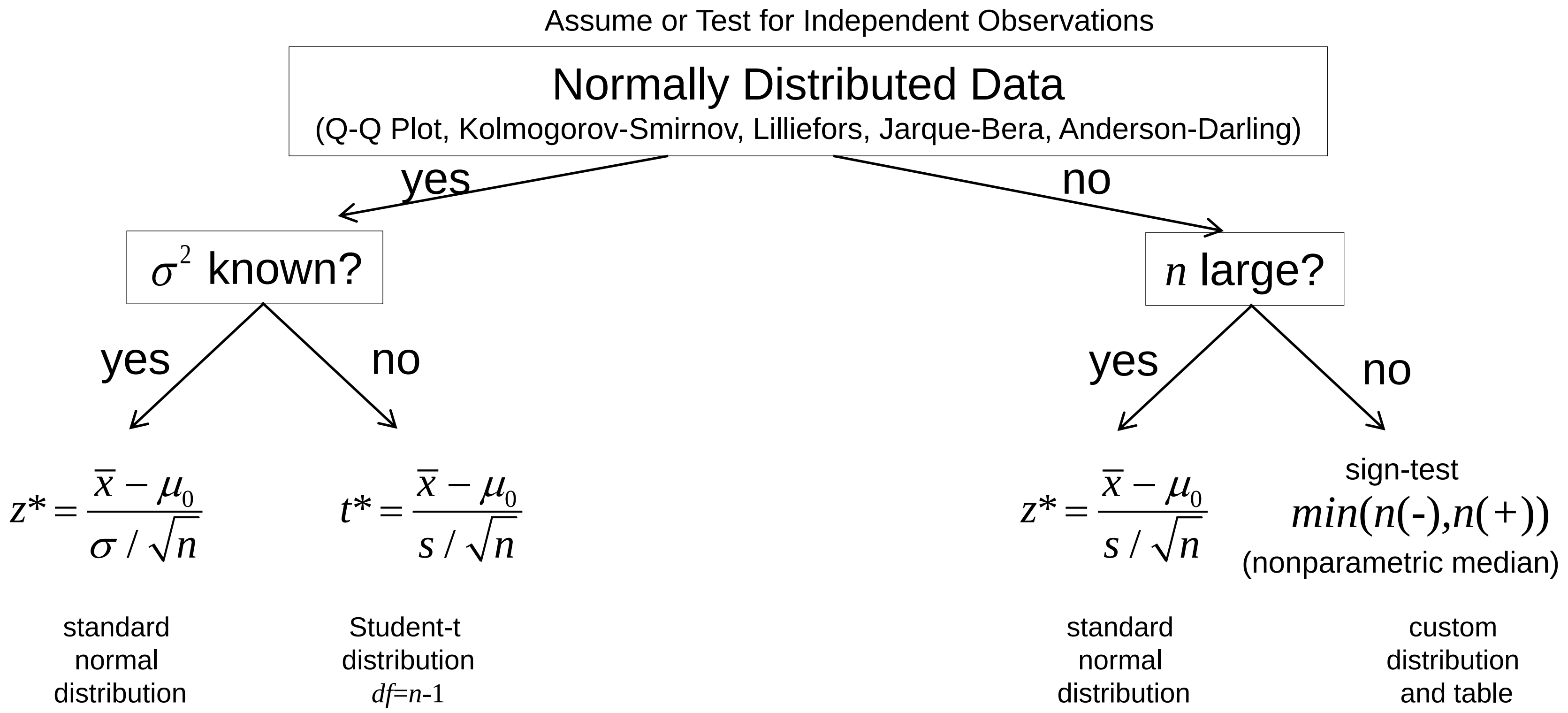
Parameters

$\mu_1, \dots, \mu_k$

$\sigma_1^2, \dots, \sigma_k^2$

$p_1, \dots, p_k$

Statistical Inference: Procedures for μ



Assume or Test for Independent Observations

Normally Distributed Data

(Q-Q Plot, Kolmogorov-Smirnov, Lilliefors, Jarque-Bera, Anderson-Darling)

yes

no

σ^2 known?

n large?

yes

no

yes

no

$$z^* = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

$$t^* = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

$$z^* = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

sign-test

$$min(n(\text{-}), n(+))$$

(nonparametric median)

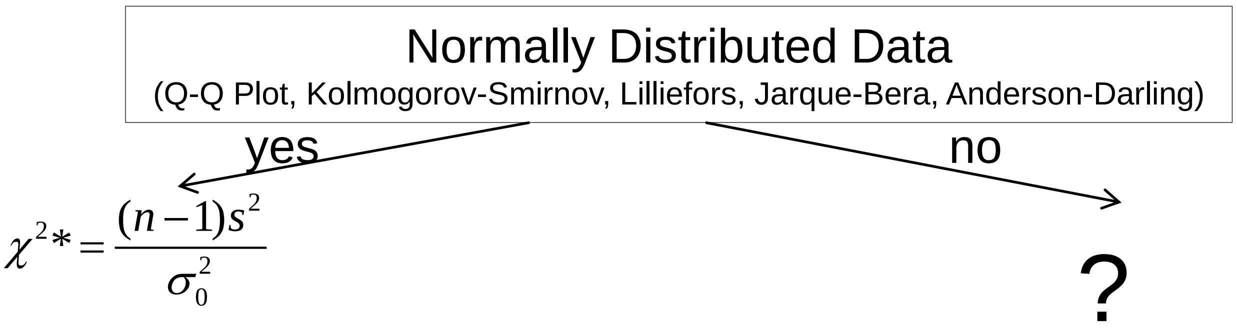
standard
normal
distribution

Student-t
distribution
$df = n-1$

standard
normal
distribution

custom
distribution
and table

Statistical Inference: Procedures for σ^2



Assume or Test for Independent Observations

Normally Distributed Data

(Q-Q Plot, Kolmogorov-Smirnov, Lilliefors, Jarque-Bera, Anderson-Darling)

yes

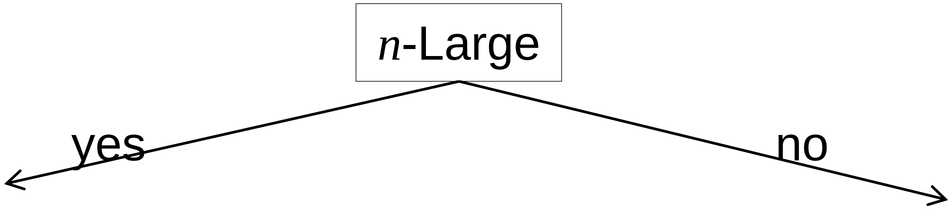
$$\chi^{2}* = \frac{(n-1)s^{2}}{\sigma_0^{2}}$$

Chi-Square distribution $df=n$-1

no

?

Statistical Inference: Procedures for *p*



n-Large

yes

no

$$z^* = \frac{p' - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Normal distribution

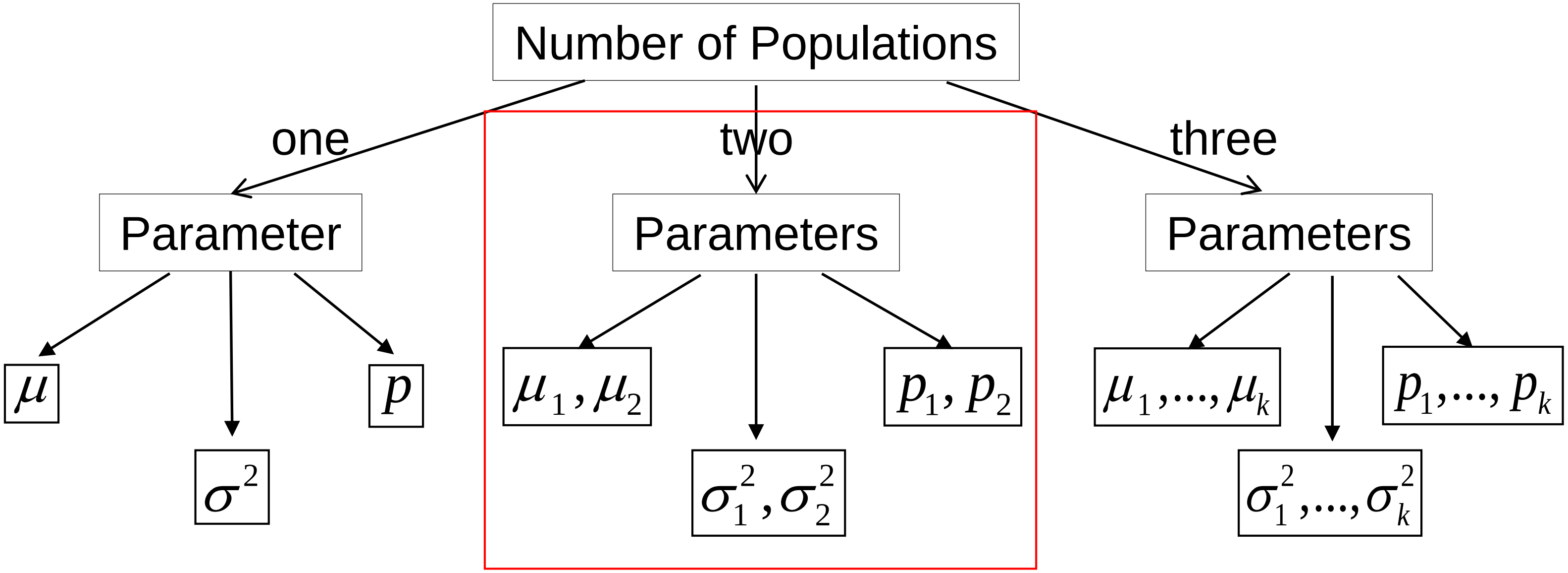
$$p' = \frac{x}{n}$$

Binomial Distribution

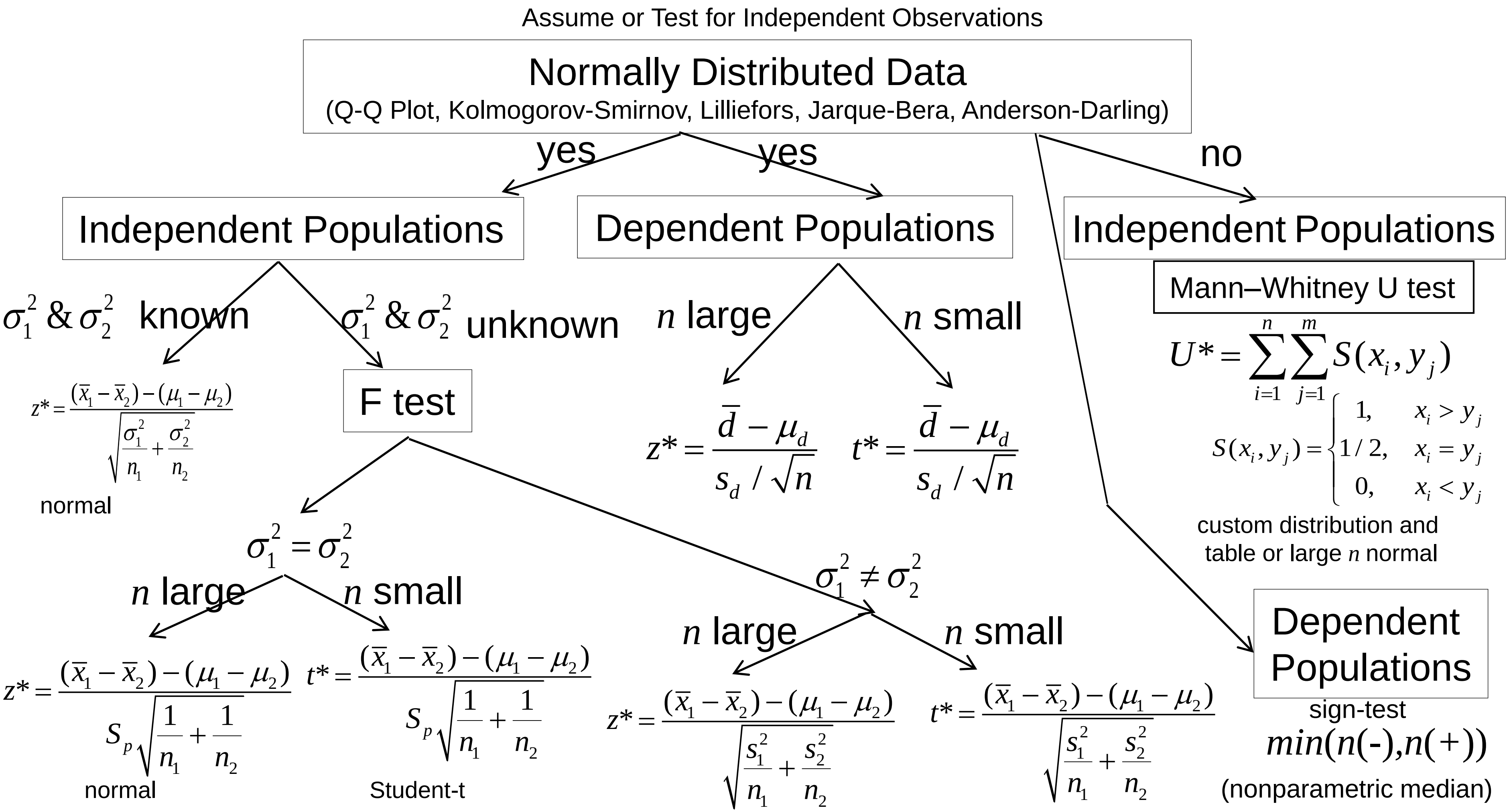
$$P(x) = \frac{n!}{x!(n-x)!} p_0^x (1-p_0)^{n-x}$$

p-value= $\sum_{x>i} \frac{n!}{i!(n-i)!} p_0^i (1-p_0)^{n-i}$

Statistical Inference: 2 Populations



Statistical Inference: Procedures for μ_1-μ_2



Assume or Test for Independent Observations

Normally Distributed Data

(Q-Q Plot, Kolmogorov-Smirnov, Lilliefors, Jarque-Bera, Anderson-Darling)

yes

yes

no

Independent Populations

σ_1^2 & σ_2^2 known

$$z^*=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

normal

σ_1^2 & σ_2^2 unknown

F test

$\sigma_1^2=\sigma_2^2$

n large

$$z^*=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

normal

n small

$$t^*=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Student-t

$\sigma_1^2\neq\sigma_2^2$

n large

$$z^*=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

n small

$$t^*=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

Dependent Populations

n large

$$z^*=\frac{\bar{d}-\mu_d}{s_d/\sqrt{n}}$$

n small

$$t^*=\frac{\bar{d}-\mu_d}{s_d/\sqrt{n}}$$

Independent Populations

Mann–Whitney U test

$$U^*=\sum_{i=1}^{n}\sum_{j=1}^{m}S(x_i,y_j)$$

$$S(x_i,y_j)=\begin{cases}1, & x_i>y_j\\ 1/2, & x_i=y_j\\ 0, & x_i<y_j\end{cases}$$

custom distribution and table or large n normal

Dependent Populations

sign-test

$$min(n(-),n(+))$$

(nonparametric median)

Statistical Inference: Procedures for σ_1^2, σ_2^2

Assume or Test for Independent Observations

Normally Distributed Data

(Q-Q Plot, Kolmogorov-Smirnov, Lilliefors, Jarque-Bera, Anderson-Darling)

yes

$$F^* = \frac{s_1^2}{s_2^2}$$

F-distribution

$df_1 = n_1 - 1$

$df_2 = n_2 - 1$

no

Levene test

$$W^* = \frac{(n-k)\sum_{i=1}^{k} n_i (\bar{A}_{i.} - \bar{A}_{..})^2}{(k-1)\sum_{i=1}^{k} \sum_{j=}^{n_i} (A_{ij} - \bar{A}_{i.})^2}$$

F-distribution

$df_1 = n_1 - 1$

$df_2 = n_2 - 1$

$$A_{ij} = | y_{ij} - \bar{y}_{i.} |$$

Pop *i* Obs *j*

“.”=sum over

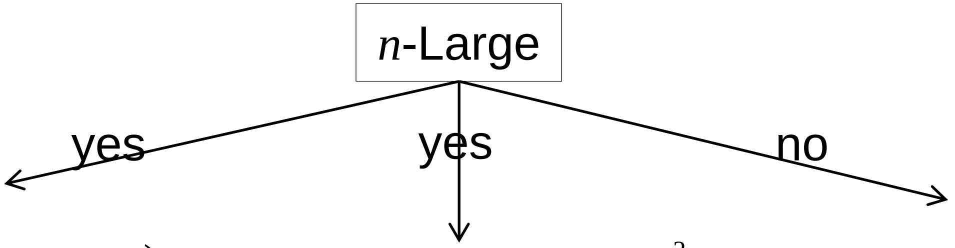
Bartlet test

$$T^* = \frac{(n-k)\ln s_p^2 - \sum_{i=1}^{k} (n_i - 1)\ln s_i^2}{1 + (1/(3(k-1)))((\sum_{i=1}^{k} 1/(n_i - 1)) - 1/(n-k)}$$

Chi-Square distribution

$df = k - 1$

Statistical Inference: Procedures for p_1-p_2



n-Large

yes

yes

no

$$z^* = \frac{(p_1' - p_2') - (p_{0,1} - p_{0,2})}{\sqrt{p_p'(1-p_p')\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \quad \overset{=}{\longleftrightarrow} \quad \chi^{2*} = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

Normal distribution

$$p_1' = \frac{x_1}{n_1} \qquad p_2' = \frac{x_2}{n_2}$$

$$p_p' = \frac{x_1 + x_2}{n_1 + n_2}$$

Chi-Square distribution
$df=k$-1

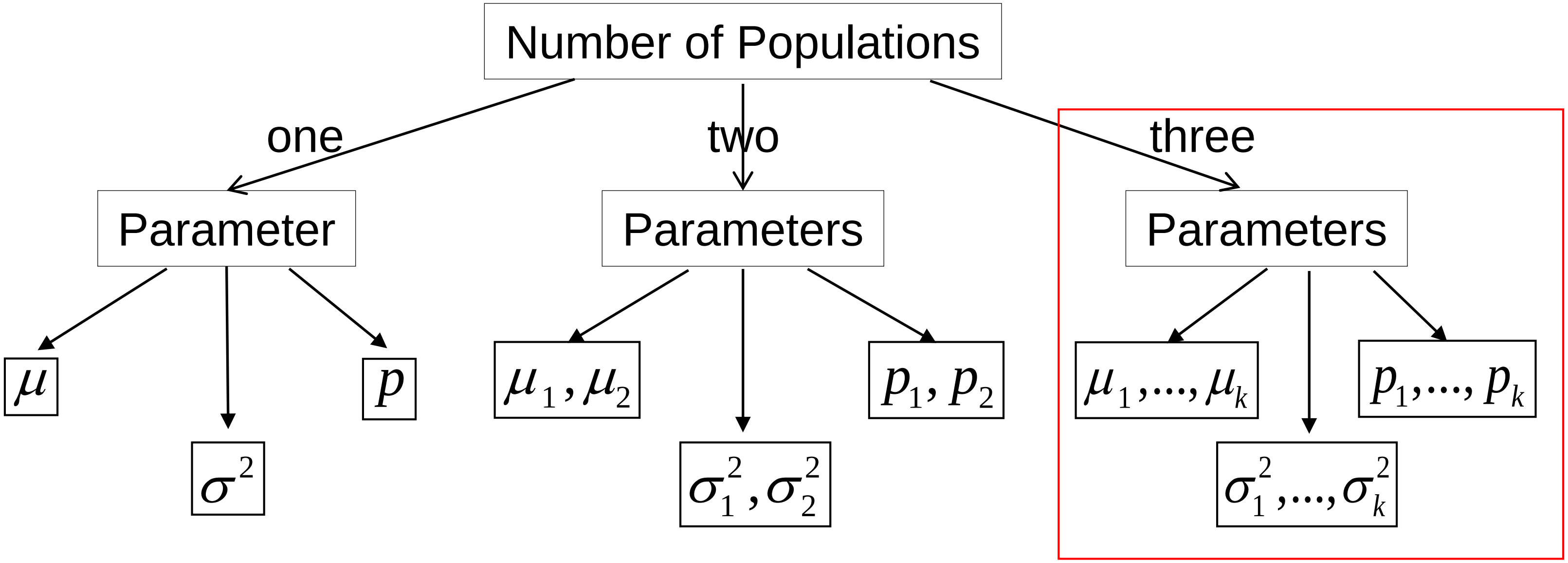
Fisher's exact test

$$p_1' - p_2'$$

$$P(x_1, x_2) = \frac{\binom{x_1 + x_2}{x_1}\binom{n_1 - x_1 + n_2 - x_2}{n_1 - x_1}}{\binom{n_1 + n_2}{n_1}}$$

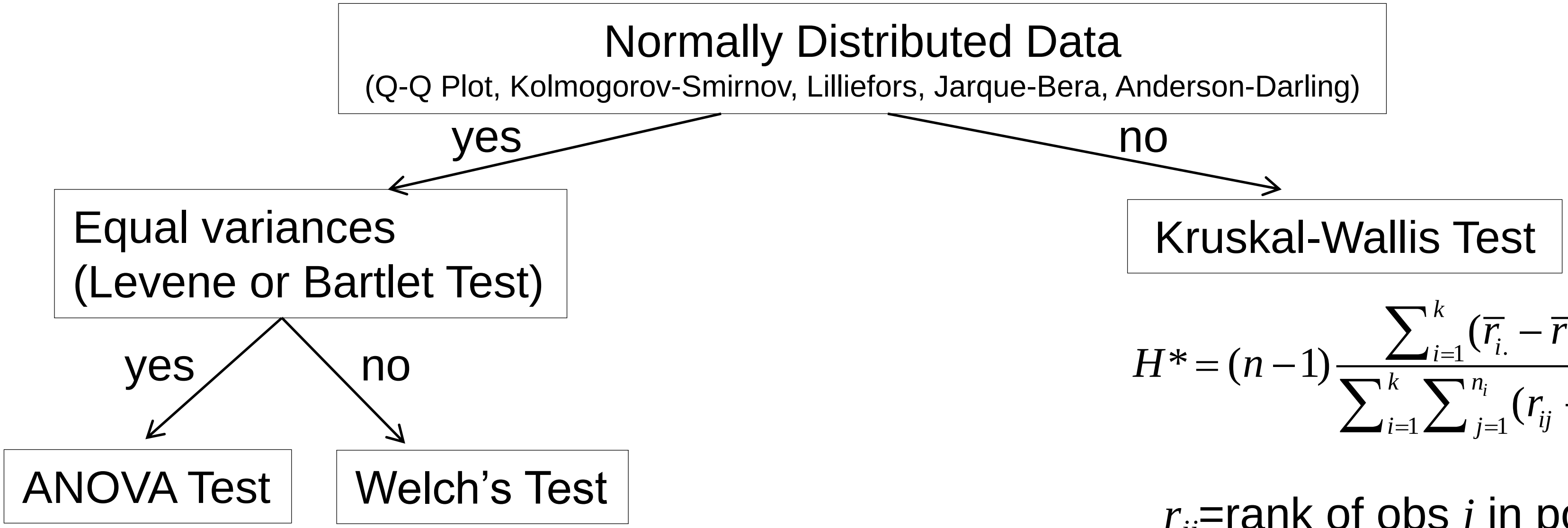
Sum probabilities of more
extreme values to get p-value

Statistical Inference: k Populations



Statistical Inference: Procedures for $\mu_1,\ldots,\mu_k$

Assume or Test for Independent Observations



Normally Distributed Data

(Q-Q Plot, Kolmogorov-Smirnov, Lilliefors, Jarque-Bera, Anderson-Darling)

yes

Equal variances

(Levene or Bartlet Test)

yes

ANOVA Test

no

Welch's Test

no

Kruskal-Wallis Test

$$F^* = \frac{s_1^2}{s_2^2}$$

F-distribution

$df_1 = n_1 - 1$

$df_2 = n_2 - 1$

$$F^* = \frac{s_1^2}{s_2^2}$$

F-distribution

$df_1 = n_1 - 1$

$df_2 = 1/\Lambda$

$$\Lambda = \frac{3\sum_{j=1}^{k}\frac{\left(1-\frac{w_j}{\sum_{j=1}^{k} w_j}\right)^2}{n_j-1}}{k^2-1}$$

$$H^* = (n-1)\frac{\sum_{i=1}^{k}(\bar{r}_{i.}-\bar{r})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(r_{ij}-\bar{r})^2}$$

r_{ij}=rank of obs j in pop i

$$\bar{r}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} r_{ij}$$

$$\bar{r} = \frac{1}{2}(n+1)$$

Special distribution and table.

Statistical Inference: Procedures for $\sigma_1^2,\ldots,\sigma_k^2$

Assume or Test for Independent Observations

Normally Distributed Data

(Q-Q Plot, Kolmogorov-Smirnov, Lilliefors, Jarque-Bera, Anderson-Darling)

yes

$$F^*=\frac{s_1^2}{s_2^2}$$

F-distribution
$df_1=n_1$-1
$df_2=n_2$-1

no

Levene test

$$W^*=\frac{(n-k)\sum_{i=1}^{k} n_i(\bar{A}_{i.}-\bar{A}_{..})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=}^{n_i}(A_{ij}-\bar{A}_{i.})^2}$$

F-distribution
$df_1=n_1$-1
$df_2=n_2$-1

$$A_{ij}=|\,y_{ij}-\bar{y}_{i.}\,|$$

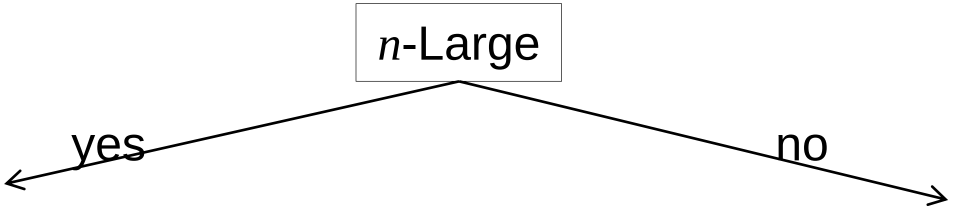
Pop i Obs j
"."=sum over

Bartlet test

$$T^*=\frac{(n-k)\ln s_p^2-\sum_{i=1}^{k}(n_i-1)\ln s_i^2}{1+(1/(3(k-1)))((\sum_{i=1}^{k}1/(n_i-1))-1/(n-k)}$$

Chi-Square distribution
$df=k$-1

Statistical Inference: Procedures for $p_1,\ldots,p_k$



$$\chi^2{*} = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

Chi-Square distribution
$df=k-1$

Freeman-Halton exact test

or

p-value via simulation

Statistical Inference

Hypothesis tests also exist for:
Regression coefficients (i.e. slope and y-intercept)
Correlation Coefficient
Temporal autocorrelation
Correlation Matrices
Two-way ANOVA
…

Statistical Inference

Questions?