The principals of statistical analysis

(Statistikk i helsefagleg forsking)

HELSTA Fall 2014

Øystein Haaland

Postdoc, Research Group for Genetic Epidemiology

Department of Global Public Health and Primary Care

University of Bergen

Chi-square test (χ^2-test)

- To test the association between two variables on a nominal level.
- Observed number (O) for each combination of values of the two variables is listed in a table.
- Expected number (E) for each cell is calculated.
- $\chi^2 = \Sigma(O - E)^2/E$ has a chi-square distribution with $(r-1)\cdot(s-1)$ degrees of freedom, where r and s are the number of catgories for each of the two variables

Chi-square test - example

Pain during labor – is there a difference in the perceived level of pain between women giving birth for the first time and women who have given birth before?

Level of pain

	Weak or moderate pain	Strong or unbearable pain	
First time mothers	**A** E = (A+B)(A+C)/n	**B** E = (A+B)(B+D)/n	A+B
Given birth before	**C** E = (A+C)(C+D)/n	**D** E = (B+D)(C+D)/n	C+D
	A+C	B+D	n = A+B+C+D

Chi-square test - example

Pain during labor – is there a difference in the perceived level of pain between women giving birth for the first time and women who have given birth before?

Level of pain

	Weak or moderate pain	Strong or unbearable pain	
First time mothers	**33** 45%	**40** 55%	73 100%
Given birth before	**46** 62%	**28** 38%	74 100%
	79	68	147

Relative risk (RR) = proportion of women with strong pain among first time mothers - divided by the proportion of women with strong pain among women having given birth before

RR = 55% / 38% = 1.45

Chi-square test - example

Level of pain

	Weak or moderate pain	Strong or unbearable pain	
First time mothers	**33** E = 39.2	**40** E = 33.8	73
Given birth before	**46** E = 39.8	**28** E = 34.2	74
	79	68	147

$\chi^2 = \Sigma(O - E)^2/E = (33-39)^2/39.2 + (40-33.8)^2/33.8 + ... = 4.25$

One degree of freedom: p-value between 0.05 and 0.025 (table A5).

Table A5 Percentage points of the χ^2 distribution.

	P-value							
d.f.	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	10.22	13.36	15.51	17.53	20.09	21.96	26.13
9	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	10.34	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	69.33	77.58	85.53	90.53	95.02	100.43	104.22	112.32
80	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	89.33	98.65	107.57	113.15	118.14	124.12	128.30	137.21
100	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

røyk * hjertesykdom Crosstabulation

			hjertesykdom		
			Ja	Nei	Total
røyk	Ja	Count	205	13958	14163
		% within røyk	1,4%	98,6%	100,0%
	Nei	Count	58	8096	8154
		% within røyk	,7%	99,3%	100,0%
Total		Count	263	22054	22317
		% within røyk	1,2%	98,8%	100,0%

H0: No difference in risk of heart disease between smokers and non-smokers
H1: Difference in risk

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	24,078[b]	1	,000		
Continuity Correction[a]	23,450	1	,000		
Fisher's Exact Test				,000	,000
N of Valid Cases	22317				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 96,09.

H0 rejected because p<0.05

Risk Estimate

	Value	95% Confidence Interval		
		Lower	Upper	
Odds Ratio for røyk (Nei / Ja)	2,050	1,530	2,747	OR
For cohort hjertesykdom = Ja	2,035	1,522	2,720	RR
N of Valid Cases	22317			

H0 rejected because 95% confidence interval did not contain 1

kjona * hjertesykdom Crosstabulation

			hjertesykdom		
			Ja	Nei	Total
kjona	Mann	Count	169	10094	10263
		% within kjona	1,6%	98,4%	100,0%
	Kvinne	Count	94	11960	12054
		% within kjona	,8%	99,2%	100,0%
Total		Count	263	22054	22317
		% within kjona	1,2%	98,8%	100,0%

H0: No difference in risk of heart disease between males and females

H1: Difference in risk

Risk Estimate

	Value	95% Confidence Interval		
		Lower	Upper	
Odds Ratio for kjona (Mann / Kvinne)	2,130	1,653	2,745	OR
For cohort hjertesykdom = Ja	2,112	1,643	2,714	RR
N of Valid Cases	22317			

H0 rejected?

kjona * hjertesykdom Crosstabulation

			hjertesykdom		
			Ja	Nei	Total
kjona	Mann	Count	169	10094	10263
		% within kjona	1,6%	98,4%	100,0%
	Kvinne	Count	94	11960	12054
		% within kjona	,8%	99,2%	100,0%
Total		Count	263	22054	22317
		% within kjona	1,2%	98,8%	100,0%

H0: No difference in risk of heart disease between males and females

H1: Difference in risk

Risk Estimate

	Value	95% Confidence Interval		
		Lower	Upper	
Odds Ratio for kjona (Mann / Kvinne)	2,130	1,653	2,745	OR
For cohort hjertesykdom = Ja	2,112	1,643	2,714	RR
N of Valid Cases	22317			

H0 rejected? Yes, because 1 is not contained in the 95% confidence interval(s)

ald1997 * hjertesykdom Crosstabulation

			hjertesykdom		
			Ja	Nei	Total
ald1997	40	Count	26	3636	3662
		% within ald1997	,7%	99,3%	100,0%
	41	Count	36	3692	3728
		% within ald1997	1,0%	99,0%	100,0%
	42	Count	38	3703	3741
		% within ald1997	1,0%	99,0%	100,0%
	43	Count	52	3692	3744
		% within ald1997	1,4%	98,6%	100,0%
	44	Count	45	3661	3706
		% within ald1997	1,2%	98,8%	100,0%
	46	Count	35	1753	1788
		% within ald1997	2,0%	98,0%	100,0%
	47	Count	31	1917	1948
		% within ald1997	1,6%	98,4%	100,0%
Total		Count	263	22054	22317
		% within ald1997	1,2%	98,8%	100,0%

H0: Age does not matter
H1: Age matters

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22,835[a]	6	,001
N of Valid Cases	22317		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 21,07.

H0 rejected?

ald1997 * hjertesykdom Crosstabulation

			hjertesykdom		
			Ja	Nei	Total
ald1997	40	Count	26	3636	3662
		% within ald1997	,7%	99,3%	100,0%
	41	Count	36	3692	3728
		% within ald1997	1,0%	99,0%	100,0%
	42	Count	38	3703	3741
		% within ald1997	1,0%	99,0%	100,0%
	43	Count	52	3692	3744
		% within ald1997	1,4%	98,6%	100,0%
	44	Count	45	3661	3706
		% within ald1997	1,2%	98,8%	100,0%
	46	Count	35	1753	1788
		% within ald1997	2,0%	98,0%	100,0%
	47	Count	31	1917	1948
		% within ald1997	1,6%	98,4%	100,0%
Total		Count	263	22054	22317
		% within ald1997	1,2%	98,8%	100,0%

H0: Age does not matter
H1: Age matters

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22,835[a]	6	,001
N of Valid Cases	22317		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 21,07.

H0 rejected? Yes, because $p<0.05$

alccat2 * hjertesykdom Crosstabulation

			hjertesykdom		
			Ja	Nei	Total
alccat2	1	Count	65	4968	5033
		% within alccat2	1,3%	98,7%	100,0%
	2	Count	22	1932	1954
		% within alccat2	1,1%	98,9%	100,0%
	3	Count	34	3110	3144
		% within alccat2	1,1%	98,9%	100,0%
	4	Count	24	2853	2877
		% within alccat2	,8%	99,2%	100,0%
	5	Count	39	3159	3198
		% within alccat2	1,2%	98,8%	100,0%
	6	Count	35	2774	2809
		% within alccat2	1,2%	98,8%	100,0%
	7	Count	13	619	632
		% within alccat2	2,1%	97,9%	100,0%
	8	Count	1	61	62
		% within alccat2	1,6%	98,4%	100,0%
Total		Count	233	19476	19709
		% within alccat2	1,2%	98,8%	100,0%

H0: Alcohol use does not matter

H1: Alcohol use matters

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8,197[a]	7	,316
N of Valid Cases	19709		

a. 1 cells (6,3%) have expected count less than 5. The minimum expected count is ,73.

H0 rejected?

alccat2 * hjertesykdom Crosstabulation

			hjertesykdom		
			Ja	Nei	Total
alccat2	1	Count	65	4968	5033
		% within alccat2	1,3%	98,7%	100,0%
	2	Count	22	1932	1954
		% within alccat2	1,1%	98,9%	100,0%
	3	Count	34	3110	3144
		% within alccat2	1,1%	98,9%	100,0%
	4	Count	24	2853	2877
		% within alccat2	,8%	99,2%	100,0%
	5	Count	39	3159	3198
		% within alccat2	1,2%	98,8%	100,0%
	6	Count	35	2774	2809
		% within alccat2	1,2%	98,8%	100,0%
	7	Count	13	619	632
		% within alccat2	2,1%	97,9%	100,0%
	8	Count	1	61	62
		% within alccat2	1,6%	98,4%	100,0%
Total		Count	233	19476	19709
		% within alccat2	1,2%	98,8%	100,0%

H0: Alcohol use does not matter

H1: Alcohol use matters

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8,197[a]	7	,316
N of Valid Cases	19709		

a. 1 cells (6,3%) have expected count less than 5. The minimum expected count is ,73.

H0 rejected? No, because p>0.05

Logistic regression

$$\textbf{logit}\,(p_i) = \ln\left(p_i/(1-p_i)\right) = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i}$$

$$p_i/(1-p_i) = \exp(\beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i})$$

Two outcomes: Event or non-event (1 or 0, alive or dead)
x_1 ... x_k are k independent variables (age, sex,…)
p_i is the probability of the event happening for person no. i,
β_0, β_1, ..., β_k are regression coefficients

$$\exp(\beta_k) = \text{OR for outcomes between } x_k \text{ og } (x_k + 1)$$

Logistic regression

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	røyk	1,037	,175	35,153	1	,000	2,820
	kjønn	,843	,144	34,256	1	,000	2,324
	alder	,135	,030	20,307	1	,000	1,144
	alccat2	-,090	,036	6,146	1	,013	,914
	Constant	-13,032	1,347	93,613	1	,000	,000

a. Variable(s) entered on step 1: røyk, kjønn, alder, alccat2.

All p<0.05 OR

alccat2 * røyk Crosstabulation

			røyk		
			1,00	2,00	Total
alccat2	1	Count	2458	2557	5015
		% within alccat2	49,0%	51,0%	100,0%
	2	Count	801	1150	1951
		% within alccat2	41,1%	58,9%	100,0%
	3	Count	1163	1976	3139
		% within alccat2	37,1%	62,9%	100,0%
	4	Count	982	1887	2869
		% within alccat2	34,2%	65,8%	100,0%
	5	Count	933	2260	3193
		% within alccat2	29,2%	70,8%	100,0%
	6	Count	729	2075	2804
		% within alccat2	26,0%	74,0%	100,0%
	7	Count	145	485	630
		% within alccat2	23,0%	77,0%	100,0%
	8	Count	13	48	61
		% within alccat2	21,3%	78,7%	100,0%
Total		Count	7224	12438	19662
		% within alccat2	36,7%	63,3%	100,0%

Nei Ja

alccat2 * kjona Crosstabulation

			kjona		
			Menn	Kvinner	Total
alccat2	1	Count	1659	3374	5033
		% within alccat2	33,0%	67,0%	100,0%
	2	Count	688	1266	1954
		% within alccat2	35,2%	64,8%	100,0%
	3	Count	1263	1881	3144
		% within alccat2	40,2%	59,8%	100,0%
	4	Count	1351	1526	2877
		% within alccat2	47,0%	53,0%	100,0%
	5	Count	1729	1469	3198
		% within alccat2	54,1%	45,9%	100,0%
	6	Count	1993	816	2809
		% within alccat2	71,0%	29,0%	100,0%
	7	Count	533	99	632
		% within alccat2	84,3%	15,7%	100,0%
	8	Count	60	2	62
		% within alccat2	96,8%	3,2%	100,0%
Total		Count	9276	10433	19709
		% within alccat2	47,1%	52,9%	100,0%

Assocation between education and risk of multiple sclerosis

Education short * Status Crosstabulation

Count

		Status		
		controls	cases	Total
Education short	Grunnskole (up to 10 yrs)	201	153	354
	Videregående (11-13 yrs)	601	385	986
	Høyskole/ Universitet (>14 yrs)	890	402	1292
Total		1692	940	2632

Assocation between education and risk of multiple sclerosis

Education short * Status Crosstabulation

			Status		
			controls	cases	Total
Education short	Grunnskole (up to 10 yrs)	Count	201	153	354
		% within Status	11,9%	16,3%	13,4%
	Videregående (11-13 yrs)	Count	601	385	986
		% within Status	35,5%	41,0%	37,5%
	Høyskole/ Universitet (>14 yrs)	Count	890	402	1292
		% within Status	52,6%	42,8%	49,1%
Total		Count	1692	940	2632
		% within Status	100,0%	100,0%	100,0%

Assocation between education and risk of multiple sclerosis

Education short * Status Crosstabulation

			Status		
			controls	cases	Total
Education short	Grunnskole (up to 10 yrs)	Count	201	153	354
		Expected Count	227,6	126,4	354,0
	Videregående (11-13 yrs)	Count	601	385	986
		Expected Count	633,9	352,1	986,0
	Høyskole/ Universitet (>14 yrs)	Count	890	402	1292
		Expected Count	830,6	461,4	1292,0
Total		Count	1692	940	2632
		Expected Count	1692,0	940,0	2632,0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	25,362[a]	2	,000
Likelihood Ratio	25,328	2	,000
Linear-by-Linear Association	24,549	1	,000
N of Valid Cases	2632		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 126,43.

Educational level and the risk of MS – adjusting for confounding factors

EnvIMS – Norwegian data

	Model 1 P* OR [95%CI]
Education	< 0.0001
Compulsory	1
Intermediate	**0.82** [0.64 – 1.06]
Tertiery	**0.56** [0.43 – 0.72]

P* overall p-value

Educational level and the risk of MS – adjusting for confounding factors

EnvIMS – Norwegian data

	Model 1 P* OR [95%CI]
Education	< 0.0001
Compulsory	1
Intermediate	**0.82** [0.64 – 1.06]
Tertiery	**0.56** [0.43 – 0.72]
Smoking	
No	
Yes	
Infectious mononucleosis	
No	
Yes	
Sun exposure	
Max. vs min. exposure	
Fatty fish consumption	
Max. vs min. exposure	
Overweight	
Max. vs min. exposure	

P* overall p-value

Educational level and the risk of MS – adjusting for confounding factors

EnvIMS – Norwegian data

	Model 1	Model 2
	P*	P*
	OR [95%CI]	OR [95%CI]
Education	< 0.0001	0.0013
Compulsory	1	1
Intermediate	**0.82** [0.64 – 1.06]	**0.94** [0.72 – 1.23]
Tertiery	**0.56** [0.43 – 0.72]	**0.69** [0.53 – 0.90]
Smoking		< 0.0001
No		1
Yes		**2.13** [1.77 – 2.56]
Infectious mononucleosis		
No		
Yes		
Sun exposure		
Max. vs min. exposure		
Fatty fish consumption		
Max. vs min. exposure		
Overweight		
Max. vs min. exposure		

P* overall p-value

Educational level and the risk of MS – adjusting for confounding factors

EnvIMS – Norwegian data

	Model 1	Model 2	Model 3
	P*	P*	P*
	OR [95%CI]	OR [95%CI]	OR [95%CI]
Education	< 0.0001	0.0013	< 0.0001
Compulsory	1	1	1
Intermediate	**0.82** [0.64 – 1.06]	**0.94** [0.72 – 1.23]	**0.90** [0.68 – 1.18]
Tertiery	**0.56** [0.43 – 0.72]	**0.69** [0.53 – 0.90]	**0.64** [0.49 – 0.82]
Smoking		< 0.0001	< 0.0001
No		1	1
Yes		**2.13** [1.77 – 2.56]	2.13 [1.76 – 2.57]
Infectious mononucleosis			< 0.0001
No			1
Yes			**2.28** [1.77 – 2.96]
Sun exposure			
Max. vs min. exposure			
Fatty fish consumption			
Max. vs min. exposure			
Overweight			
Max. vs min. exposure			

P* overall p-value

Educational level and the risk of MS – adjusting for confounding factors

EnvIMS – Norwegian data

	Model 1	Model 2	Model 3	Model 4
	P*	P*	P*	P*
	OR [95%CI]	OR [95%CI]	OR [95%CI]	OR [95%CI]
Education	< 0.0001	0.0013	< 0.0001	< 0.0001
Compulsory	1	1	1	1
Intermediate	**0.82** [0.64 – 1.06]	**0.94** [0.72 – 1.23]	**0.90** [0.68 – 1.18]	**0.88** [0.66 – 1.17]
Tertiery	**0.56** [0.43 – 0.72]	**0.69** [0.53 – 0.90]	**0.64** [0.49 – 0.82]	**0.62** [0.47 – 0.83]
Smoking		< 0.0001	< 0.0001	< 0.0001
No		1	1	1
Yes		**2.13** [1.77 – 2.56]	2.13 [1.76 – 2.57]	**2.10** [1.73 – 2.54]
Infectious mononucleosis			< 0.0001	< 0.0001
No			1	1
Yes			**2.28** [1.77 – 2.96]	**2.30** [1.77 – 2.98]
Sun exposure				0.0013
Max. vs min. exposure				**1.89** [1.28 – 2.76]
Fatty fish consumption				
Max. vs min. exposure				
Overweight				
Max. vs min. exposure				

P* overall p-value

Educational level and the risk of MS – adjusting for confounding factors

EnvIMS – Norwegian data

	Model 1	Model 2	Model 3	Model 4	Model 5
	P*	P*	P*	P*	P*
	OR [95%CI]	OR [95%CI]	OR [95%CI]	OR [95%CI]	OR [95%CI]
Education	< 0.0001	0.0013	< 0.0001	< 0.0001	< 0.0001
Compulsory	1	1	1	1	1
Intermediate	**0.82** [0.64 – 1.06]	**0.94** [0.72 – 1.23]	**0.90** [0.68 – 1.18]	**0.88** [0.66 – 1.17]	**0.85** [0.61 – 1.19]
Tertiery	**0.56** [0.43 – 0.72]	**0.69** [0.53 – 0.90]	**0.64** [0.49 – 0.82]	**0.62** [0.47 – 0.83]	**0.62** [0.44 – 0.87]
Smoking		< 0.0001	< 0.0001	< 0.0001	< 0.0001
No		1	1	1	1
Yes		**2.13** [1.77 – 2.56]	2.13 [1.76 – 2.57]	**2.10** [1.73 – 2.54]	**2.01** [1.63 – 2.49]
Infectious mononucleosis			< 0.0001	< 0.0001	< 0.0001
No			1	1	1
Yes			**2.28** [1.77 – 2.96]	**2.30** [1.77 – 2.98]	**2.34** [1.77 – 3.10]
Sun exposure				0.0013	0.0061
Max. vs min. exposure				**1.89** [1.28 – 2.76]	**1.83** [1.19 – 2.82]
Fatty fish consumption					0.060
Max. vs min. exposure					**2.05** [0.97 – 4.34]
Overweight					
Max. vs min. exposure					

P* overall p-value

Educational level and the risk of MS – adjusting for confounding factors

EnvIMS – Norwegian data

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	P*	P*	P*	P*	P*	p*
	OR [95%CI]	OR [95%CI]	OR [95%CI]	OR [95%CI]	OR [95%CI]	OR [95%CI]
Education	< 0.0001	0.0013	< 0.0001	< 0.0001	< 0.0001	0.0032
Compulsory	1	1	1	1	1	1
Intermediate	**0.82** [0.64 – 1.06]	**0.94** [0.72 – 1.23]	**0.90** [0.68 – 1.18]	**0.88** [0.66 – 1.17]	**0.85** [0.61 – 1.19]	**0.83** [0.59 – 1.16]
Tertiery	**0.56** [0.43 – 0.72]	**0.69** [0.53 – 0.90]	**0.64** [0.49 – 0.82]	**0.62** [0.47 – 0.83]	**0.62** [0.44 – 0.87]	**0.61** [0.44 – 0.86]
Smoking		< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
No		1	1	1	1	1
Yes		**2.13** [1.77 – 2.56]	2.13 [1.76 – 2.57]	**2.10** [1.73 – 2.54]	**2.01** [1.63 – 2.49]	**1.92** [1.55 – 2.39]
Infectious mononucleosis			< 0.0001	< 0.0001	< 0.0001	< 0.0001
No			1	1	1	
Yes			**2.28** [1.77 – 2.96]	**2.30** [1.77 – 2.98]	**2.34** [1.77 – 3.10]	**2.29** [1.72 – 3.05]
Sun exposure				0.0013	0.0061	0.029
Max. vs min. exposure				**1.89** [1.28 – 2.76]	**1.83** [1.19 – 2.82]	**1.63** [1.05 – 2.55]
Fatty fish consumption					0.060	0.034
Max. vs min. exposure					**2.05** [0.97 – 4.34]	**2.28** [1.07 – 4.88]
Overweight						0.029
Max. vs min. exposure						**1.88** [1.07 – 3.32]

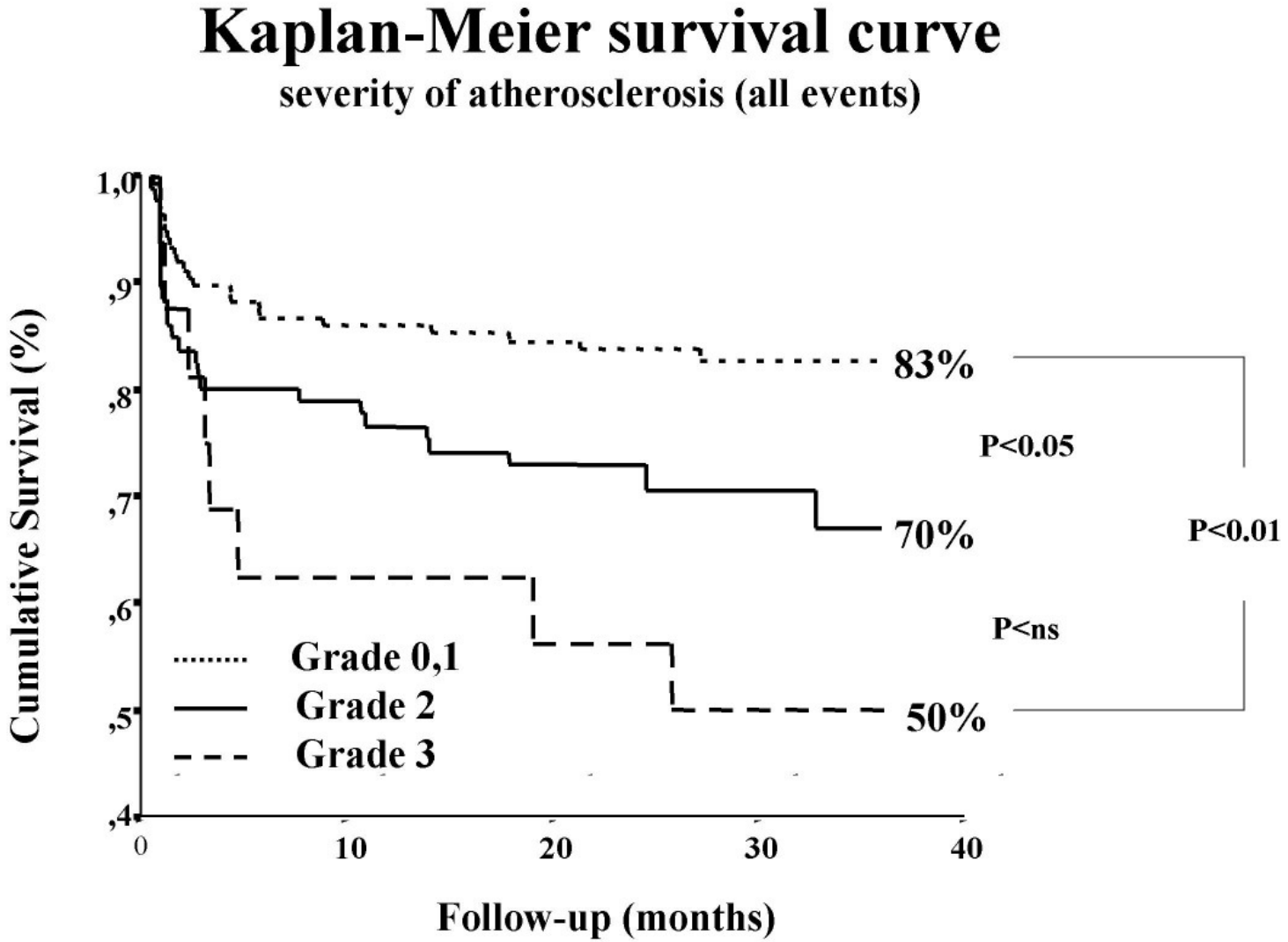
P* overall p-value

Kaplan Meier survival curves

Shows percentage of study individuals over time who are still alive (have not reached a defined endpoint/event).

Cox regression survival analysis

Regression analysis with time to event as the dependent variable. Assumes constant risk over time between levels of exposure variables (independent variables) – “proportional hazard”



A Kaplan-Meier curve showing the association between the severity of the transesophageally detected aortic plaques and the long term survival including all events. It can be clearly seen that more severe the atherosclerosis on the descending aorta was, higher the probability of future cardiovascular events.

Varga et al. Cardiovascular Ultrasound 2004 ***2****:21 doi:10.1186/1476-7120-2-21*

Kaplan Meier survival curves

Shows percentage of study individuals still alive (not reached a defined endpoint) over time.

Cox regression survival analysis

$$\lambda(t|X) = \lambda_0(t)\exp(\beta_1 X_1 + \cdots + \beta_p X_p)$$

Regression analysis with time to event as the dependent variable. Assumes constant risk over time between levels of exposure variables (independent variables) – “proportional hazard”

Qual Life Res
DOI 10.1007/s11136-013-0453-2

Poor self-rated health associated with an increased risk of subsequent development of lung cancer

Hilde Kristin Refvik Riise · Trond Riise · Gerd Karin Natvig · Anne Kjersti Daltveit

Accepted: 30 May 2013

Abstract
Purpose Self-rated health has shown to be a strong predictor of mortality and some major chronic diseases. The purpose of this study was to investigate whether poor self-rated health also was related to an increased risk of subsequent development of cancer.
Methods Information on self-rated health, life-style factors, and other health-related risk factors was ascertained in a cohort of 25,532 persons participating in the Hordaland Health Study in 1997–1999. Information on development findings need to be repeated before elaborate interpretations can be made.

Keywords Self-rated health · Follow-up study · HUSK · Lung cancer · Health behavior · Life-style

Introduction

Environmental, genetical, and lifestyle factors are in

Table 3 Distribution of significant covariates (crude) for lung cancer among 24,487 respondents in the HUSK study 1997–1999

	N (%)	Lung cancer (%)	Crude[a] HR (CI 95 %)	*p*-Trend/*p*-overall	Adjusted[b] HR (CI 95 %)	*p*-Trend/*p*-overall
Self-rated health						
Very good	4,651 (19.0)	7 (0.2)	1 (ref.)	0.001/0.007	1 (ref.)	0.038/0.20
Good	15,727 (64.2)	45 (0.3)	1.50 (0.67–3.34)		1.57 (0.61–4.02)	
Not so good	3,594 (14.7)	23 (0.6)	2.54 (1.07–6.03)		2.13 (0.77–5.89)	
Poor	276 (1.1)	4 (1.5)	6.13 (1.77–21.19)		3.88 (0.97–15.50)	
Light physical activity						
3 h or more per week	10,660 (43.5)	30 (0.5)	1 (ref.)	0.002/0.005	1 (ref.)	0.077/0.16
Between 1 and 2 h per week	8,664 (35.4)	23 (0.3)	1.16 (0.67–2.00)		1.09 (0.61–1.93)	
<1 h per week	3,218 (13.1)	11 (0.3)	1.59 (0.79–3.20)		1.13 (0.52–2.43)	
No such activity	1,106 (4.5)	11 (1.0)	3.41 (1.71–6.81)		2.36 (1.11–5.01)	
Pack years						
None smokers	9,266 (37.8)	6 (0.1)	1 (ref.)	<0.001/<0.001	1 (ref.)	<0.001/<0.001
0–5 pack years	3,889 (15.8)	5 (0.1)	2.60 (0.79–8.56)		3.21 (0.86–12.01)	
5–15 pack years	6,027 (24.6)	17 (0.3)	5.83 (2.28–14.912)		6.03 (1.98–18.32)	
15–30 pack years	3,999 (16.3)	33 (0.8)	16.57 (6.87–39.94)		15.85 (5.49–45.76)	
>30 pack years	707 (2.9)	15 (2.1)	25.55 (9.65–67.65)		24.98 (7.87–79.31)	
Alcohol per 2 weeks						
None drinker/teetotaler	5,966 (24.4)	19 (0.3)	1 (ref.)	0.044/0.13	1 (ref.)	0.040/0.58
Low	13,861 (56.6)	41 (0.3)	1.42 (0.82–2.45)		1.33 (0.74–2.40)	
Medium and high	2,941 (12.0)	13 (0.4)	2.11 (1.02–4.40)		1.43 (0.65–3.12)	
BMI						
20–25	11,125 (45.4)	36 (0.3)	1 (ref.)	–/<0.001	1 (ref.)	–/<0.001
0–19.99	1,027 (4.2)	11 (1.1)	3.68 (1.85–7.31)		3.14 (1.55–6.36)	
>26	12,289 (50.2)	31 (0.3)	0.61 (0.38–0.99)		0.63 (0.37–1.07)	

[a] Adjusted for age cohort and gender

[b] Adjusted for age cohort, gender, light physical activity, pack years, BMI, alcohol, and SRH

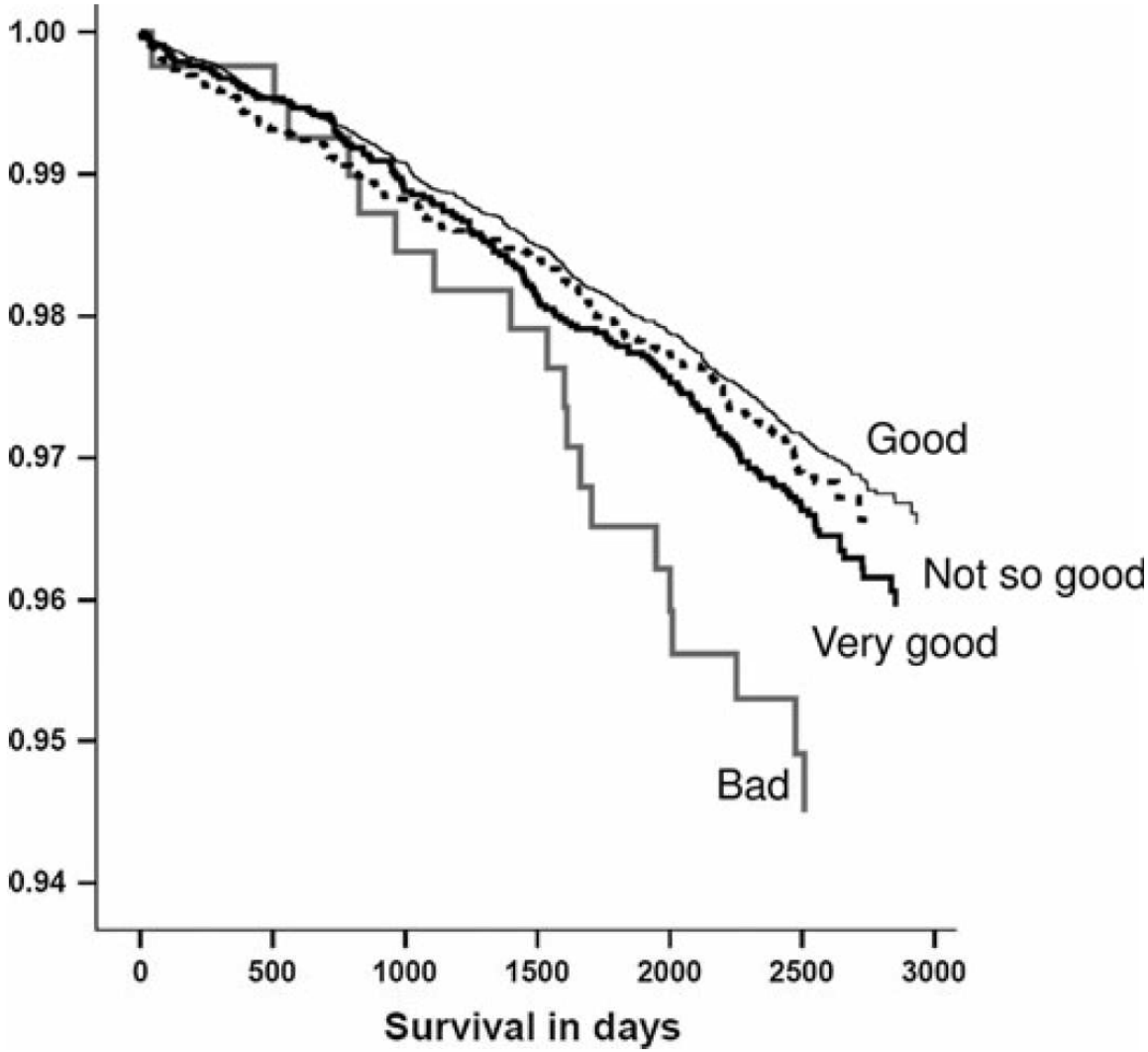


Fig. 1 Cumulative risk of overall cancer according to self-rated health. Survival curves from Cox regression analysis for the different categories of self-rated health according to overall cancer among 24,487 respondents in the HUSK study 1997–1999. Adjusted for age cohort, gender, light physical activity, pack years, alcohol, and BMI

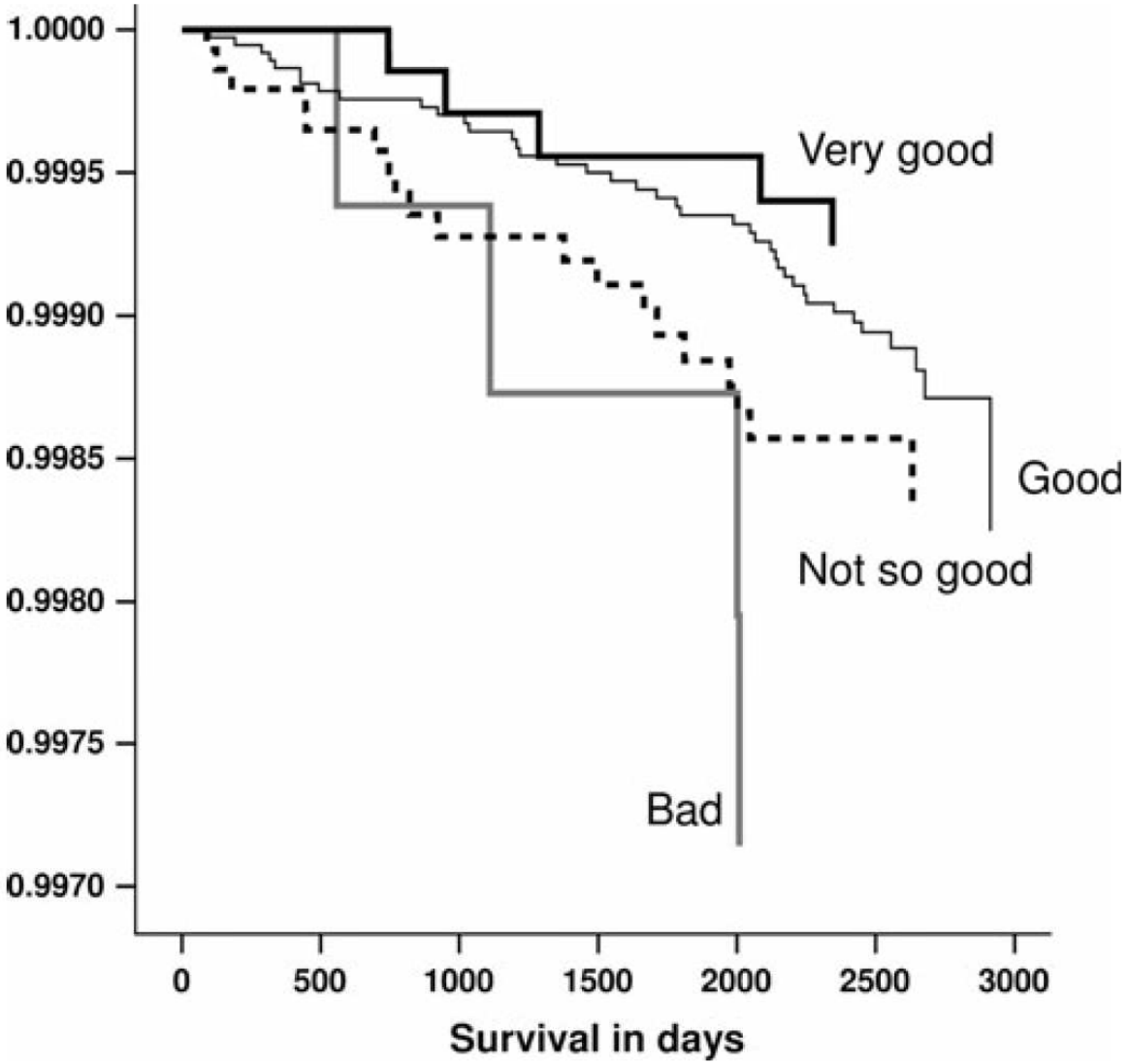


Fig. 2 Cumulative risk of lung cancer according to self-rated health. Survival curves from Cox regression analysis for the different categories of self-rated health according to lung cancer among 24,487 respondents in the HUSK study 1997–1999. Adjusted for age cohort, gender, light physical activity, pack years, alcohol, and BMI

Power and sample size calculations

- β = function of α, Δ and n, where
  - β is the statistical power of the test
  - α is the level of significance
  - Δ is the effect size (difference between the groups)
  - n is the number of cases in the study

- α is also called type I error
- 1- β is also called type II error

Sample size calculation - example

Sample sizes for testing difference in percentage of "active lesion free" patients.

$\alpha = 0.05, \beta = 0.80$

Percent of patients without lesions: In treatment group	In placebo group	Relative "risk" of no lesion	Number of patients needed in both groups
30	15	2.0	95
45	15	3.0	28
60	15	4.0	14

What statistical test to use (hypothesis testing)

Dependent variable (outcome variable)	**Research question**	**Model**
Ordinal/continuous	Comparing a mean and a fixed number (H_o: $\mu = \mu_o$)	One-sample t-test
(may be ranked)	Comparing the means of two groups (H_o: $\mu_1 = \mu_2$)	Two-sample t-test
	Comparing the mean in two paired groups (i.e. before and after treatment) $d = X_{after} - X_{before}$ (H_o: $\mu_d = 0$)	Paired t-test
	Comparing the means of 3 and more groups (H_o: $\mu_1 = \mu_2 = \mu_3$)	Analysis of variance (ANOVA)
	Comparing the means of groups according to two or more categorical variables (e.g. marital status and sex) (H_o: $\mu_1 = \mu_2$, $v_1 = v_2$)	Multiway ANOVA
	Association with another ordinal independent variable (dose-response) dep.var. = α+ β*(indep.var.) (H_o: $\beta = 0$)	Regression analysis, Correlation analysis
Nominal (categorical)	Comparing a proportion and a fixed number (H_o: $\pi = \pi_o$)	Binomial test
	Comparing proportions in subgroups (Crosstabs) (H_o: $\pi_1 = \pi_2$)	Chi-square test Fishers exact test
Two categories only	Association with several ordinal and nominal independent variables (e.g. dependent variable “dead/alive” – “disease(yes/no)”	Logistic regression
Time to an event	Between groups. Particularly useful for prospective design Censoring.	Life tables Kaplan-Meier curves
	Association with several ordinal and nominal independent variables	Cox-regression

Example of a statistical test I

- Problem: To what extent will a congenital heart malformation (non-cyanotic) influence the motor development of a child?

- A study is performed where 18 children with congenital heart malformation are followed for observation of when the children first were able to walk.
  The mean age in months was 14.1

Congenital heart malformation and motor development – test II

- From large studies of normal children it has been shown that the mean age of children at their first steps alone is 13 months with a standard deviation of 1.75 months.

- Based on the problem in question a null-hypothesis (H_0) is defined:
  - Children with congenital heart malformation has the same mean age when they learn to walk
  - H_0: $\mu = 13$ months

Congenital heart malformation and motor development – test III

- Assume a normal distribution of the mean and assume that the H_0 is true, then:

- Calculate the probability that a random sample of 18 children has a mean “as far away from” 13 as 14.1. This is the p-value.

- If the p-value is small (less than 0.05), this means that what we have observed is unlikely to observe. Self-contradiction. Reject the underlying assumption.

t-test

- When the standard error cannot be assumed known, we will use the standard error in the sample (SD) as an estimate of the standard deviation of the population (σ).
- Instead of a z-value, we will compute a t-value for calculating the p-value.
- $t = (\overline{X} - \mu_0)/(SD/\sqrt{n}))$
- P-value is found in a t-table using (n-1) degrees of freedom

What statistical test to use (hypothesis testing)

Dependent variable (outcome variable)	**Research question**	**Model**
Ordinal/continuous	Comparing a mean and a fixed number (H_o: $\mu = \mu_o$)	One-sample t-test
(may be ranked)	Comparing the means of two groups (H_o: $\mu_1 = \mu_2$)	Two-sample t-test
	Comparing the mean in two paired groups (i.e. before and after treatment) $d = X_{after} - X_{before}$ (H_o: $\mu_d = 0$)	Paired t-test
	Comparing the means of 3 and more groups (H_o: $\mu_1 = \mu_2 = \mu_3$)	Analysis of variance (ANOVA)
	Comparing the means of groups according to two or more categorical variables (e.g. marital status and sex) (H_o: $\mu_1 = \mu_2$, $v_1 = v_2$)	Multiway ANOVA
	Association with another ordinal independent variable (dose-response) dep.var. = α+ β*(indep.var.) (H_o: β = 0)	Regression analysis, Correlation analysis
Nominal (categorical)	Comparing a proportion and a fixed number (H_o: $\pi = \pi_o$)	Binomial test
	Comparing proportions in subgroups (Crosstabs) (H_o: $\pi_1 = \pi_2$)	Chi-square test Fishers exact test
Two categories only	Association with several ordinal and nominal independent variables (e.g. dependent variable “dead/alive” – “disease(yes/no)”	Logistic regression
Time to an event	Between groups. Particularly useful for prospective design Censoring.	Life tables Kaplan-Meier curves
	Association with several ordinal and nominal independent variables	Cox-regression

What statistical test to use (hypothesis testing)

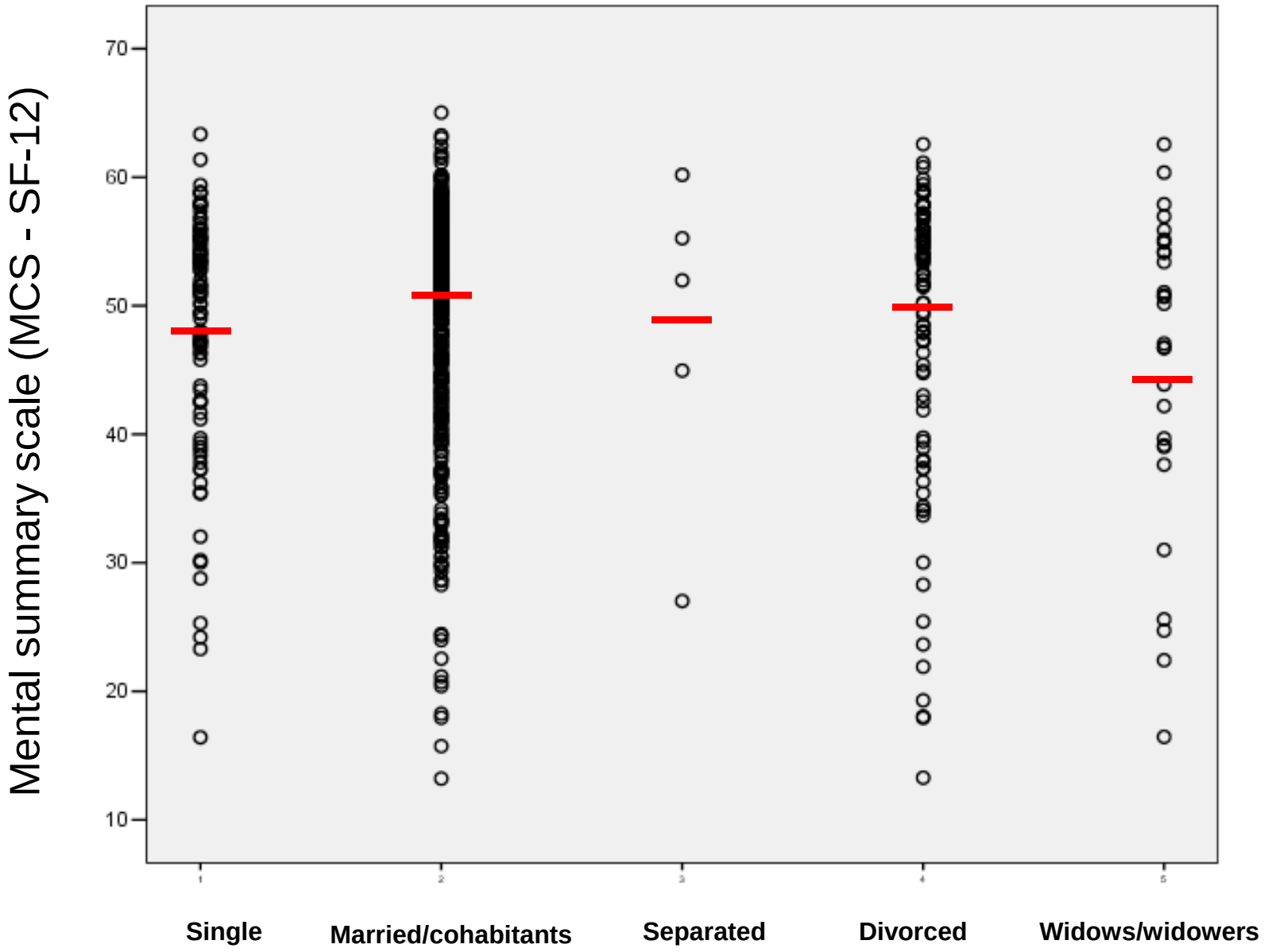
Dependent variable (outcome variable)	**Research question**	**Model**
Ordinal/continuous (may be ranked)	Comparing a mean and a fixed number (H_o: $\mu = \mu_o$)	One-sample t-test
	Comparing the means of two groups (H_o: $\mu_1 = \mu_2$)	Two-sample t-test
	Comparing the mean in two paired groups (i.e. before and after treatment) $d = X_{after} - X_{before}$ (H_o: $\mu_d = 0$)	Paired t-test
	Comparing the means of 3 and more groups (H_o: $\mu_1 = \mu_2 = \mu_3$)	Analysis of variance (ANOVA)
	Comparing the means of groups according to two or more categorical variables (e.g. marital status and sex) (H_o: $\mu_1 = \mu_2$, $v_1 = v_2$)	Multiway ANOVA
	Association with another ordinal independent variable (dose-response) dep.var. = α+ β*(indep.var.) (H_o: $\beta = 0$)	Regression analysis, Correlation analysis
Nominal (categorical)	Comparing a proportion and a fixed number (H_o: $\pi = \pi_o$)	Binomial test
	Comparing proportions in subgroups (Crosstabs) (H_o: $\pi_1 = \pi_2$)	Chi-square test Fishers exact test
Two categories only	Association with several ordinal and nominal independent variables (e.g. dependent variable “dead/alive” – “disease(yes/no)”	Logistic regression
Time to an event	Between groups. Particularly useful for prospective design Censoring.	Life tables Kaplan-Meier curves
	Association with several ordinal and nominal independent variables	Cox-regression

Paired t-test

- Paired samples *t*-tests are used when we have a sample of matched pairs of similar units, or one group of units that has been tested twice (a "repeated measures" *t*-test).
- $X_{diff} = X_{after} - X_{before}$
- We use a one sample t-test on X_{diff} with (n-1) d.f.
- $t = (\bar{X} - \mu_0)/(SD/\sqrt{n}))$

What statistical test to use (hypothesis testing)

Dependent variable (outcome variable)	**Research question**	**Model**
Ordinal/continuous (may be ranked)	Comparing a mean and a fixed number (H_o: $\mu = \mu_o$)	One-sample t-test
	Comparing the means of two groups (H_o: $\mu_1 = \mu_2$)	Two-sample t-test
	Comparing the mean in two paired groups (i.e. before and after treatment) $d = X_{after} - X_{before}$ (H_o: $\mu_d = 0$)	Paired t-test
	Comparing the means of 3 and more groups (H_o: $\mu_1 = \mu_2 = \mu_3$)	Analysis of variance (ANOVA)
	Comparing the means of groups according to two or more categorical variables (e.g. marital status and sex) (H_o: $\mu_1 = \mu_2$, $v_1 = v_2$)	Multiway ANOVA
	Association with another ordinal independent variable (dose-response) dep.var. = α+ β*(indep.var.) (H_o: $\beta = 0$)	Regression analysis, Correlation analysis
Nominal (categorical)	Comparing a proportion and a fixed number (H_o: $\pi = \pi_o$)	Binomial test
	Comparing proportions in subgroups (Crosstabs) (H_o: $\pi_1 = \pi_2$)	Chi-square test Fishers exact test
Two categories only	Association with several ordinal and nominal independent variables (e.g. dependent variable “dead/alive” – “disease(yes/no)”	Logistic regression
Time to an event	Between groups. Particularly useful for prospective design Censoring.	Life tables Kaplan-Meier curves
	Association with several ordinal and nominal independent variables	Cox-regression


Mental summary scale (MCS - SF-12)
70
60
50
40
30
20
10
Single
Married/cohabitants
Separated
Divorced
Widows/widowers

Analysis of variance (ANOVA)

Descriptive Statistics

Dependent Variable: **Mental health MCS**

sivilst	Mean	Std. Deviation	N
Single	48,5461	9,22855	100
Gift/samboer	50,8525	8,11733	640
Separert	47,8762	12,90328	5
Skilt	48,8690	10,95987	104
Enke/enkmann	45,3697	12,19819	28
Total	50,1622	8,87514	877

Tests of Between-Subjects Effects

Dependent Variable: **Mental health MCS**

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1409,270[a]	4	352,318	4,545	,001
Intercept	227055,384	1	227055,384	2929,248	,000
sivilst	1409,270	4	352,318	4,545	,001
Error	67591,520	872	77,513		
Total	2275752,944	877			
Corrected Total	69000,790	876			

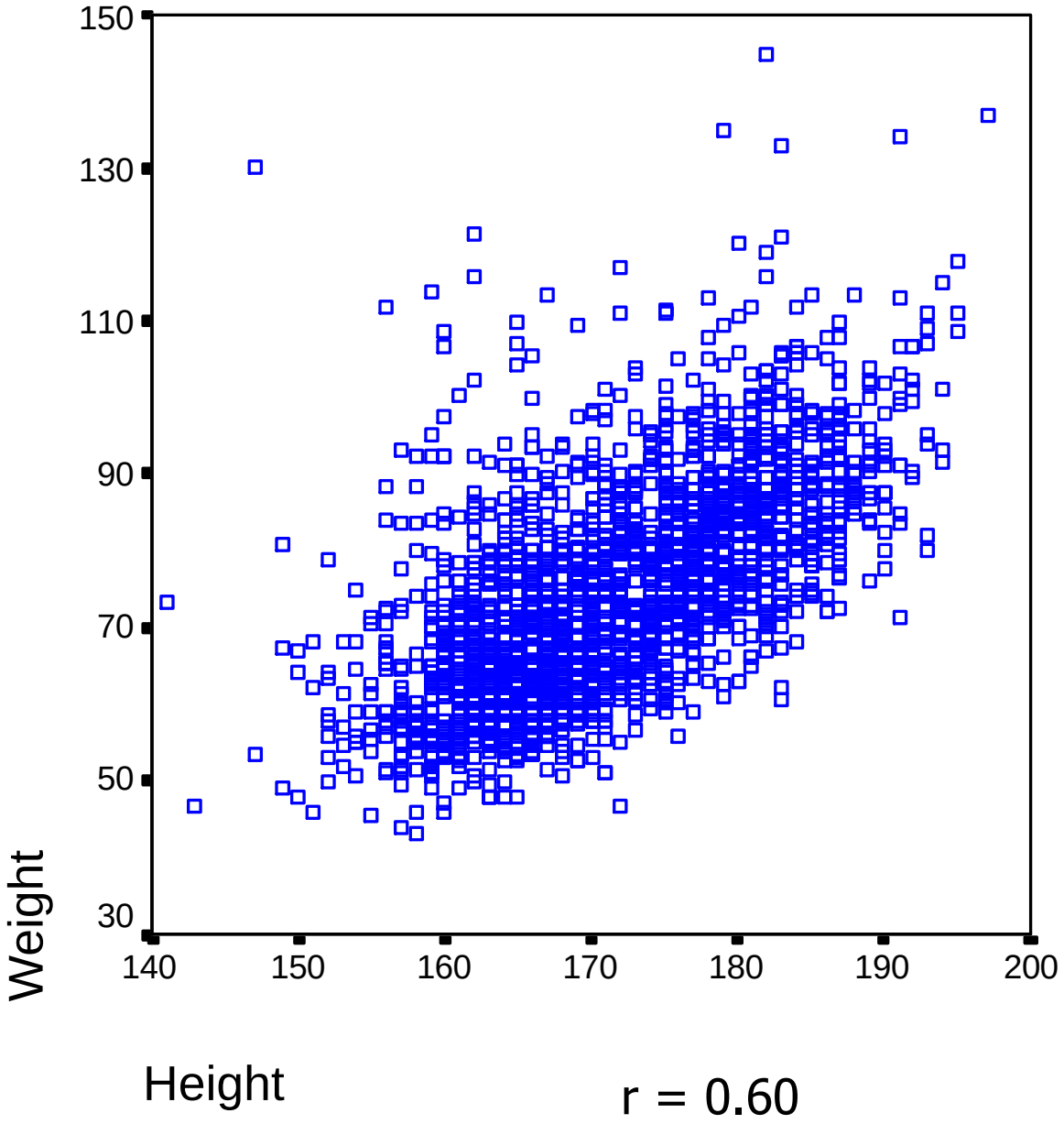
a. R Squared = ,020 (Adjusted R Squared = ,016)

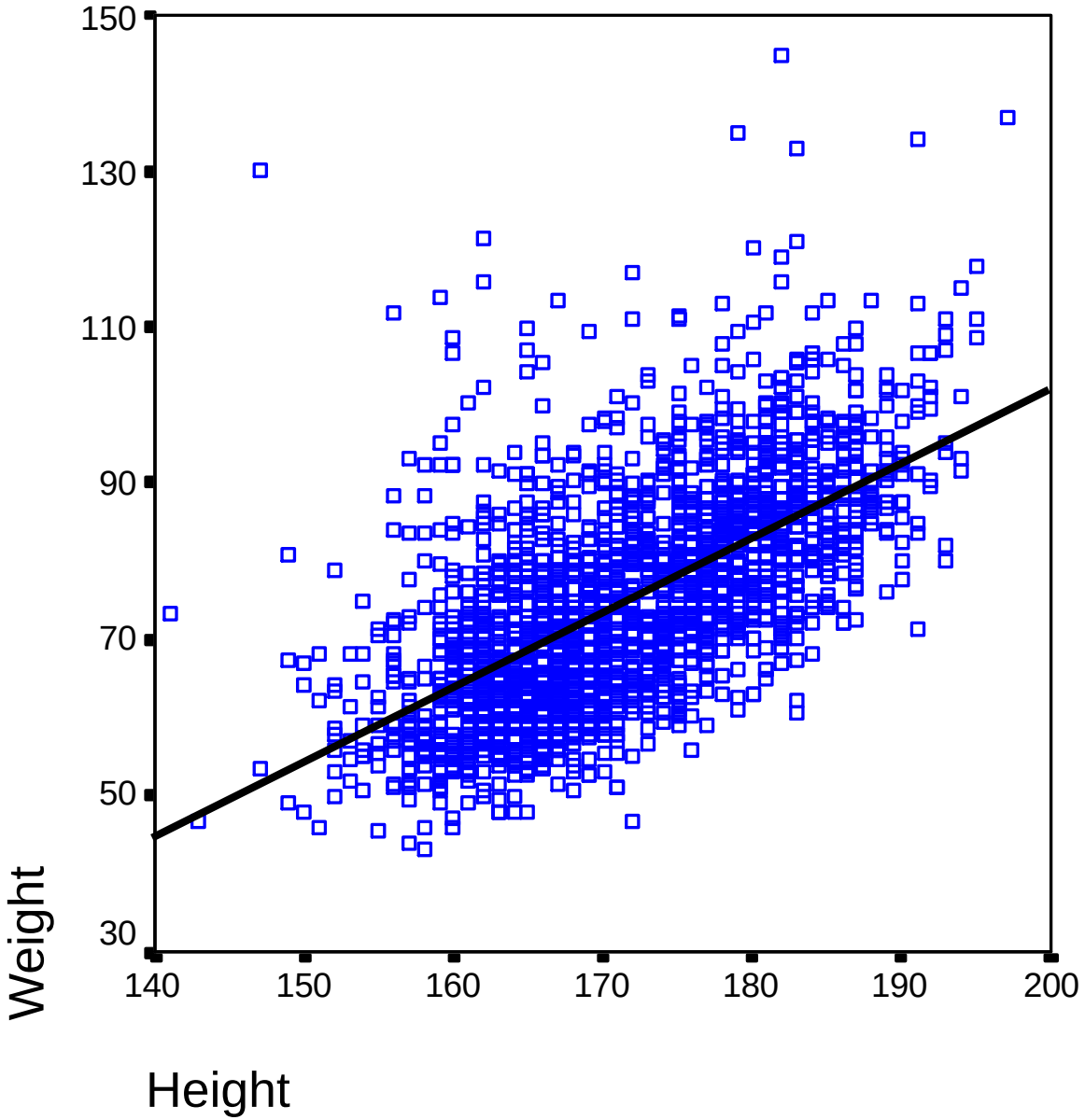
What statistical test to use (hypothesis testing)

Dependent variable (outcome variable)	**Research question**	**Model**
Ordinal/continuous	Comparing a mean and a fixed number (H_o: $\mu = \mu_o$)	One-sample t-test
(may be ranked)	Comparing the means of two groups (H_o: $\mu_1 = \mu_2$)	Two-sample t-test
	Comparing the mean in two paired groups (i.e. before and after treatment) $d = X_{after} - X_{before}$ (H_o: $\mu_d = 0$)	Paired t-test
	Comparing the means of 3 and more groups (H_o: $\mu_1 = \mu_2 = \mu_3$)	Analysis of variance (ANOVA)
	Comparing the means of groups according to two or more categorical variables (e.g. marital status and sex) (H_o: $\mu_1 = \mu_2$, $v_1 = v_2$)	Multiway ANOVA
	Association with another ordinal independent variable (dose-response) dep.var. = α+ β*(indep.var.) (H_o: $\beta = 0$)	Regression analysis, Correlation analysis
Nominal (categorical)	Comparing a proportion and a fixed number (H_o: $\pi = \pi_o$)	Binomial test
	Comparing proportions in subgroups (Crosstabs) (H_o: $\pi_1 = \pi_2$)	Chi-square test Fishers exact test
Two categories only	Association with several ordinal and nominal independent variables (e.g. dependent variable “dead/alive” – “disease(yes/no)”	Logistic regression
Time to an event	Between groups. Particularly useful for prospective design Censoring.	Life tables Kaplan-Meier curves
	Association with several ordinal and nominal independent variables	Cox-regression

What statistical test to use (hypothesis testing)

Dependent variable (outcome variable)	**Research question**	**Model**
Ordinal/continuous	Comparing a mean and a fixed number (H_o: $\mu = \mu_o$)	One-sample t-test
(may be ranked)	Comparing the means of two groups (H_o: $\mu_1 = \mu_2$)	Two-sample t-test
	Comparing the mean in two paired groups (i.e. before and after treatment) $d = X_{after} - X_{before}$ (H_o: $\mu_d = 0$)	Paired t-test
	Comparing the means of 3 and more groups (H_o: $\mu_1 = \mu_2 = \mu_3$)	Analysis of variance (ANOVA)
	Comparing the means of groups according to two or more categorical variables (e.g. marital status and sex) (H_o: $\mu_1 = \mu_2$, $v_1 = v_2$)	Multiway ANOVA
	Association with another ordinal independent variable (dose-response) dep.var. = α+ β*(indep.var.) (H_o: β = 0)	Regression analysis, Correlation analysis
Nominal (categorical)	Comparing a proportion and a fixed number (H_o: $\pi = \pi_o$)	Binomial test
	Comparing proportions in subgroups (Crosstabs) (H_o: $\pi_1 = \pi_2$)	Chi-square test Fishers exact test
Two categories only	Association with several ordinal and nominal independent variables (e.g. dependent variable “dead/alive” – “disease(yes/no)”	Logistic regression
Time to an event	Between groups. Particularly useful for prospective design Censoring.	Life tables Kaplan-Meier curves
	Association with several ordinal and nominal independent variables	Cox-regression


150
130
110
90
70
50
30
Weight
140
150
160
170
180
190
200
Height
r = 0.60



r = 0.60

Weight = -97 + 0.94 · Height

Regression analysis

- A regression analysis is modelling an association between one (or more) response variables (dependent variables) and predictor variables (independent variable).
- The association is estimated as regression coefficients.
- The regression coefficient indicates how much the dependent variable is changing when the independent variable is changing one unit.

$$Y = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i}$$

What statistical test to use (hypothesis testing)

Dependent variable (outcome variable)	**Research question**	**Model**
Ordinal/continuous	Comparing a mean and a fixed number (H_o: $\mu = \mu_o$)	One-sample t-test
(may be ranked)	Comparing the means of two groups (H_o: $\mu_1 = \mu_2$)	Two-sample t-test
	Comparing the mean in two paired groups (i.e. before and after treatment) $d = X_{after} - X_{before}$ (H_o: $\mu_d = 0$)	Paired t-test
General linear model	Comparing the means of 3 and more groups (H_o: $\mu_1 = \mu_2 = \mu_3$)	Analysis of variance (ANOVA)
	Comparing the means of groups according to two or more categorical variables (e.g. marital status and sex) (H_o: $\mu_1 = \mu_2$, $v_1 = v_2$)	Multiway ANOVA
	Association with another ordinal independent variable (dose-response) dep.var. = α+ β*(indep.var.) (H_o: $\beta = 0$)	Regression analysis, Correlation analysis
Nominal (categorical)	Comparing a proportion and a fixed number (H_o: $\pi = \pi_o$)	Binomial test
	Comparing proportions in subgroups (Crosstabs) (H_o: $\pi_1 = \pi_2$)	Chi-square test Fishers exact test
Two categories only	Association with several ordinal and nominal independent variables (e.g. dependent variable “dead/alive” – “disease(yes/no)”	Logistic regression
Time to an event	Between groups. Particularly useful for prospective design Censoring.	Life tables Kaplan-Meier curves
	Association with several ordinal and nominal independent variables	Cox-regression

What statistical test to use (hypothesis testing)

Dependent variable (outcome variable)	**Research question**	**Model**
Ordinal/continuous	Comparing a mean and a fixed number (H_o: $\mu = \mu_o$)	One-sample t-test
(may be ranked)	Comparing the means of two groups (H_o: $\mu_1 = \mu_2$)	Two-sample t-test
	Comparing the mean in two paired groups (i.e. before and after treatment) $d = X_{after} - X_{before}$ (H_o: $\mu_d = 0$)	Paired t-test
	Comparing the means of 3 and more groups (H_o: $\mu_1 = \mu_2 = \mu_3$)	Analysis of variance (ANOVA)
	Comparing the means of groups according to two or more categorical variables (e.g. marital status and sex) (H_o: $\mu_1 = \mu_2$, $v_1 = v_2$)	Multiway ANOVA
	Association with another ordinal independent variable (dose-response) dep.var. = α+ β*(indep.var.) (H_o: $\beta = 0$)	Regression analysis, Correlation analysis
Nominal (categorical)	Comparing a proportion and a fixed number (H_o: $\pi = \pi_o$)	Binomial test
	Comparing proportions in subgroups (Crosstabs) (H_o: $\pi_1 = \pi_2$)	Chi-square test Fishers exact test
Two categories only	Association with several ordinal and nominal independent variables (e.g. dependent variable “dead/alive” – “disease(yes/no)”	Logistic regression
Time to an event	Between groups. Particularly useful for prospective design Censoring.	Life tables Kaplan-Meier curves
	Association with several ordinal and nominal independent variables	Cox-regression

Horizontal: *job satisfaction*. Vertical: *age* (years). Each cell shows number of persons and percentage across rows (2.line) and across columns (3.line).

	Very good	Good	Not so good	Total
50 or below	16 (20) (50)	12 (15) (23)	52 (65) (33)	80 (100) (33)
Above 50	16 (10) (50)	40 (25) (77)	104 (65) (67)	160 (100) (67)
Total	32 (13.3) (100)	52 (21.7) (100)	156 (65) (100)	240 (100) (100)

Which percentages would you choose to present?
How would you render this table with words?
A chi-square test in this table gives a χ^2 of 6.46. What is the p-value?
How do you interpret this p-value?

Is there any other test you could use with these data to test whether there is a difference in job satisfaction between the young and older employees?

What statistical test to use (hypothesis testing)

Dependent variable (outcome variable)	**Research question**	**Model**
Ordinal/continuous	Comparing a mean and a fixed number (H_o: $\mu = \mu_o$)	One-sample t-test
(may be ranked)	Comparing the means of two groups (H_o: $\mu_1 = \mu_2$)	Two-sample t-test
	Comparing the mean in two paired groups (i.e. before and after treatment) $d = X_{after} - X_{before}$ (H_o: $\mu_d = 0$)	Paired t-test
	Comparing the means of 3 and more groups (H_o: $\mu_1 = \mu_2 = \mu_3$)	Analysis of variance (ANOVA)
	Comparing the means of groups according to two or more categorical variables (e.g. marital status and sex) (H_o: $\mu_1 = \mu_2$, $v_1 = v_2$)	Multiway ANOVA
	Association with another ordinal independent variable (dose-response) dep.var. = α+ β*(indep.var.) (H_o: $\beta = 0$)	Regression analysis, Correlation analysis
Nominal (categorical)	Comparing a proportion and a fixed number (H_o: $\pi = \pi_o$)	Binomial test
	Comparing proportions in subgroups (Crosstabs) (H_o: $\pi_1 = \pi_2$)	Chi-square test Fishers exact test
Two categories only	Association with several ordinal and nominal independent variables (e.g. dependent variable “dead/alive” – “disease(yes/no)”	Logistic regression
Time to an event	Between groups. Particularly useful for prospective design Censoring.	Life tables Kaplan-Meier curves
	Association with several ordinal and nominal independent variables	Cox-regression

Logistic regression

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	røyk	1,037	,175	35,153	1	,000	2,820
	kjønn	,843	,144	34,256	1	,000	2,324
	alder	,135	,030	20,307	1	,000	1,144
	alccat2	-,090	,036	6,146	1	,013	,914
	Constant	-13,032	1,347	93,613	1	,000	,000

a. Variable(s) entered on step 1: røyk, kjønn, alder, alccat2.

What statistical test to use (hypothesis testing)

Dependent variable (outcome variable)	**Research question**	**Model**
Ordinal/continuous	Comparing a mean and a fixed number (H_o: $\mu = \mu_o$)	One-sample t-test
(may be ranked)	Comparing the means of two groups (H_o: $\mu_1 = \mu_2$)	Two-sample t-test
	Comparing the mean in two paired groups (i.e. before and after treatment) $d = X_{after} - X_{before}$ (H_o: $\mu_d = 0$)	Paired t-test
	Comparing the means of 3 and more groups (H_o: $\mu_1 = \mu_2 = \mu_3$)	Analysis of variance (ANOVA)
	Comparing the means of groups according to two or more categorical variables (e.g. marital status and sex) (H_o: $\mu_1 = \mu_2$, $v_1 = v_2$)	Multiway ANOVA
	Association with another ordinal independent variable (dose-response) dep.var. = α+ β*(indep.var.) (H_o: $\beta = 0$)	Regression analysis, Correlation analysis
Nominal (categorical)	Comparing a proportion and a fixed number (H_o: $\pi = \pi_o$)	Binomial test
	Comparing proportions in subgroups (Crosstabs) (H_o: $\pi_1 = \pi_2$)	Chi-square test Fishers exact test
Two categories only	Association with several ordinal and nominal independent variables (e.g. dependent variable “dead/alive” – “disease(yes/no)”	Logistic regression
Time to an event	Between groups. Particularly useful for prospective design Censoring.	Life tables Kaplan-Meier curves
	Association with several ordinal and nominal independent variables	Cox-regression

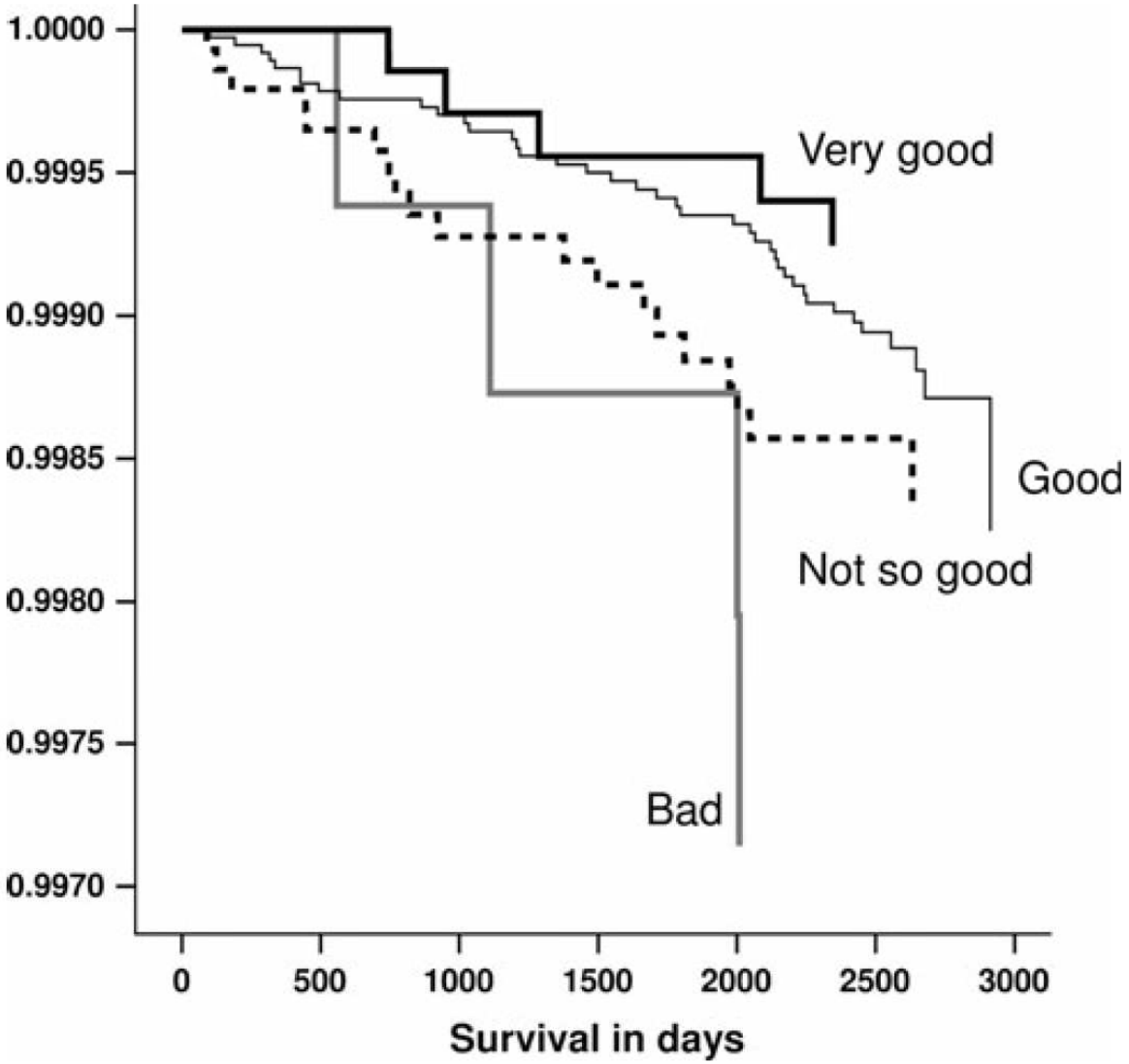


Fig. 2 Cumulative risk of lung cancer according to self-rated health. Survival curves from Cox regression analysis for the different categories of self-rated health according to lung cancer among 24,487 respondents in the HUSK study 1997–1999. Adjusted for age cohort, gender, light physical activity, pack years, alcohol, and BMI

What statistical test to use (hypothesis testing)

Dependent variable (outcome variable)	**Research question**	**Model**
Ordinal/continuous	Comparing a mean and a fixed number (H_o: $\mu = \mu_o$)	One-sample t-test
(may be ranked)	Comparing the means of two groups (H_o: $\mu_1 = \mu_2$)	Two-sample t-test
	Comparing the mean in two paired groups (i.e. before and after treatment) $d = X_{after} - X_{before}$ (H_o: $\mu_d = 0$)	Paired t-test
	Comparing the means of 3 and more groups (H_o: $\mu_1 = \mu_2 = \mu_3$)	Analysis of variance (ANOVA)
	Comparing the means of groups according to two or more categorical variables (e.g. marital status and sex) (H_o: $\mu_1 = \mu_2$, $v_1 = v_2$)	Multiway ANOVA
	Association with another ordinal independent variable (dose-response) dep.var. = α+ β*(indep.var.) (H_o: $\beta = 0$)	Regression analysis, Correlation analysis
Nominal (categorical)	Comparing a proportion and a fixed number (H_o: $\pi = \pi_o$)	Binomial test
	Comparing proportions in subgroups (Crosstabs) (H_o: $\pi_1 = \pi_2$)	Chi-square test Fishers exact test
Two categories only	Association with several ordinal and nominal independent variables (e.g. dependent variable “dead/alive” – “disease(yes/no)”	Logistic regression
Time to an event	Between groups. Particularly useful for prospective design Censoring.	Life tables Kaplan-Meier curves
	Association with several ordinal and nominal independent variables	Cox-regression

Table 3 Distribution of significant covariates (crude) for lung cancer among 24,487 respondents in the HUSK study 1997–1999

	N (%)	Lung cancer (%)	Crude[a] HR (CI 95 %)	*p*-Trend/*p*-overall	Adjusted[b] HR (CI 95 %)	*p*-Trend/*p*-overall
Self-rated health						
Very good	4,651 (19.0)	7 (0.2)	1 (ref.)	0.001/0.007	1 (ref.)	0.038/0.20
Good	15,727 (64.2)	45 (0.3)	1.50 (0.67–3.34)		1.57 (0.61–4.02)	
Not so good	3,594 (14.7)	23 (0.6)	2.54 (1.07–6.03)		2.13 (0.77–5.89)	
Poor	276 (1.1)	4 (1.5)	6.13 (1.77–21.19)		3.88 (0.97–15.50)	
Light physical activity						
3 h or more per week	10,660 (43.5)	30 (0.5)	1 (ref.)	0.002/0.005	1 (ref.)	0.077/0.16
Between 1 and 2 h per week	8,664 (35.4)	23 (0.3)	1.16 (0.67–2.00)		1.09 (0.61–1.93)	
<1 h per week	3,218 (13.1)	11 (0.3)	1.59 (0.79–3.20)		1.13 (0.52–2.43)	
No such activity	1,106 (4.5)	11 (1.0)	3.41 (1.71–6.81)		2.36 (1.11–5.01)	
Pack years						
None smokers	9,266 (37.8)	6 (0.1)	1 (ref.)	<0.001/<0.001	1 (ref.)	<0.001/<0.001
0–5 pack years	3,889 (15.8)	5 (0.1)	2.60 (0.79–8.56)		3.21 (0.86–12.01)	
5–15 pack years	6,027 (24.6)	17 (0.3)	5.83 (2.28–14.912)		6.03 (1.98–18.32)	
15–30 pack years	3,999 (16.3)	33 (0.8)	16.57 (6.87–39.94)		15.85 (5.49–45.76)	
>30 pack years	707 (2.9)	15 (2.1)	25.55 (9.65–67.65)		24.98 (7.87–79.31)	
Alcohol per 2 weeks						
None drinker/teetotaler	5,966 (24.4)	19 (0.3)	1 (ref.)	0.044/0.13	1 (ref.)	0.040/0.58
Low	13,861 (56.6)	41 (0.3)	1.42 (0.82–2.45)		1.33 (0.74–2.40)	
Medium and high	2,941 (12.0)	13 (0.4)	2.11 (1.02–4.40)		1.43 (0.65–3.12)	
BMI						
20–25	11,125 (45.4)	36 (0.3)	1 (ref.)	–/<0.001	1 (ref.)	–/<0.001
0–19.99	1,027 (4.2)	11 (1.1)	3.68 (1.85–7.31)		3.14 (1.55–6.36)	
>26	12,289 (50.2)	31 (0.3)	0.61 (0.38–0.99)		0.63 (0.37–1.07)	

[a] Adjusted for age cohort and gender

[b] Adjusted for age cohort, gender, light physical activity, pack years, BMI, alcohol, and SRH

Summary

- The sample needs to be representative in order to be generalized
- The number of non-responders should not be too high
- The conclusion of a “real” difference between two groups or a “real” association between to variables is based on: the probability that this difference/association is a chance finding is less than 0.05 ($p < 0.05$)
- A statistically significant finding does not necessarily imply causality
- Multivariate analyses are used to adjust for other factors than those of primary interest

Courses in Statistics at the Dept of Global Public Health and Primary Care

Code	Description	# days	# stud	Credits	When
HELSTA	Introductory course in statistics (master)	6 d	80	5	Fall
MEDSTA2	Regression analysis in medical science (Stata)	6 d	25	5	Spring
MEDSTATA	Introduction to Stata	2,5 d	25	3	Fall
MEDSTA3	Analysis of longitudinal data using Stata	2 d	15	2	Spring