PUBH620 Biostatistics

Week 7

Relative Risk and Odds Ratio

Lecture notes by Dr Brandon Cheong

ELECTRONIC WARNING NOTICE

Commonwealth of Australia
Copyright Act 1968
Form of notice for paragraph 135KA (a) of the *Copyright Act 1968*
Warning
This material has been copied and communicated to you by or on behalf of *Australian Catholic University under Part VA of the Copyright Act 1968 (the **Act).***
The material in this communication may be subject to copyright under the Act. Any further copying or communication of this material by you may be the subject of copyright or performers' protection under the Act.
Do not remove this notice.

In this lecture…

(clicking the links below will direct you to the topic page)

Topic Learning Objectives (TLOs)

- Understand the definition of relative risk and how it is used in a public health context
- Understand the definition of odds ratio and how it is used in a public health context

7.1 Prospective and Retrospective Studies

In order to understand Relative Risk and Odds Ratios, you must first understand the contexts in which they are used in.

This means understanding the difference between prospective and retrospective studies. Generally speaking, a prospective study is one that looks forward (in to the future). So this may be observation of the development of disease over time or looking at the impact of an intervention at various time points and comparing two or more groups. A retrospective study on the other hand is a study that looks backward (to the past). This might be a case-control study where the status of the disease is known but factors related to the disease are unknown and so historical data (ie. Hospital data and registries) need to be accessed in order to conduct the study.

In many public health studies, we are often interested in looking at the exposure of disease and comparing it to a group that have been unexposed to the disease.

The two main methods we use to quantify the risk are:

Relative Risk – which is appropriate for cross-sectional, randomised controlled trials (RCTs) and cohort studies (prospective studies) and

Odds Ratio – which is appropriate for case-control studies (retrospective study).

If you haven't done so already, watch this 9 minute 42 second video to get a quick explanation of common research study designs in public health.

https://www.youtube.com/watch?v=Jd3gFT0-C4s&feature=emb_logo

7.1 Prospective and Retrospective Studies

An example of a prospective study is a randomised controlled trial (RCT). When conducting research on a disease, there is usually an exposed group of patients (those that contain the risk factor) and an unexposed group of patients (those that do not contain the risk factor). A prospective study is one where patients are investigated into the future (prospectively) and data is collected on the number of patients in each sample, over various time points.

Some examples of this might be:

A study where pregnant women at high risk of gestational diabetes are participating in a RCT that compares a dietary intervention with a control group during the third trimester of pregnancy. Each patient is followed prospectively and the number of patients who develop gestational diabetes is recorded.

A study where patients who have been exposed to an influenza virus are monitored over time to see if interventions can be used to prevent the development of pneumonia.

A hypertension study where pharmaceutical interventions are used to prevent patients at risk of developing cardiovascular disease.

Other examples of prospective studies include cross-sectional and cohort studies.

7.1 Prospective and Retrospective Studies

An example of a retrospective study is a case-control study. Unlike a prospective study where patients are followed in to the future, a retrospective study requires the researchers to look back (retrospectively) at patients who may already contain the disease and determine the causes or risk factors leading to the development of the disease.

An example of this might be:

Determining whether smoking is a risk factor of a particular disease. A case-control study employs “cases” (those who have the disease) and “controls” (those who do not have the disease). Each patient would be asked their smoking status retrospectively and these were recorded. A research question for this might be “Do smokers have a higher chance of getting the disease?”.

7.2 Relative Risk

The Relative Risk (RR) is essentially a ratio of the risk of developing the disease among patients in the exposed group to the risk of developing the disease among the patients in the unexposed group.

To understand this, we need to calculate the proportions of patients who have the disease in the exposed and unexposed groups. Then we divide the proportion for the exposed group by that of the unexposed group to get the required RR.

It helps to use a contingency table to help with visualisation:

7.2 Relative Risk

Risk Factor	Outcomes		Total
	Disease	Healthy	
Exposed	d_e	h_e	$d_e + h_e$
Unexposed	d_u	h_u	$d_u + h_u$
Total	$d_e + d_u$	$h_e + h_u$	$N = d_e + h_e + d_u + h_u$

Note here that the outcome variable is binary (there are only two categories), this is essential as understanding risks and odds ratios will help you to understand binary logistic regression, which is where the dependent variable (outcome) is binary. Having said this however, the exposure variable could have two or more categories but for simplicity's sake, the table above shows only one binary exposure (exposed or unexposed). N is the total number of patients and we can see that d_e represents those who have developed the disease and were exposed to the disease and h_e represents those who were exposed to the disease but remained healthy. From this table, we can determine the proportions of those who were exposed to the disease to those who weren't exposed to it.

7.2 Relative Risk

The formula to calculate relative risk is:

$$\mathrm{RR} = \frac{d_e/(d_e + h_e)}{d_u/(d_u + h_u)} = \frac{d_e(d_u + h_u)}{d_u(d_e + h_e)}$$

The proportions can be determined by realising the number of diseased patients who were exposed is $d_e/(d_e+h_e)$ hence the proportion for number of diseased patients who were unexposed would be $d_u/(d_u+h_u)$. The relative risk is simply the ratio of these two quantities. Often we find RR ranges between 0 and a positive value but it can generally have three possibilities.

- RR = 1 indicates that the risk is the same in both the exposed and unexposed groups.

- RR > 1 indicates risk is greater in the exposed group than the unexposed group

- RR < 1 indicates risk is less in the exposed group than the unexposed group.

7.2 Relative Risk

Now that we know how to calculate the RR, we need to find whether this value is significant or not. In order to do this we can conduct a hypothesis test and calculate the 95% confidence interval. One thing to note however, is that RR does not follow a normal distribution so in order to make this viable, we need to do a logarithmic transformation. This will make it easier to conduct the hypothesis test. In this case, a natural logarithm (*base 'e'* often denoted by '*ln*') is used.

The formula for the 95% confidence interval for logarithm of RR is:

- 95% CI for ln(RR): ln(RR) $\pm$ 1.96 x SE (standard error)
- We remember that for a normal distribution with α = 0.05, two-tailed test, 1.96 is the value to use for a normal distribution.
    - Lower limit for ln(RR) = ln(RR) – 1.96 x *SE*
    - Upper limit for ln(RR) = ln(RR) + 1.96 x *SE*

    - *But how do we calculate SE?*

7.2 Relative Risk

The formula for SE of ln(RR) is:

$$SE = \sqrt{\frac{1}{d_e} - \frac{1}{d_e+h_e} + \frac{1}{d_u} - \frac{1}{d_u+h_u}}$$

Remember that since we did a log transform of RR, the 95% CI formula gives confidence intervals for the RR in terms of logarithms of the RR. This doesn't make a lot of sense to us so we need to do an additional mathematical operation. The *actual* confidence interval for RR is obtained by anti-logging (e^x or exp) the lower and upper limits. Thus, the lower and upper limits of the confidence interval for RR are obtained as:

- Lower limit for RR = exp[lower limit of ln(RR)]
- Upper limit for RR = exp[upper limit for ln(RR)]

7.2 Relative Risk

One way to understand the significance is by looking at the range of the confidence interval. If the range does not include the value of 1, then we can say that we have statistical significance for the risk of developing disease in the exposed and unexposed groups. Be reminded that a RR value that is different than 1 represents either a higher or lower risk of having the disease.

In order to calculate the *p*-value, we need to establish a null hypothesis.

- Null hypothesis: There is no difference in the risk of disease between exposed and unexposed groups in the study population. In other words, study population RR = 1.
- Alternative hypothesis: There is a difference in the risk of disease between exposed and unexposed groups in the study population. In other words, the study population RR $\neq$ 1

The test statistic for testing these hypotheses is the Z-statistic, which follows the standard normal distribution and is given by:

$$Z\text{-statistic} = \ln(RR)/SE$$

7.2 Relative Risk

Up until now we haven't needed to calculate the p-value ourselves because SPSS does this for us. However, we will see that when running risks in SPSS, we do not get a p-value and need to use a table (z-table) to determine it. The p-value is obtained using the normal distribution probability table, shown on the next slide.

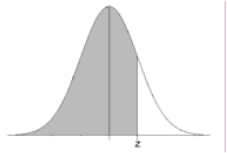
As always, if the p-value is less than 0.05, we reject the null hypothesis and say that the RR is significantly different from 1. If the p-value is greater than 0.05, we accept the null hypothesis and say that the RR is not significantly different from 1.

Once we have a z-statistic, we can determine the p-value. Let's take an arbitrary value of 1.74 for our z-statistic to see how we can get the p-value.

Remember we have a two-tailed test, which looks like this:

7.2 Relative Risk

Standard Normal Cumulative Probability Table



Cumulative probabilities for POSITIVE z-values are shown in the following table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

https://www.math.arizona.edu/~jwatkins/normal-table.pdf

Normal distribution probability table

For z-statistic of 1.74, the supposed p-value is 0.9591 however, this value represents the shaded area in a normal distribution and we are interested in the tails. This means we need to subtract 0.9591 from 1 and multiply the result by 2 (for two tail test).

Hence, 1 – 0.9591 = 0.0409 and multiply this value by 2 gives us a true p-value of 0.082. This value is greater than 0.05 hence we do not have statistical significance.

Another table we could use actually shows the p-value for a two-tailed test directly.

Note also that this link provided has a table for negative z-values.

7.2 Relative Risk

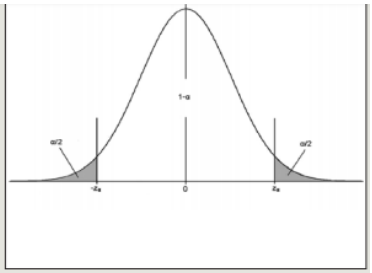


TABLE 1. Extract from two-tailed standard Normal table. Values tabulated are *P*-values corresponding to particular cut-offs and are for *z* values calculated to two decimal places.

TABLE 1

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	1.0000	0.9920	0.9840	0.9761	0.9681	0.9601	0.9522	0.9442	0.9362	0.9283
0.10	0.9203	0.9124	0.9045	0.8966	0.8887	0.8808	0.8729	0.8650	0.8572	0.8493
0.20	0.8415	0.8337	0.8259	0.8181	0.8103	0.8206	0.7949	0.7872	0.7795	0.7718
0.30	0.7642	0.7566	0.7490	0.7414	0.7339	0.7263	0.7188	0.7114	0.7039	0.6965
0.40	0.6892	0.6818	0.6745	0.6672	0.6599	0.6527	0.6455	0.6384	0.6312	0.6241
0.50	0.6171	0.6101	0.6031	0.5961	0.5892	0.5823	0.5755	0.5687	0.5619	0.5552
0.60	0.5485	0.5419	0.5353	0.5287	0.5222	0.5157	0.5093	0.5029	0.4965	0.4902
0.70	0.4839	0.4777	0.4715	0.4654	0.4593	0.4533	0.4473	0.4413	0.4354	0.4295
0.80	0.4237	0.4179	0.4122	0.4065	0.4009	0.3953	0.3898	0.3843	0.3789	0.3735
0.90	0.3681	0.3628	0.3576	0.3524	0.3472	0.3421	0.3371	0.3320	0.3271	0.3222
1.00	0.3173	0.3125	0.3077	0.3030	0.2983	0.2837	0.2891	0.2846	0.2801	0.2757
1.10	0.2713	0.2670	0.2627	0.2585	0.2543	0.2501	0.2460	0.2420	0.2380	0.2340
1.20	0.2301	0.2263	0.2225	0.2187	0.2150	0.2113	0.2077	0.2041	0.2005	0.1971
1.30	0.1936	0.1902	0.1868	0.1835	0.1802	0.1770	0.1738	0.1707	0.1676	0.1645
1.40	0.1615	0.1585	0.1556	0.1527	0.1499	0.1471	0.1443	0.1416	0.1389	0.1362
1.50	0.1336	0.1310	0.1285	0.1260	0.1236	0.1211	0.1188	0.1164	0.1141	0.1118
1.60	0.1096	0.1074	0.1052	0.1031	0.1010	0.0989	0.0969	0.0949	0.0930	0.0910
1.70	0.0891	0.0873	0.0854	0.0836	0.0819	0.0801	0.0784	0.0767	0.0751	0.0735
1.80	0.0719	0.0703	0.0688	0.00672	0.0658	0.0643	0.0629	0.0615	0.0601	0.0588
1.90	0.0574	0.0561	0.0549	0.0536	0.0524	0.0512	0.0500	0.0488	0.0477	0.0466
2.00	0.0455	0.0444	0.0434	0.0424	0.0414	0.0404	0.0394	0.0385	0.0375	0.0366
2.10	0.0357	0.0349	0.0340	0.0332	0.0324	0.0316	0.0308	0.0300	0.0293	0.0285
2.20	0.0278	0.0271	0.0264	0.0257	0.0251	0.0244	0.0238	0.0232	0.0226	0.0220
2.30	0.0214	0.0209	0.0203	0.0198	0.0193	0.0188	0.0183	0.0178	0.0173	0.0168
2.40	0.0164	0.0160	0.0155	0.0151	0.0147	0.0143	0.0139	0.0135	0.0131	0.0128
2.50	0.0124	0.0121	0.0117	0.0114	0.0111	0.0108	0.0105	0.0102	0.0099	0.0096
2.60	0.0093	0.0091	0.0088	0.0085	0.0083	0.0080	0.0078	0.0076	0.0074	0.0071
2.70	0.0069	0.0067	0.0065	0.0063	0.0061	0.0060	0.0058	0.0056	0.0054	0.0053
2.80	0.0051	0.0050	0.0048	0.0047	0.0045	0.0044	0.0042	0.0041	0.0040	0.0039
2.90	0.0037	0.0036	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028
3.00	0.0027	0.0026	0.0025	0.0024	0.0024	0.0023	0.0022	0.0021	0.0021	0.0020

https://www.sheffield.ac.uk/polopoly_fs/1.43999!/file/tutorial-10-reading-tables.pdf

Normal distribution probability table

As can be seen, we can confirm that the p-value is indeed 0.082 with this table. Therefore, we accept the null hypothesis and say that there is no difference in the risk of disease between exposed and unexposed groups. ie. RR = 1.

7.2 Relative Risk

Example

Let's consider an example of a gestational diabetes study. 78 pregnant women at high risk of gestational diabetes participate in a RCT comparing a dietary treatment program to a non-treatment program during the third trimester (typically 28 to 40 weeks) of pregnancy. 7 out 41 patients on the diet program developed gestational diabetes, and 20 out of 37 patients without treatment developed gestational diabetes. The research question of interest is whether pregnant women at risk of gestational diabetes can lower the risk of gestational diabetes through a dieting program.

This data is summarised in a table below:

Treatment	Gestational Diabetes		Total
	Developed	Did not develop	
Dietary program	7 (d_e)	34 (h_e)	**41**
Non-treatment program	20 (d_u)	17 (h_u)	**37**
Total	**27**	**51**	**78**

7.2 Relative Risk

Let's first calculate the RR

In this study, 7 out of the 41 patients on the diet program developed gestational diabetes and 20 out of 37 not on the diet program developed gestational diabetes. If we were to summarise this in regards to the RR formula…

$\mathrm{RR} = \frac{d_e(d_u+h_u)}{d_u(d_e+h_e)}$

$d_e = 7$, $h_e = 34$, therefore $d_e + h_e = 41$

$d_u = 12$, $h_u = 25$, therefore $d_u + h_u = 37$

$$\mathrm{RR} = \frac{7\ (37)}{20\ (41)} = 0.3158536585$$

The RR is obtained by dividing the risk in the exposed group by the risk in the unexposed group, that is, RR = 0.3158536585 (good to use all decimal values calculator gives you to get more accurate result).

If we were to take the reciprocal of this RR value, we can interpret the meaning of RR.

1/0.3158536585 = 3.17

This means the risk of developing gestational diabetes for non-treatment pregnant women is approximately 3 times higher than those who were on diet. Hence, we can say that the dieting program helps to reduce the risk of having gestational diabetes.

7.2 Relative Risk

Calculating the 95% Confidence Interval

The 95% CI for the RR is obtained as follows:

- 95% CI for ln(RR): ln(RR) $\pm$ 1.96 x SE
- RR = 0.316
- ln(RR) = -1.152476279
- The standard error of the natural logarithm of the RR is:

$$SE = \sqrt{\frac{1}{7} - \frac{1}{41} + \frac{1}{20} - \frac{1}{37}} = 0.3760849265$$

  - Lower limit for ln(RR) = ln(RR) – 1.96 x *SE*
  - Upper limit for ln(RR) = ln(RR) + 1.96 x *SE*
- Lower limit for ln(RR) = -1.152476279 – 1.96 x 0.3760849265 = -1.8896
- Upper limit for ln(RR) = -1.152476279 + 1.96 x 0.3760849265 = -0.4153

7.2 Relative Risk

But remember, we need to use the exponential (anti-log) of these values to get the true lower and upper limits…

Upper limit = exp(-1.8896) = 0.1511

Lower limit = exp(-0.4153) = 0.6601

Hence the confidence interval is (0.151, 0.660)

This means that we are 95% confident that the risk of developing gestational diabetes among patients on a diet is between 0.151 and 0.660 times less compared to patients who were not on a diet program. This interval excludes the value of 1, hence the risk of developing gestational diabetes is significant.

7.2 Relative Risk

Determining the p-value from a z-statistic

The significance of the RR (0.316) can be evaluated using a hypothesis test. The null hypothesis to be tested is that the population risks of developing gestational diabetes in both groups are the same against the alternative hypothesis that the risks are different. The test statistic is:

Z-statistic = ln(RR)/SE

= -1.152476279 /0.3760849265

= -3.06

Looking at our table of negative z-values (https://www.math.arizona.edu/~jwatkins/normal-table.pdf),

we see that the p-value is 0.0011 but remember we need to multiply this by 2 to get the true p-value for a two-tailed test hence our p-value = 0.0022. As this is < 0.05, we reject the null hypothesis.

Interpreting and reporting the results:

The risk of developing gestational diabetes among the patients on a diet program is significantly lower than patients on a non-treatment program (RR = 0.32, p-value = .002, 95% CI: 0.15 – 0.66). Therefore we can say that dieting reduces the risk of gestational diabetes in pregnant women in their third trimester for this study population.

7.3 Odds Ratio

The Odds Ratio (OR) is generally associated with case control research designs. In case control studies the data can also be summarised in a table however, there is a key difference between case control studies and randomised controlled trials. In a RCT, patients are randomly allocated to a group whether this is a treatment group or a control group. In a case control study, there are ‘cases’ (people who have the disease) and there are ‘controls’ (people who don’t have the disease). Thus the selection of patients is based on their disease status. Case control studies are retrospective and hence we calculate the OR as opposed to the RR.

Risk Factor	Outcomes		Total
	Diseased	Healthy	
Exposed	d_e	h_e	$d_e + h_e$
Unexposed	d_u	h_u	$d_u + h_u$
Total	$d_e + d_u$	$h_e + h_u$	$N = d_e + h_e + d_u + h_u$

7.3 Odds Ratio

When it comes to defining the Odds Ratio, we are essentially looking at the odds of cases among the exposed and unexposed groups. In other words, the odds of an event can be defined as the proportion of the event happening over the proportion of the event not happening. Here we realise we are just taking ratios over ratios over ratios! Let's take a look:

In the exposed group the odds of cases would be d_e/h_e

In the unexposed group the odds of cases would be d_u/h_u

We therefore realise that the OR works out to be:

$$OR = \frac{d_e/h_e}{d_u/h_u} = \frac{d_e h_u}{h_e d_u}$$

Similar to RR, the OR ranges from zero to some positive number but generally we encounter three scenarios:

7.3 Odds Ratio

- An OR of 1 indicates that the odds of cases are the same in the exposed and unexposed groups. No association between exposure and outcome.
- An OR of greater than 1 shows that the odds of cases are higher in the exposed group than the odds of cases in the unexposed group.
- An OR of less than 1 means that the odds of cases in the exposed group are less than the odds of cases in the unexposed group.

We can evaluate the significance of OR by calculating the 95% confidence intervals and the p-value by performing a hypothesis test. Similar to RR, the sampling distribution of OR is not normally distributed; however, the natural logarithm of OR (ln(OR)) follows the normal distribution. The SE for the ln(OR) is:

$$SE = \sqrt{\frac{1}{d_e} + \frac{1}{h_e} + \frac{1}{d_u} + \frac{1}{h_u}}$$

7.3 Odds Ratio

Thus the 95% confidence interval for the ln(OR) is obtained using the following formulas:

- 95% CI for ln(OR): ln(OR) $\pm$ 1.96 x SE
- Lower limit (for ln(OR)) = ln(OR) – 1.96 x SE.
    - Hence lower limit for OR = exp[lower limit for ln(OR)]
- Upper limit (for ln(OR)) = ln(OR) + 1.96 x SE.
    - Hence upper limit for OR = exp[upper limit for ln(OR)]

Remember that we need to use the anti-log operation so that the confidence interval ranges make sense!

7.3 Odds Ratio

Example

What are the odds of developing heart disease for smokers and non-smokers? How do we quantify the odds?

Smoking status	Disease Status		Total
	Case	Control	
Smoker	63	22	**85**
Non-smoker	20	97	**117**
Total	**83**	**119**	**202**

7.3 Odds Ratio

First let's calculate OR:

$$\text{OR} = d_e h_u / h_e d_u$$

$$= (63 \times 97) / (22 \times 20)$$

$$= 13.9$$

This means that the odds of developing cardiovascular disease among smokers are 13.9 times higher than for non-smokers.

7.3 Odds Ratio

Now let's calculate the 95% CI for OR

First, we calculate the 95% confidence interval for the logarithm of the odds ratio (ln(OR)) and then take the anti-log (exponential) of this interval for obtaining the confidence interval for the odds ratio. The steps are as follows:

- 95% CI for ln(OR): ln(OR) $\pm$ 1.96 x SE
- OR = 13.9
- ln(OR) = ln(13.9) = 2.63.
- SE = $\sqrt{\frac{1}{d_e}+\frac{1}{h_e}+\frac{1}{d_u}+\frac{1}{h_u}} = \sqrt{\frac{1}{63}+\frac{1}{22}+\frac{1}{20}+\frac{1}{97}} = 0.348764734$
- Lower limit for ln(OR) = 2.63 – 1.96 x 0.348764734 = 1.947492099
    - Lower limit for OR = exp(1.947492099) = 7.0111
- Upper limit for ln(OR) = 2.63 + 1.96 x 0.348764734 = 3.314649857
    - Upper limit for OR = exp(3.314649857) = 27.5128
- The 95% CI is (7.011, 27.513)

We can be 95% confident that the odds of having cardiovascular disease among smokers is between 7.011 and 27.513 times higher than for non-smokers; the interval excludes the value of 1, so the odds of having cardiovascular disease among the smokers is significant.

7.3 Odds Ratio

Determining the p-value from the z-statistic:

A hypothesis test can also be performed for further evidence on significance of OR.

Null hypothesis: There is no difference in the odds of developing the disease among smokers and non-smokers in the study population.

Alternative hypothesis: There is a difference in the odds of developing the disease among smokers and non-smokers in the study population.

The Z statistic:

$$\text{Z-statistic} = \ln(\text{OR})/\text{SE}$$
$$= 2.631/0.348764734$$
$$= 7.55$$

Using the same table we did for RR (https://www.math.arizona.edu/~jwatkins/normal-table.pdf), we can determine the p-value. However, our z-statistic goes beyond the z values on the table so we have to use the largest value on the table, which is z = 3.49 hence our supposed p-value is 0.9998 but remember that this represents the shaded area and so we need to subtract this value from 1 and multiply it by 2 (for two-tailed test). Hence:

P-value = 2 x (1 – 0.9998) = a very small number (4×10^{-4}). Therefore $p < 0.001$ and the null hypothesis is rejected.

7.3 Odds Ratio

Interpreting and reporting OR

The odds of developing cardiovascular disease among smokers is significantly higher than for non-smokers (OR = 13.90, *p*-value < .001, 95% CI: 7.01 – 27.51). Smoking is related to the development of cardiovascular disease in this study population. (Which we probably already knew!)

7.4 Relative Risk and Odds Ratio in SPSS

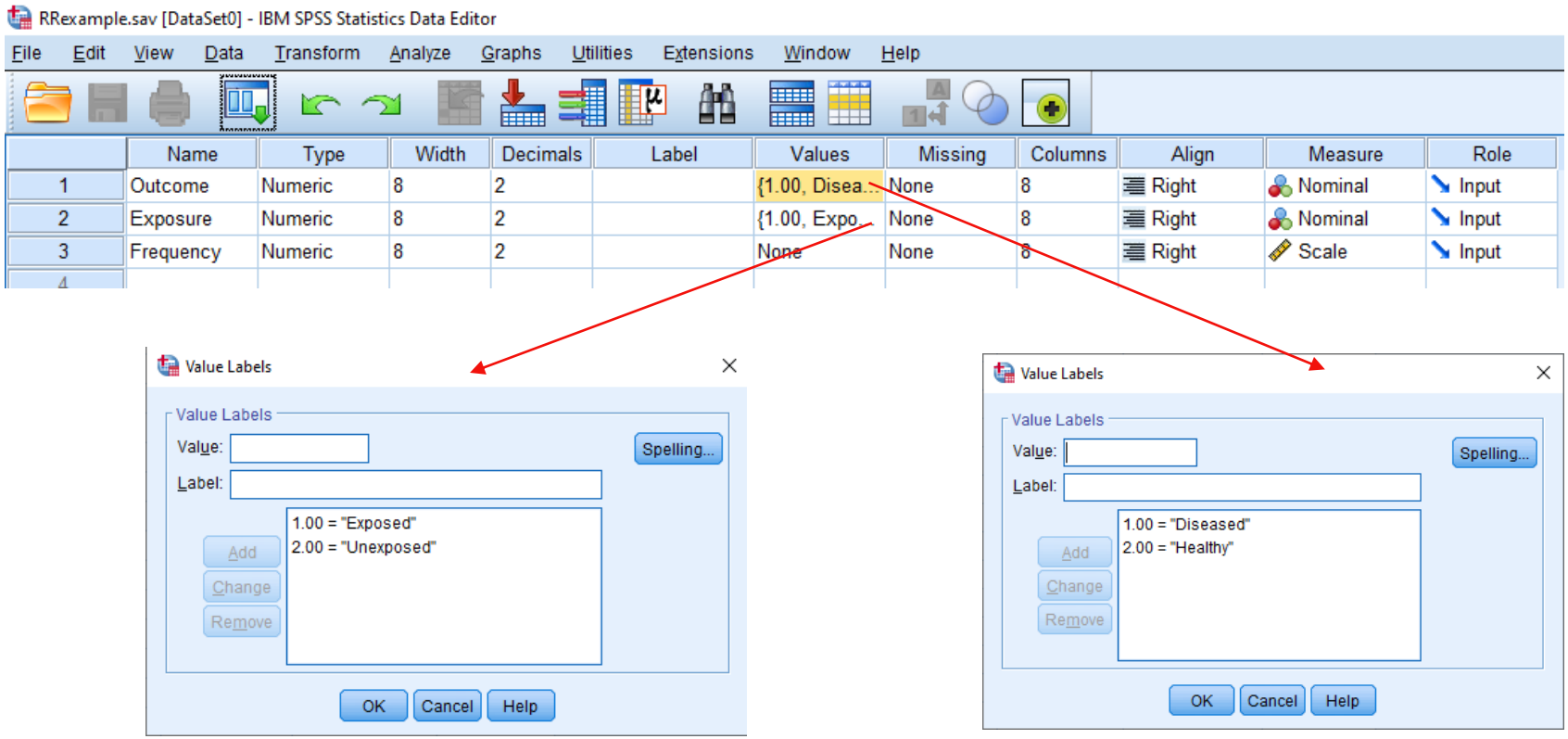
This section will contain how to do RR and OR in SPSS. This section is combined because the same procedure in SPSS is done for both measures of association.

Let's first look at the RR example (gestational diabetes)…recall:

Treatment	Gestational Diabetes		Total
	Developed	Did not develop	
Dietary program	7 (d_e)	34 (h_e)	**41**
Non-treatment program	20 (d_u)	17 (h_u)	**37**
Total	**27**	**51**	**78**

7.4 Relative Risk and Odds Ratio in SPSS

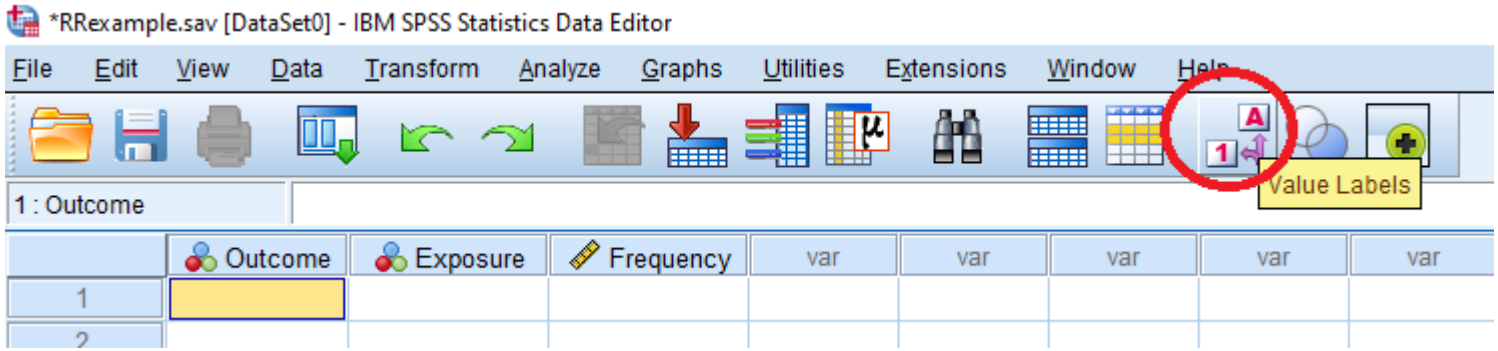
Set this up in SPSS:

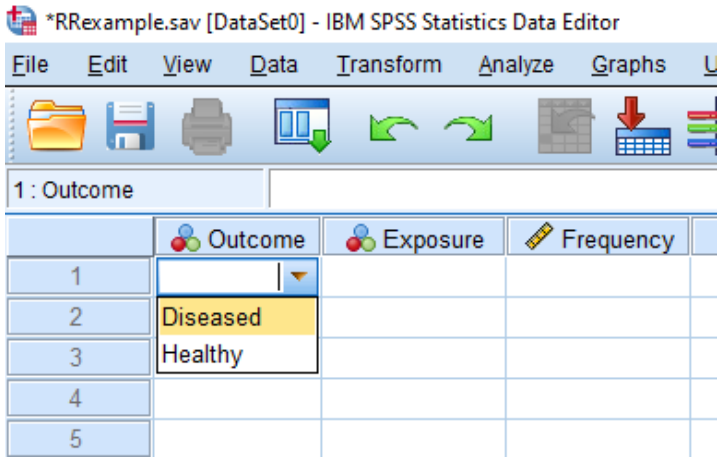


7.4 Relative Risk and Odds Ratio in SPSS

Set this up in SPSS:

This values box allows you to use your value labels you have already set up.

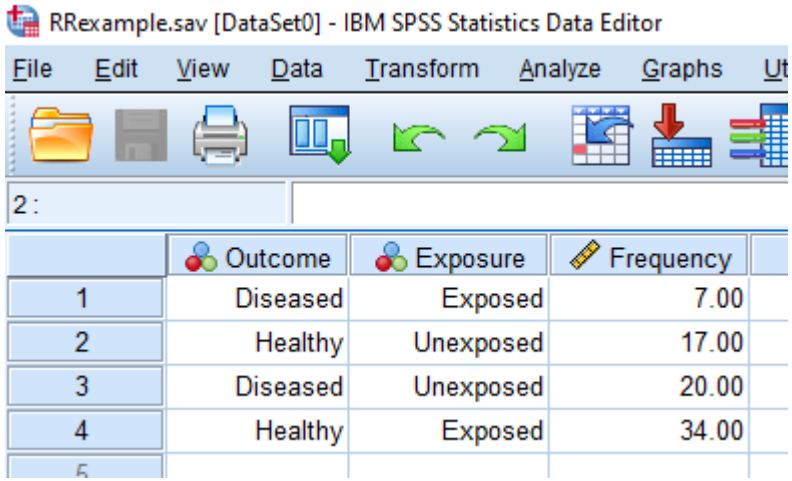




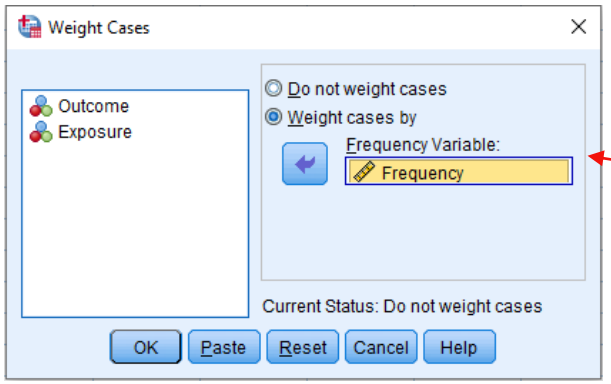
Now when you click the cell in data view, you will see a dropdown menu with your value labels.

7.4 Relative Risk and Odds Ratio in SPSS

Set up your data view like so:



	Outcome	Exposure	Frequency
1	Diseased	Exposed	7.00
2	Healthy	Unexposed	17.00
3	Diseased	Unexposed	20.00
4	Healthy	Exposed	34.00



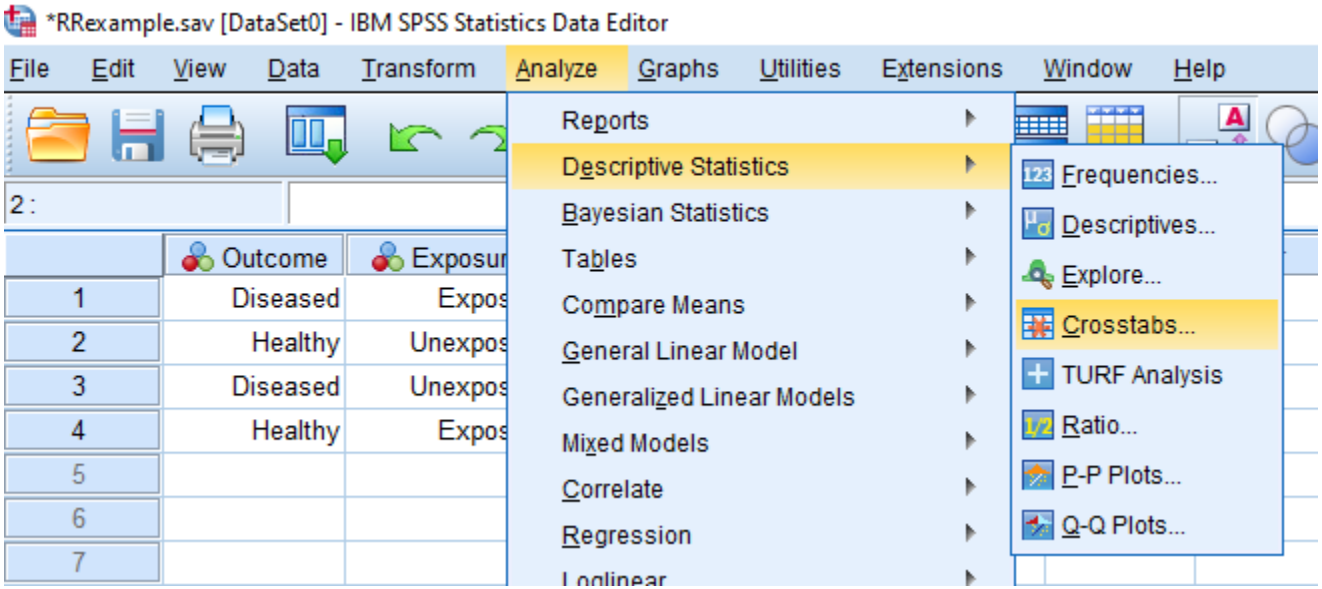
We are using a contingency table here, which might look familiar to you from doing chi-square tests! We have a frequency column and so we need to weight cases to tell SPSS that the numbers actually represent how many cases there are. Another way to explain this is that SPSS sees two “diseased” and two “healthy” and so it would assume that there are only 4 cases in total. By weighing cases we are telling SPSS that the numbers in the Frequency column represent the cases and not how many rows it counts. Go to <Data>, <Weight Cases> and weight cases by Frequency.

7.4 Relative Risk and Odds Ratio in SPSS

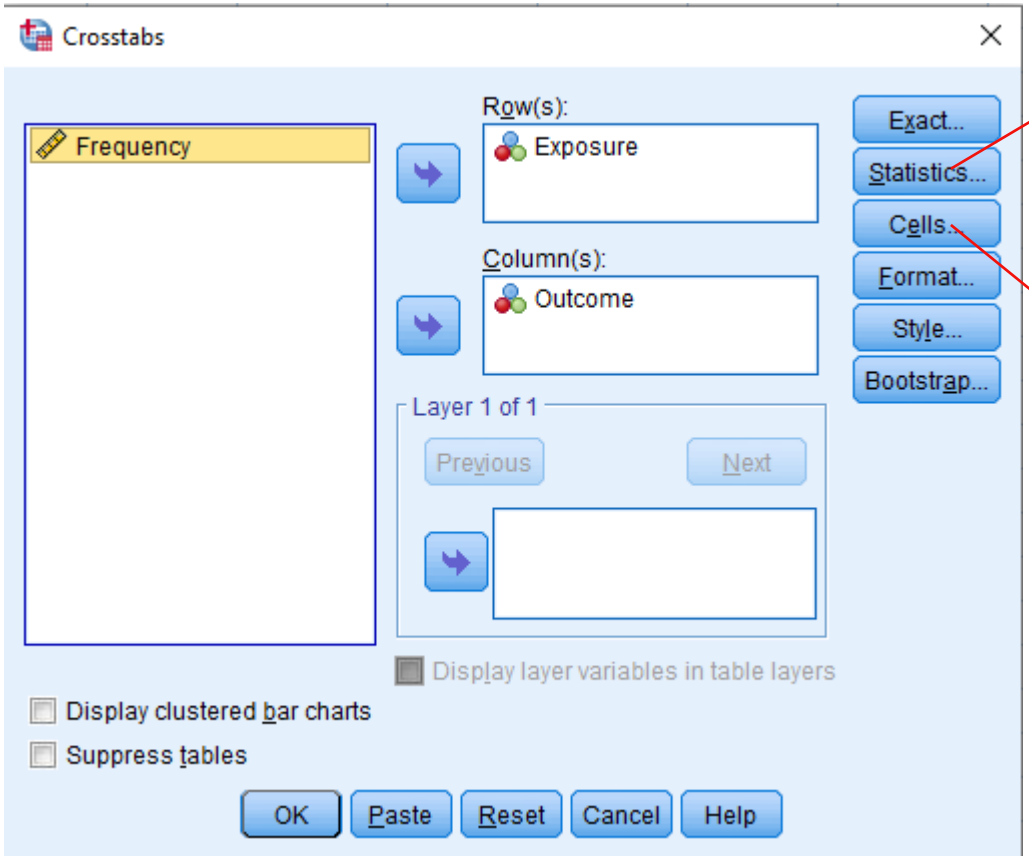
<Analyze>

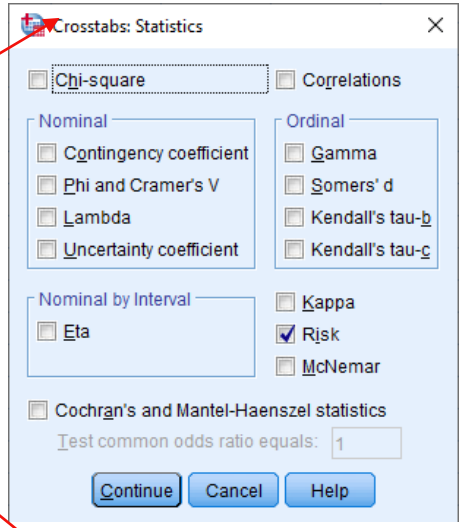
<Descriptive Statistics>

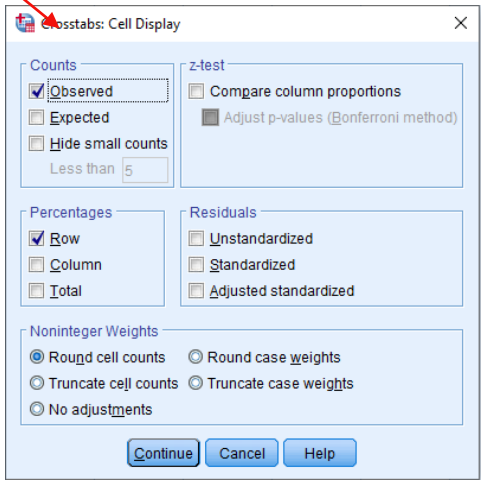
<Crosstabs>



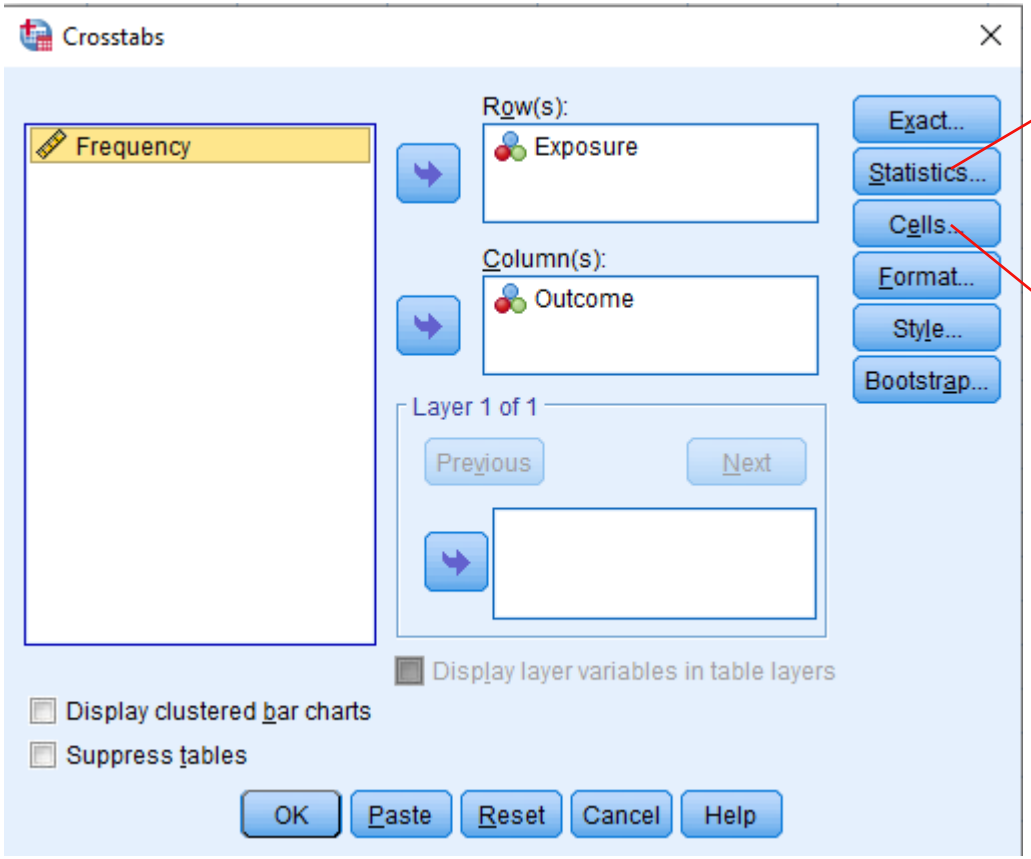
7.4 Relative Risk and Odds Ratio in SPSS

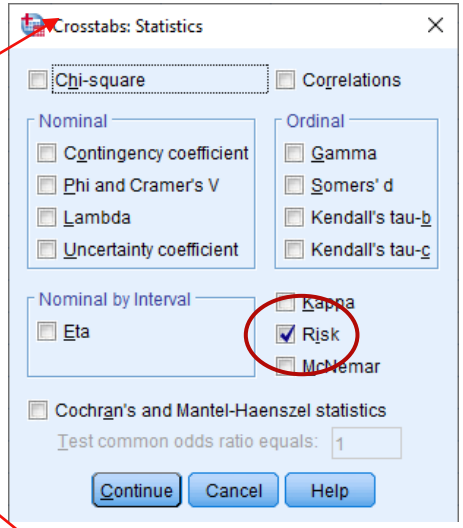


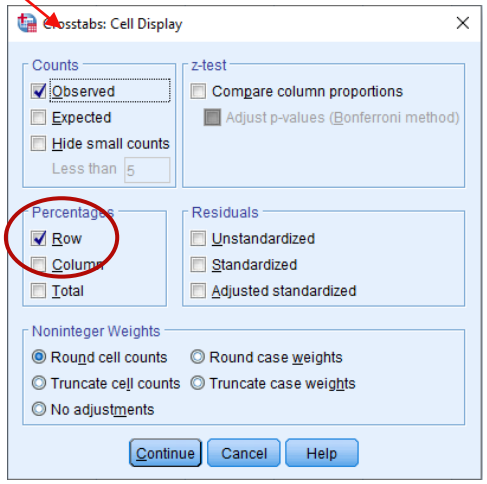




7.4 Relative Risk and Odds Ratio in SPSS







7.4 Relative Risk and Odds Ratio in SPSS

SPSS Output

Exposure * Outcome Crosstabulation

			Outcome		
			Diseased	Healthy	Total
Exposure	Exposed	Count	7	34	41
		% within Exposure	17.1%	82.9%	100.0%
	Unexposed	Count	20	17	37
		% within Exposure	54.1%	45.9%	100.0%
Total		Count	27	51	78
		% within Exposure	34.6%	65.4%	100.0%

The first table shows us the summary of the 2 x 2 layout of exposed/unexposed and diseased/healthy categories. We also see the % within Exposure ie. 17.1% is found by taking 7/41 and multiplying by 100.

Risk Estimate

		95% Confidence Interval	
	Value	Lower	Upper
Odds Ratio for Exposure (Exposed / Unexposed)	.175	.062	.495
For cohort Outcome = Diseased	.316	.151	.660
For cohort Outcome = Healthy	1.805	1.239	2.629
N of Valid Cases	78		

The second table shows our RR value, which in this case is “For cohort Outcome = Diseased”. RR = 0.316, which is what we calculated!

We also get the 95% CI in this table.

7.4 Relative Risk and Odds Ratio in SPSS

Odds ratio example (cardiovascular disease)…recall:

Smoking status	Disease Status		Total
	Case	Control	
Smoker	63	22	**85**
Non-smoker	20	97	**117**
Total	**83**	**119**	**202**

Note: setting up an OR in SPSS is the same as a RR…

7.4 Relative Risk and Odds Ratio in SPSS

Odds ratio example (cardiovascular disease):

ORexample.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities

6 : Frequency

	Exposure	Outcome	Frequency
1	Case	Smoker	63.00
2	Control	Smoker	22.00
3	Case	Non-smoker	20.00
4	Control	Non-smoker	97.00

Don't forget to weight cases according to the Frequency column.

7.4 Relative Risk and Odds Ratio in SPSS

Run the crosstabs and get the output:

Exposure * Outcome Crosstabulation

			Outcome		
			Smoker	Non-smoker	Total
Exposure	Case	Count	63	20	83
		% within Exposure	75.9%	24.1%	100.0%
	Control	Count	22	97	119
		% within Exposure	18.5%	81.5%	100.0%
Total		Count	85	117	202
		% within Exposure	42.1%	57.9%	100.0%

Risk Estimate

		95% Confidence Interval	
	Value	Lower	Upper
Odds Ratio for Exposure (Case / Control)	13.889	7.011	27.512
For cohort Outcome = Smoker	4.106	2.762	6.102
For cohort Outcome = Non-smoker	.296	.200	.437
N of Valid Cases	202		

And so the OR = 13.9, which is what we calculated.

Again, we also have the 95% CI in this table, which matches what we calculated.