Al-Farabi Kazakh National University

Statistical methods in epidemiology

FA.Iskakova

Department of Epidemiology, Biostatistics and Evidence-Based Medicine

2020

Objectives:

- Measures of disease occurrence;
- Measures of association between risk factors and health outcomes
- Estimation of attributable fractions of disease
- Demonstrate proficiency in making refined interpretations of statistical results

Epidemiological statistics

- **Epidemiology:** The branch of medicine dealing with the **incidence and prevalence of disease in large populations and with detection of the source and cause** of epidemics of infectious disease.

- **Epidemiology statistics:** Epidemiological statistics is the science that is primarily concerned with **making inferences about population parameters using sampled measurement, statistical methods** provide the tools for epidemiological research.

USING CATEGORICAL DATA IN MEASURE OF DISEASE OCCURENCE

- the **proportions** (or **percentages**)
- **rate**

Total number of infants	Number of infants with colic	Proportion	Percentage
360	68	68/360 = 0.188	(68/360)x100 = 18.8

From this sample the **proportion** of infants with colic is 0.188 and the equivalent percentage is 18.8%
The rate of colic is 0.188 or 18.8 per 100, or 188 per 1000.

USING OF Numerical Data

Summarizing numeric data depend on the distribution of the data

- **The Mean** is the most widely known measure of average

For example, there are 60 children who were drug withdrawn at birth, to calculate their mean we need to add together the 60 CD4 measurements from these children:

$$Mean = \frac{0.39 + 0.51 + 0.89 + \cdots + 7.49 + 7.99 + 10.19}{60}$$
$$= 3.256 \times 10^3 cells\ per\ mm^{-3}$$

- **MEDIAN** is the middle value when a data set is ordered from least to greatest.

- **The MODE** is the number that occurs most often in a data set.

Measure of disease occurrence

Ratios	Quantifies the magnitude of one occurrence X, in relation to another event Y as X/Y	e.g Ratio of TB cases in community A to B is 1:10
Proportions	Ratio of TB cases in community A to B is 1:10	e.g proportion of TB cases in community A is 10%
Rates:	a proportion with time element It measure the occurrence of an event overtime	e.g # measles cases in 2000/ # population in 2000

TYPES OF RATES

1. ***Crude rates***: Apply to the total population in a given area
2. ***Specific rates***: Apply to specific subgroups in the population (age, sex etc) or specific diseases
3. ***Standardized rates***: used to permit comparisons of rates in population which differ in structure (e.g age structure)

TYPES OF RATES

MORBIDITY RATES:

- Incidence rates(Cumulative incidence, incidence density)
- Prevalence (Period prevalence, point prevalence)

MORTALITY RATES:

- Crude death rate
- Age-specific mortality rate
- Sex-specific mortality rate
- Cause-specific mortality rate
- Proportionate mortality ratio
- Case fatality rate
- Fetal death rate

Measures of association

Chi-square statistics
OR – ODDS Ratio
RR – Relative Ratio

Chi-square statistics

- Chi-square tests whether there is an association between two categorical variables

Ho: There is no association between row & column variables

Ha: There is an association between row and column variables

Chi-square statistic has a degree of freedom (r-1)(c-1), where r is number of rows & c number of columns

$X^2 = \Sigma \frac{(O - E)^2}{E}$	**O: Observed cells E: Expected cells**	**Expected value = $\frac{\text{(Row total)X(Column total)}}{\text{Grand total}}$**	**$X^2 = \frac{(/ad-bc/-n/2)^2 n}{(a+b)(a+c)(c+d)(b+d)}$**

Odds ratio (OR)

Odds ratio is the ratio of odds of exposure among diseased to odds of exposure among non-diseased

Odds of an event E is the ratio of probability of the event to its complement

Odds of exposure among exposed=a/c

Odds of exposure among non-diseased=b/d

$$OR = \frac{\text{Odds of exposure among diseased}}{\text{Odds of exposure among non-diseased}}$$

OR= (a/c)/(b/d);   OR= ad/bc

Relative risk (RR)

- Expresses risk of developing a diseases in exposed group (a + b) as compared to non-exposed group (c + d)

$$RR = \frac{\text{Incidence (risk) among exposed}}{\text{Incidence (risk) among non-exposed}}$$

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

What does a RR of 2 mean? Thus a relative risk of 2 means the exposed group is two times at a higher risk when compared to non-exposed

Strength of association: High if $RR \geq 3$

Moderate if RR is between 1.5 & 2.9

Weak if RR is between 1.2 & 1.4

Attributable Risk (AR)

- AR indicates how much of the risk is due to /attributable/ to the exposure
- Quantifies the excess risk in the exposed that can be attributable to the exposure by removing the risk of the disease occurred due to other causes

  AR= Risk (incidence) in exposed- Risk (incidence) in non-exposed

  AR= {a/(a+b)} / {c/(c+d)}

  Attributable risk is also called risk difference

- What does attributable risk of 10 mean?

10 of the exposed cases are attributable to the exposure

By removing the exposure one can prevent 10 cases from getting the disease

Attributable risk percent (AR%)

- Estimates the proportion of disease among the exposed that is attributable to the exposure

- The proportion of the disease in the exposed that can be eliminated by eliminating the exposure

- $$AR\% = \left(\frac{\text{Risk in exposed} - \text{Risk in non-exposed}}{\text{Risk in non-exposed}}\right) \times 100\%$$

What does AR% of 10% mean?

10% of the disease can be attributed to the exposure

10% of the disease can be eliminated if we avoid the exposure

Population Attributable Risk (PAR)

- Estimates the rate of disease in total population that is attributable to the exposure

PAR = Risk in population – Risk in unexposed

PAR = ARX prevalence rate of exposure

- Estimates the proportion of disease in the study population that is attributable to exposure and thus could be eliminated if the exposure were eliminated

$$PAR\% = \frac{\text{Risk in population} - \text{Risk in unexposed}}{\text{Risk in population}}$$

- **Possible outcomes in studying the relationship between exposure & disease**

No association	Positive association	Negative association
RR>1	RR=1	RR<1 (fraction)
AR=0	AR>0	AR<0 (Negative)

Common statistical tests

- Independent samples t-test

✓ Used to assess whether a statistically significant difference exists in the mean of a continuous outcome variable between two independent groups.

Common statistical tests

- Paired samples t-test

✓ Used on paired or matched samples; that is, for each data point from one sample there is a corresponding data point from second sample, and both data points are collected from same source.

Common statistical tests

- One-way analysis of variance (ANOVA)

✓ An extension of independent samples t-test; used when you wish to compare at least 3 group means. *"One-way"* indicates a single factor or characteristic (independent variable) is being investigated.

Common statistical tests

- Linear correlation coefficient

✓ Used to determine whether a statistically significant linear relationship exists between two continuous variables (i.e. between pairs of (x, y) data in a sample).

Common statistical tests

- Chi-square test

✓ Used to assess whether an association exists between two categorical variables (or to test whether these two variables are independent of each other).

Exercise #1: Relationship between gender & prevalent hypertension

- What test(s) should be performed?
- Answer: Chi-square/Odds Ratio (OR)
- Why? Chi-square is employed to determine whether an association exists between two categorical variables; the OR shows the ***strength*** and ***direction*** of the association.

Exercise #1: Results

			prevhyp1 No	prevhyp1 Yes	Total
sex1	F	Count	1691	799	2490
		% within sex1	67.9%	32.1%	100.0%
		% within prevhyp1	56.3%	55.9%	56.2%
	M	Count	1313	631	1944
		% within sex1	67.5%	32.5%	100.0%
		% within prevhyp1	43.7%	44.1%	43.8%
Total		Count	3004	1430	4434
		% within sex1	67.7%	32.3%	100.0%
		% within prevhyp1	100.0%	100.0%	100.0%

32.1% F vs 32.5% M have prevhyp1

Chi-Square Tests

	Value	df	p-value
Pearson Chi-Square	.069	1	.793

This is the Chi-Square test value

Since $p > 0.05$, insufficient evidence to conclude gender difference in prevalent hypertension

Risk Estimate

	Value	95% CI Lower	95% CI Upper
OR for sex1 (Female / Male)	1.017	.896	1.155

OR is nonsignificant because 95% CI includes value of one!

Exercise #2: Blood pressure taken on bare arm versus over clothing

- What test should be performed?
- Answer: Paired-samples t-test
- Why? We have related samples of continuous data; that is, the subjects are the same group with two measurements (bare and sleeved arm) collected on each.

Exercise #2: Results

Paired Samples Statistics

	Mean	N	SD	SEM
bare	138.68	25	10.032	2.006
sleeve	140.20	25	10.017	2.003

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
				95% CI				
	Mean	SD	SEM	Lower	Upper			
bare - sleeve	-1.520	3.331	.666	-2.895	-.145	-2.282	24	.032

This is the value of the paired t-test.

Since $p < 0.05$, sufficient evidence to indicate difference in SBP between bare and sleeved arm conditions.

Exercise #3:Total weight loss

- What test should be performed?
- Answer: One-way analysis of variance
- Why? Because we are interested in analyzing differences in the mean of a continuous variable (weight loss) that has four independent groups.

Exercise #3: Results

Descriptives

wtloss

	N	Mean	SD	SEM
group1	5	.7260	1.22712	.54879
group2	5	2.7200	1.81375	.81114
group3	5	1.6340	1.49011	.66640
group4	5	4.6260	1.58281	.70785
Total	20	2.4265	2.05584	.45970

$p < 0.05$ indicates statistically significant results; thus subsequent post-hoc comparisons test needed

ANOVA

wtloss

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	42.218	3	14.073	5.912	.006
Within Groups	38.085	16	2.380		
Total	80.303	19			

This is the ANOVA F- test statistic

Exercise #3: Results

Multiple Comparisons

Dependent Variable: wtloss

Bonferroni

(I) group	(J) group	Mean Difference (I-J)	Sig.
group1	group2	-1.99400	.347
	group3	-.90800	1.000
	group4	-3.90000*	.006
group2	group1	1.99400	.347
	group3	1.08600	1.000
	group4	-1.90600	.411
group3	group1	.90800	1.000
	group2	-1.08600	1.000
	group4	-2.99200*	.044
group4	group1	3.90000*	.006
	group2	1.90600	.411
	group3	2.99200*	.044

*. The mean difference is significant at the 0.05 level.

Higher mean weight loss for group 4 vs: group 1 (p = 0.006) & group 3 (p = 0.044).

Exercise #4: Average steps per day

- What test should be performed?
- Answer: Independent-samples t-test
- Why? The objective is to determine whether there is a difference in average steps per day (continuous outcome) for two independent groups—public versus private HS students.

Exercise #4: Results

Group Statistics

		N	Mean	SD	SEM
steps/ day	Public	30	10791.03	3097.633	565.548
	Private	30	12447.53	2620.132	478.368

	t	df	Sig. (2-tail)	Mean Diff.	95% CI of Diff.	
					Lower	Upper
= variances assumed	-2.236	58	.029	-1656.500	-3139.232	-173.768
= variances not assumed	-2.236	56.447	.029	-1656.500	-3140.101	-172.899

This is the t-test value

Since $p < 0.05$, sufficient evidence of difference in steps per day for public versus private students

Exercise #5: Patients with hypertriglyceridemia

- What test should be performed?
- Answer: Linear correlation coefficient
- Why? *You have two continuous variables and would like to know if they are related*

Exercise #5: Results

Correlations

		chol	trig
chol	Pearson Correlation	1	.650*
	Sig. (2-tailed)		.042
		10	10
		.650*	1
		.042	
		10	10

*. Correlation is significant at the 0.05 level (2-tailed).

Pearson's r = 0.650

Results are significant

Since small sample size, go with Spearman. Also, Spearman value much smaller than Pearson, indicating influence of outlier(s) that make Pearson appear to be larger than it should.

Correlations

			chol	trig
Spearman	chol	Correlation Coefficient	1	.418
		Sig. (2-tailed)		.229
		N		10
	trig	Correlation Coe		1
		Sig. (2-tailed)	.229	
		N	10	10

r_s (Spearman rho) = 0.418

Results are nonsignificant

Which correlation should be reported?

Review Problem #1: Glucose concentration in the eyes of dogs

A. What test should be performed?

- Answer: Paired samples t-test; this test compares two means that are from the same individual, object, or related units.

B. Interpretation of 95% CI:(-0.728, 1.288)?

- Answer: Since CI includes zero (value specified in the null hypothesis), insufficient evidence to claim a difference exists in the mean glucose concentrations between the two eyes. Results are not statistically significant.

Review Problem #2: Tamoxifen and cancer

A. 2 x 2 table:

	Breast Cancer		
Treatment	Yes	No	Totals
Tamoxifen	89	6592	6681
Placebo	175	6532	6707
Totals	264	12124	13388

Review Problem #2: Tamoxifen and cancer

B. Test to determine relationship between treatment and cancer?

- Answer: Chi-square/Relative Risk (RR)

- Why? Chi-square is employed to determine whether an association exists between two categorical variables; the RR shows the ***strength*** and ***direction*** of the association.

Review Problem #2: Tamoxifen and cancer

C. Calculate and interpret epidemiologic measure of association.

- Answer: RR

$$\mathbf{RR} = \frac{I_{exposed}}{I_{unexposed}} = \frac{\left(\frac{89}{6681}\right)}{\left(\frac{175}{6707}\right)} = \frac{1.332\%}{2.609\%} = 0.5106$$

<u>Interpretation</u>: Women on tamoxifen have a 49% reduced risk of breast cancer versus women on placebo (RR = 0.5106; 95% CI= (0.3965, 0.6575). Significant protective effect exists since CI excludes one.

Review Problem #3: Sample computer output

Describe the example and conclusion based on the computer output.

An independent samples t-test was performed to determine whether a difference exists in mean number of drinks in previous week for treatment versus controls. From 95% CI, we can conclude treatment group (n=244, M=13.62, SD=12.39) consumed anywhere between 0.92 to 5.56 fewer drinks than controls (n= 238, M=16.86, SD=13.49). Results are statistically significant since zero is excluded from CI.

Review Problem #4: Intracellular calcium & blood pressure

Independent samples t-test; comparing mean of continuous variable (calcium concentration) between two independent groups (normal versus high blood pressure).

Review Problem #4: Results from OpenEpi

Two-Sample Independent *t* Test

Must present results from "unequal variance" row.

Input Dat

This small p-value indicates equal variances cannot be assumed!

Two-sided confidence interval 95%

	Sample size	Mean	Std. Dev.	Std. Error
Group-1	38	107.9	16.1	
Group-2	45	168.2	31.7	

Result	*t* statistics	*df*	p-value[1]	Mean Difference	Lower Limit	Upper Limit
Equal variance	-10.6195	81	<0.0000001	-60.3	-71.598	-49.002
Unequal variance	-11.1682	67	<0.0000001	-60.3	-71.0771	-49.5229

	F statistics	*df*(numerator,denominator)	p-value[1]
Test for equality of variance[2]	3.87674	44,37	0.00005305

Statistically significant difference exists in platelet calcium concentration between participants with normal (M=107.9 nM, SD=16.1) versus high (M=168.2 nM, SD=31.7) blood pressure; $t_{\text{unequal variances}}$ (67) = -11.168, p <0.001).

References

- Gordis: Epidemiology, 5th Edition, Saunders 2013
- Lectures of Jhon Hopkins University, Bloomberg School of Public Health
- Wolfgang, A. Handbook of Epidemiology. Vol.1//Ahrens Wolfgang, Peugeot Iris. - 2 ed.- Springer Reference, 2014.- 469 p.
- Principles and methods of Epidemiology. 3-d Edition. R. Dicker Ooffice of epidemiologic program СДС, USAID. -2012.-457 P.
- Statistical methods in epidemiology by Michael A. Joseph, PhD, MPH Associate Professor and Vice Chair, Department of Epidemiology & Biostatistics, SUNY Downstate School of Public Health