Hypothesis Testing and Confidence Intervals

Håkon K. Gjessing

Professor/Principal Investigator

Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo

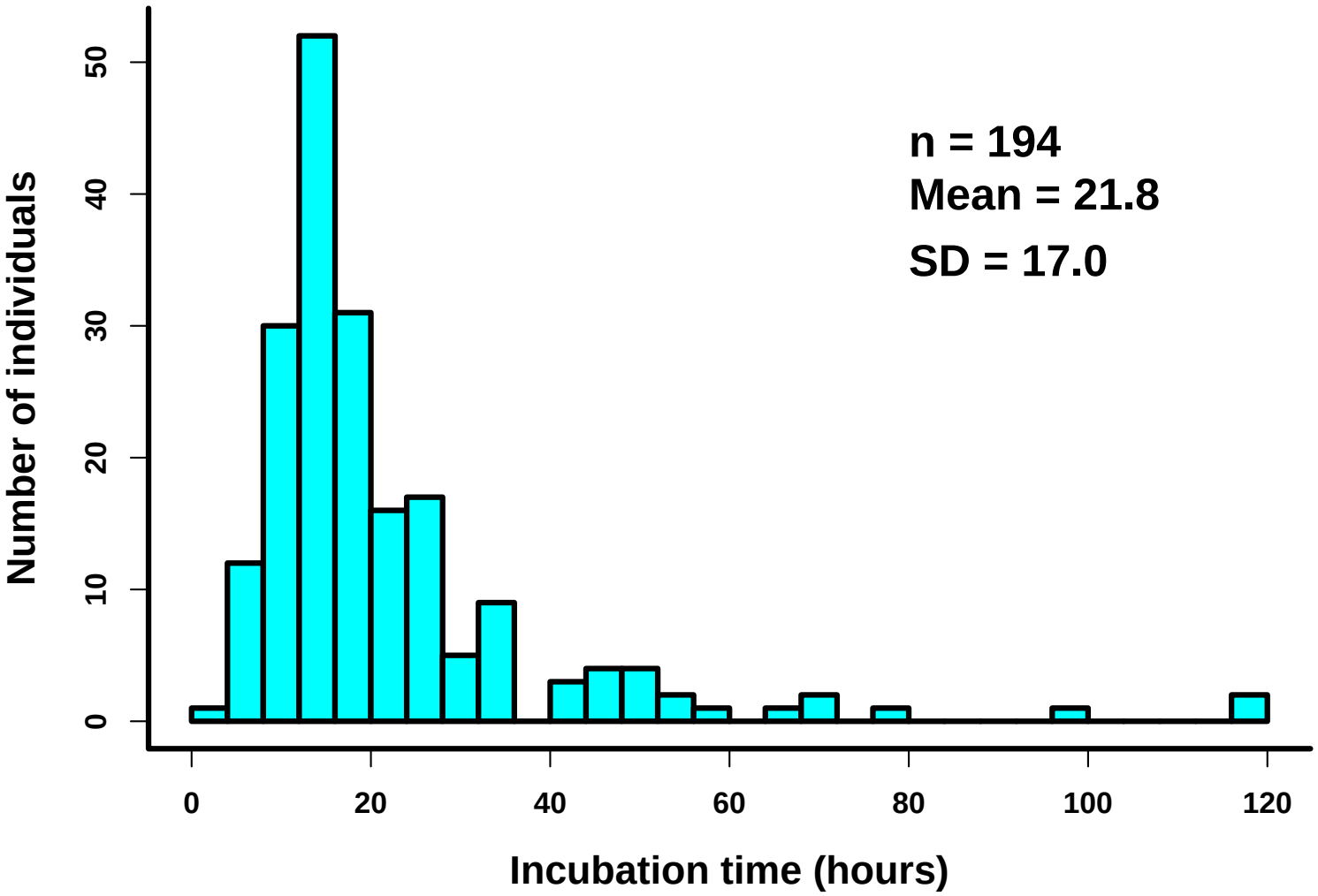
Department of Global Public Health and Primary Care, University of Bergen

Makerere

Wednesday, 7 June 2023

CENTRE FOR FERTILITY AND HEALTH

HISTOGRAM OF INCUBATION TIMES FOR 194 INDIVIDUALS



CENTRE FOR FERTILITY AND HEALTH

Confidence Intervals (CI)

Confidence interval based on the Normal distribution:

- $\overline{X}$ is approximately normally distributed ... so the interval

$$(\overline{X} - 1.96 \cdot \text{SE}(\overline{X}), \quad \overline{X} + 1.96 \cdot \text{SE}(\overline{X}))$$

has a 95% chance of covering the "true" μ.

Lower limit:

$$\overline{X} - 1.96 \cdot \text{SE}(\overline{X}) = 21.8 - 1.96 \cdot 1.22 = 19.4$$

Upper limit:

$$\overline{X} + 1.96 \cdot \text{SE}(\overline{X}) = 21.8 + 1.96 \cdot 1.22 = 24.2$$

95% confidence interval: (19.4, 24.2)

CENTRE FOR FERTILITY AND HEALTH

HYPOTHESIS TESTING, NORMAL DISTRIBUTION

One-sided test:

- Null hypothesis $H_0 : \mu \leqslant 18$ hours
- Alternative hypothesis $H_A : \mu > 18$ hours

- If $\mu = 18$ happened to be true, we should have $\overline{X} \approx 18$
- We actually *observe* $\overline{X} = 21.8$
- Remember: $SE(\overline{X}) \approx 1.22$

P-VALUE

The probability of observing $\overline{X}$ *at least* 21.8

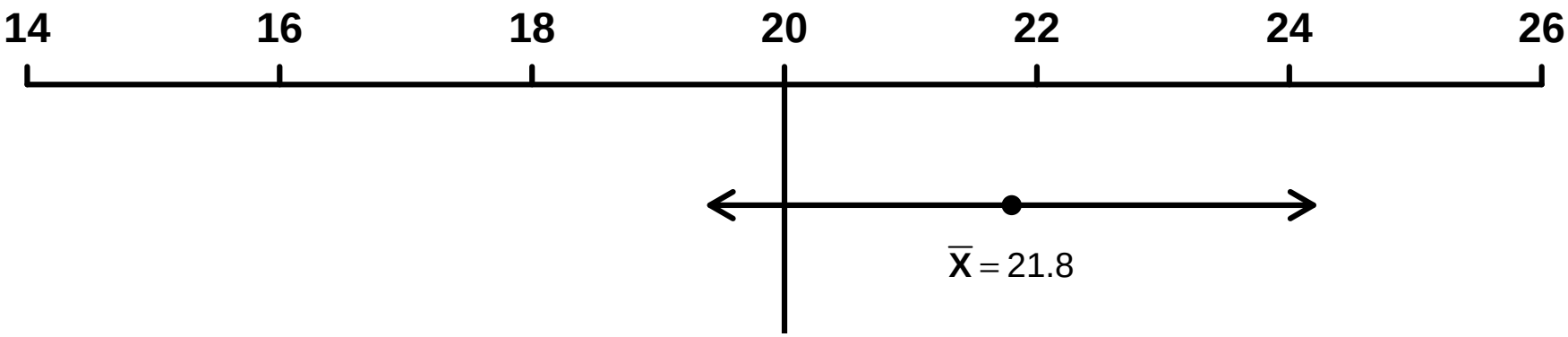
if $H_0 : \mu = 18$ were true

So, is $\overline{X}$ larger than just random variation away from $\mu = 18$?

CENTRE FOR FERTILITY AND HEALTH

CONFIDENCE INTERVALS (95% CI)

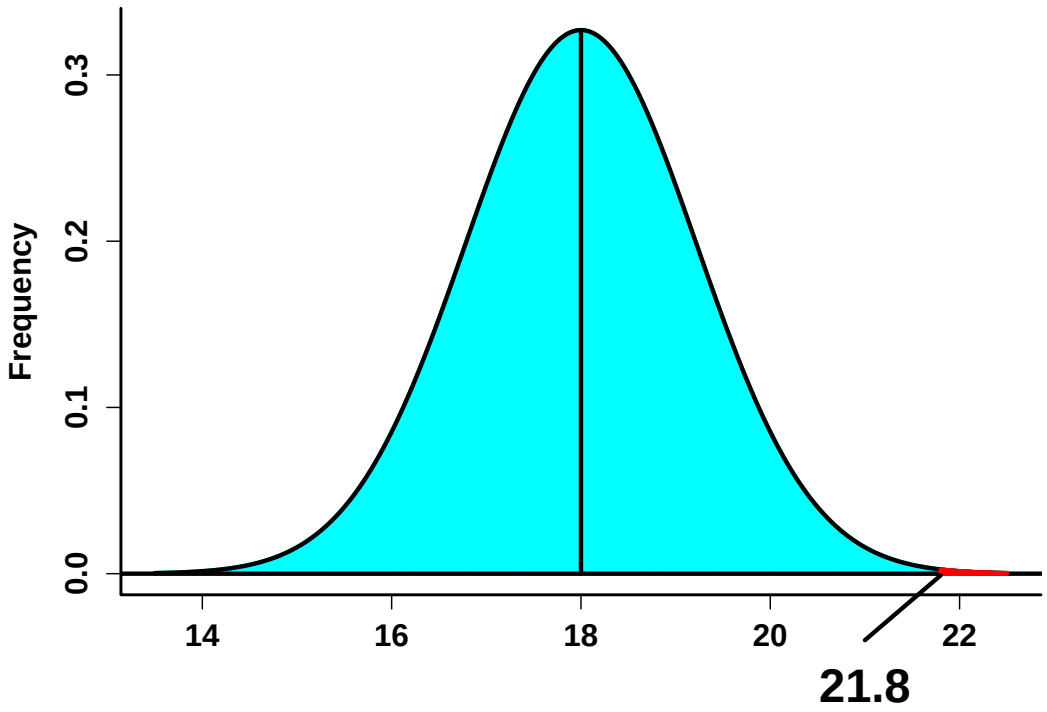
Remember from previous lecture:



- It seems clear that 18 is unlikely given the 95% confidence interval.
- But can we be more precise in *how* unlikely??

CENTRE FOR FERTILITY AND HEALTH

HYPOTHESIS TESTING: COMPUTE THE Z-SCORE AND P-VALUE



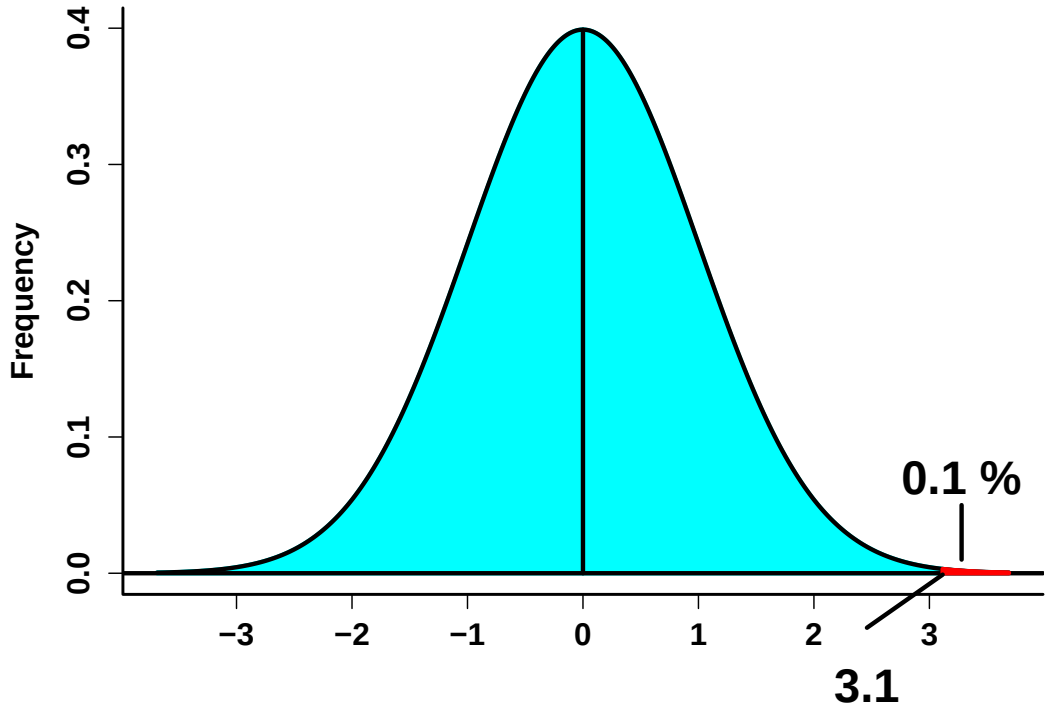
Z-score:

$$Z = \frac{21.8 - 18}{1.22} = 3.1$$

That is, $\overline{X}$ is more than 3 SE away from $\mu = 18$

CENTRE FOR FERTILITY AND HEALTH

HYPOTHESIS TESTING: COMPUTE THE Z-SCORE AND P-VALUE



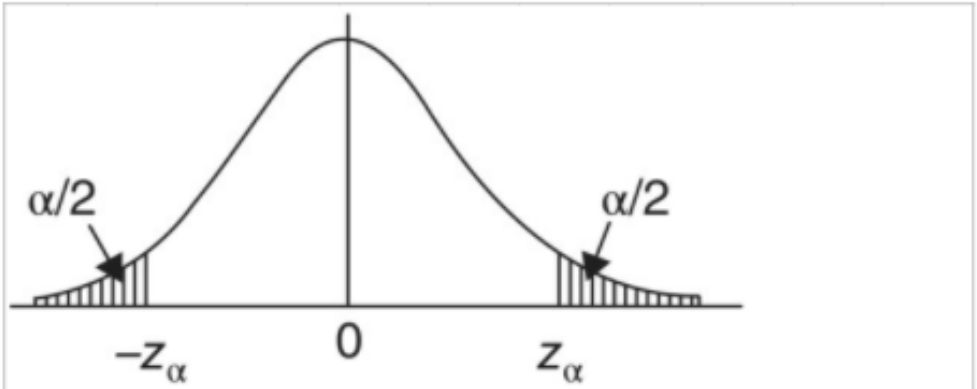
P-value = probability that $Z \geqslant 3.1$ in a standard Normal distribution

CENTRE FOR FERTILITY AND HEALTH

TEXTBOOK: THE NORMAL DISTRIBUTION

Appendix
Statistical Tables

Table T1 The Standard Normal distribution. The value tabulated is the probability, α, that a random variable, Normally distributed with mean zero and standard deviation one, will b



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	0.8026	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

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
3.10	0.0019	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014
3.20	0.0014	0.0013	0.0013	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010
3.30	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.40	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	0.0005

P-VALUE

From tables of the Normal distribution (see textbook):

P-value $\approx$ **0.0020**$/$**2** $=$ **0.001**

Not likely to be a coincidence: **H_0 rejected!**

- P-value is the probability of an extreme result *if* H_0 is true.
- But where should we draw the line for "coincidence"??

LEVEL OF THE TEST

Typical limit for "extreme": **5%**

... this level makes sense..., but mostly chosen because of tradition.

CENTRE FOR FERTILITY AND HEALTH

TWO-SIDED TEST

- Null hypothesis $\mathrm{H}_0 : \mu = 18$ hours
- Alternative hypothesis $\mathrm{H}_A : \mu \neq 18$ hours

Two-sided p-value = 2 * One-sided p-value

P-value = 2 * 0.001 = 0.002

We reject H_0 since p-value is less than 5%

But there is a **DILEMMA:**

Two-sided tests *always* reject H_0 if enough data!

(A bit weird... but still sometimes useful)

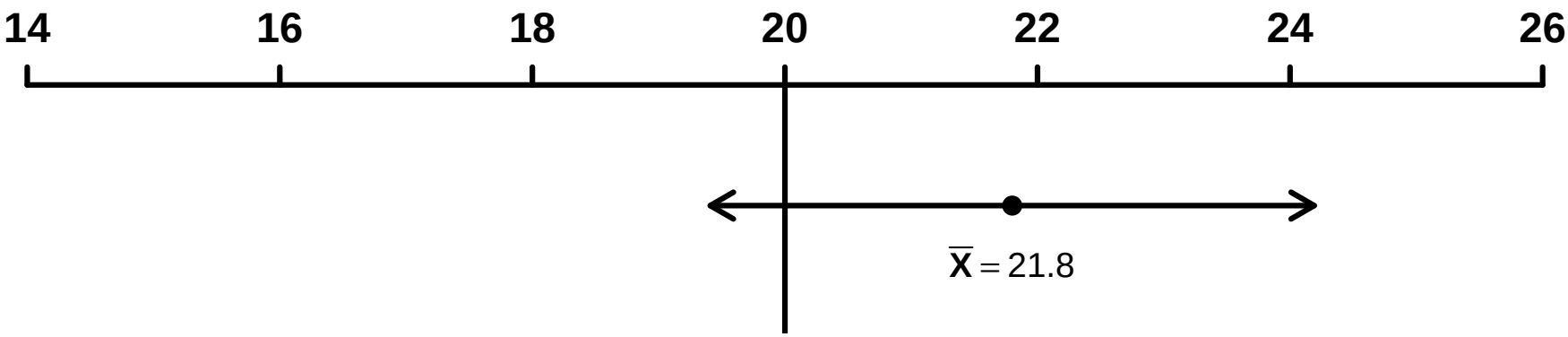
CENTRE FOR FERTILITY AND HEALTH

Confidence intervals for testing?

- Can confidence intervals be used for hypothesis testing?
- *Should* confidence intervals be used for hypothesis testing?

CENTRE FOR FERTILITY AND HEALTH

CONFIDENCE INTERVALS (95% CI)



- Test by checking if H_0 is outside 95% CI.
- This corresponds to a *two-sided* test with level 5%.
- **But** test and CI can sometimes give different conclusions.

Recommendation:

- Use confidence intervals to measure precision!
- Use hypothesis tests (with p-value) for testing!

(Sounds logical, doesn't it)

CENTRE FOR FERTILITY AND HEALTH

Causes of death in three Norwegian counties[1]

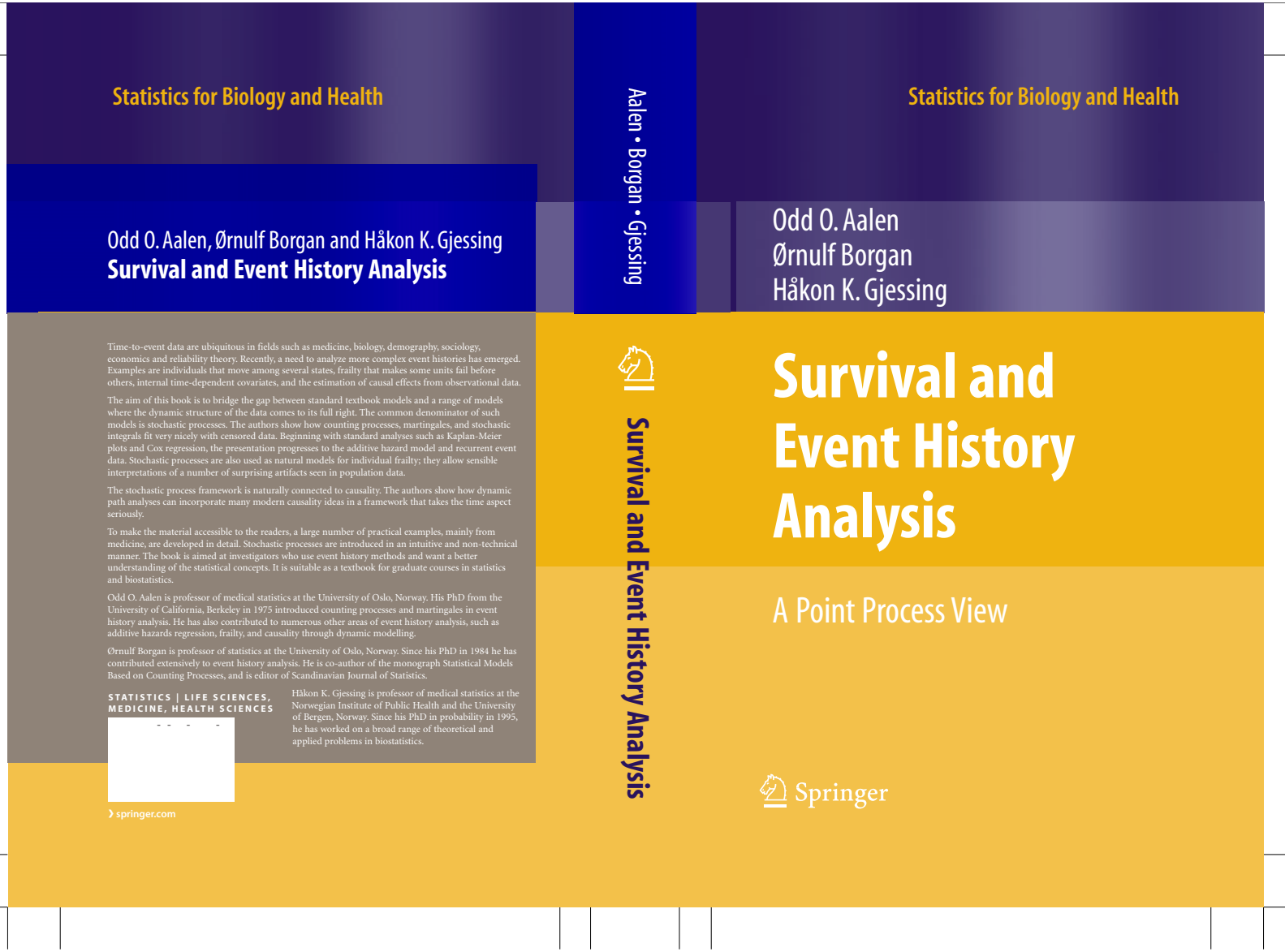
- Data collection 1974–78
- All men and women aged 35–49 years
- Three rural Norwegian counties:
  Oppland, Sogn og Fjordane, and Finmark
- Cardiovascular health screening examination
- More than 90% participation
- Self-report on past and current **smoking habits**
- Mortality in cohort followed-up to the end in year 2000
- Linked to **cause of death** registry at Statistics Norway
- A subset: 4000 (random) out of about 50000
  (http://folk.uio.no/borgan/abg-2008/data/data.html)

- **Focus: Mortality 40-70 years**

1: Vollset SE, Tverdal A, Gjessing HK. Smoking and deaths between 40 and 70 years of age in women and men. Annals of internal medicine. 2006;144(6):381.

CENTRE FOR FERTILITY AND HEALTH

TOTALLY INDECENT SELF-PROMOTION



(Includes the example data analyzed using survival analysis)

CENTRE FOR FERTILITY AND HEALTH

SINGLE PROPORTION

- Risk of dying age 40–70, all causes (ignoring length of follow-up)
- X = 586 total number of deaths
- n = 4000 total number of individuals
- p is the proportion (risk) of dying, estimated by $\widehat{p}$:

$$\widehat{p} = \frac{X}{n} = \frac{586}{4000} = 0.1465 = 14.7\%$$

CENTRE FOR FERTILITY AND HEALTH

Single proportion

- Risk of dying age 40–70, all causes (ignoring length of follow-up)
- X = 586 total number of deaths
- n = 4000 total number of individuals
- p is the proportion (risk) of dying, estimated by $\widehat{p}$:

$$\widehat{p} = \frac{X}{n} = \frac{586}{4000} = 0.1465 = 14.7\%$$

What is the **uncertainty** of this estimate? **Standard Error:**

$$SE(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \sqrt{\frac{0.1465(1-0.1465)}{4000}} = 0.00559$$

CENTRE FOR FERTILITY AND HEALTH

SINGLE PROPORTION

- Risk of dying age 40–70, all causes (ignoring length of follow-up)
- X = 586 total number of deaths
- n = 4000 total number of individuals
- p is the proportion (risk) of dying, estimated by $\widehat{p}$:

$$\widehat{p} = \frac{X}{n} = \frac{586}{4000} = 0.1465 = 14.7\%$$

What is the **uncertainty** of this estimate? **Standard Error:**

$$SE(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \sqrt{\frac{0.1465(1-0.1465)}{4000}} = 0.00559$$

Standard 95% Confidence Interval for p

(Normal approximation, Wald interval):

$$\widehat{p} \pm 1.96 \cdot SE(\widehat{p}) = 0.1465 \pm 1.96 \cdot 0.00559 = (0.136, 0.157) = (13.6\%, 15.7\%)$$

CENTRE FOR FERTILITY AND HEALTH

Two proportions: Compare males and females

- Overall death risk, by gender:

	Males	**Females**
Population proportions:	p_M	p_F
X (deaths)	398	188
n (total)	2086	1914
$\widehat{p}$	19.1%	9.8%
$SE(\widehat{p})$	0.00860	0.00680
95% CI each group:	(17.4%, 20.8%)	(8.6%, 11.2%)

CENTRE FOR FERTILITY AND HEALTH

Two proportions: Compare males and females

- Non-overlapping confidence intervals means a significant difference (**almost** always...)
- **Warning:** overlapping does NOT mean non-significant

- BUT: Would also like to **measure** gender difference (not only whether significant or non-significant)

CENTRE FOR FERTILITY AND HEALTH

Risk Difference (RD)

	Males	Females
X (deaths)	398	188
n (total)	2086	1914
$\widehat{p}$	19.1%	9.8%
$SE(\widehat{p})$	0.00860	0.00680

Risk Difference:

$$RD = \widehat{p}_M - \widehat{p}_F = 19.1\% - 9.8\% = 9.3\%$$

CENTRE FOR FERTILITY AND HEALTH

Risk Difference (RD)

	Males	Females
X (deaths)	398	188
n (total)	2086	1914
$\widehat{p}$	19.1%	9.8%
$SE(\widehat{p})$	0.00860	0.00680

Risk Difference:

$$RD = \widehat{p}_M - \widehat{p}_F = 19.1\% - 9.8\% = 9.3\%$$

- Uncertainty of RD (i.e. standard error of RD):

$$SE(RD) = \sqrt{SE(\widehat{p}_M)^2 + SE(\widehat{p}_F)^2} = \sqrt{0.00860^2 + 0.00680^2} = 0.011$$

CENTRE FOR FERTILITY AND HEALTH

NOTE ON TEXTBOOK FORMULAS

- SE for confidence intervals:

$$\mathrm{SE}(\mathrm{RD}) = \sqrt{\mathrm{SE}(\hat{p}_1)^2 + \mathrm{SE}(\hat{p}_2)^2}$$

$$= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- SE for hypothesis testing:

$$\mathrm{SE}(\mathrm{RD}) = \sqrt{\hat{p}(1-\hat{p})\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}$$

You can use the first BOTH for confidence intervals AND hypothesis testing. (But it illustrates that Standard Errors can be calculated in many different ways, depending on assumptions)

CENTRE FOR FERTILITY AND HEALTH

Risk Difference (RD): Confidence Interval

$$\text{RD} = \widehat{p}_M - \widehat{p}_F = 19.1\% - 9.8\% = 9.3\%$$

$$\text{SE}(\text{RD}) = \sqrt{\text{SE}(\widehat{p}_M)^2 + \text{SE}(\widehat{p}_F)^2} = \sqrt{0.00860^2 + 0.00680^2} = 0.011$$

95% Confidence Interval of RD (Wald interval):

$$\text{RD} \pm 1.96 \cdot \text{SE}(\text{RD}) = 0.093 \pm 1.96 \cdot 0.011$$

(7.1%, 11.4%)

CENTRE FOR FERTILITY AND HEALTH

Risk Difference (RD): Significance Test

$$\text{RD} = \hat{p}_M - \hat{p}_F = 19.1\% - 9.8\% = 9.3\%$$

$$\text{SE}(\text{RD}) = \sqrt{\text{SE}(\hat{p}_M)^2 + \text{SE}(\hat{p}_F)^2} = \sqrt{0.00860^2 + 0.00680^2} = 0.011$$

Null hypothesis:

$$H_0 : p_M = p_F \qquad \text{that is,} \qquad \text{RD}_0 = p_M - p_F = 0$$

Test statistic:

$$\textbf{Z-score} = \frac{\text{RD} - \text{RD}_0}{\text{SE}(\text{RD})} = \frac{0.093 - 0}{0.011} \approx 8.5$$

- Is RD larger than random variation away from $\text{RD}_0 = 0$ should allow?

CENTRE FOR FERTILITY AND HEALTH

Test of significance, Conclusion

- RD is 8.5 standard deviations away from zero.
- Extremely unlikely by chance (WAY outside table in book).
- H_0 rejected.
- Higher mortality among men than women.

CENTRE FOR FERTILITY AND HEALTH

Is significance really significant?

- P-value measures degree of chance.
- NOT the actual **importance** of the result.
- Statistical significance is not the same as biological significance.
- Confidence intervals better for deciding biological importance.

In our example, more than 9 percentage points higher mortality is both statistically and biologically significant!

CENTRE FOR FERTILITY AND HEALTH