The following is based on a paper by Cranor (1990)

Cranor, C.F. 1990. Some moral issues in risk assessment. *Ethics*, 101:123-143. (available form JSTOR – viva MMU library electronic journals)

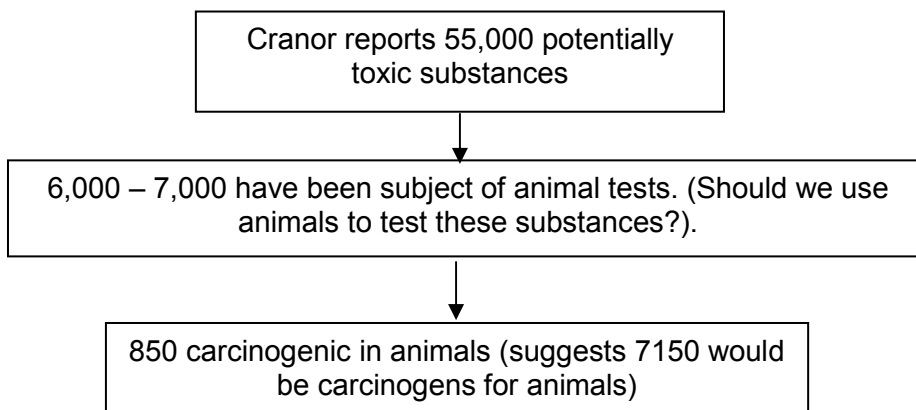US Cancer deaths – approximately 400,000 per year

Major causes (approximate figure, with ranges)
Tobacco consumption   120,000 (100,000 – 160,000)
Alcohol consumption   16,000 (12,000 – 20,000)

Thus, approximately 140,000 may be 'self-inflicted', people can reduce their risk by changing habits.

Workplace exposure   100,000 (50,000 – 150,000)
Air pollution   8,000
Diet   140,000

Thus, most cancers are related to substances that people are involuntarily exposed to. It is not clear how many diet-related cancers are due to 'natural' substances, and how may are due to artificial chemical additives.

```
┌─────────────────────────────────────┐
│   Cranor reports 55,000 potentially  │
│            toxic substances          │
└─────────────────────────────────────┘
                    │
                    ▼
┌─────────────────────────────────────────────────┐
│ 6,000 – 7,000 have been subject of animal tests. │
│   (Should we use animals to test these           │
│   substances?).                                  │
└─────────────────────────────────────────────────┘
                    │
                    ▼
┌─────────────────────────────────────────────┐
│ 850 carcinogenic in animals (suggests 7150   │
│      would be carcinogens for animals)       │
└─────────────────────────────────────────────┘
```

Other tests suggest higher proportion are carcinogens.

How can we assess which chemicals are carcinogens?
- Molecular structure analysis.
- Cell tissue experiments
- Short and/or long-term bioassays in laboratory animals
- Human epidemiological studies.

Human epidemiological studies provide the best evidence for harmful effects on humans – *as long as the study is well designed*.

These studies involve comparisons between populations: a control and an 'at-risk' population – retrospective (historical) or prospective (follow forward in time from the start of a study) cohort observational studies

Cranor argues that '*the design and outcome of epidemiological studies may be the product of obvious and potentially controversial equivalents of moral judgements when:*
- *background disease rate is low (e.g. leukaemia < 1/10,000) or*
- *sample populations are small (cost considerations or only a small number are exposed to a toxic risk) or*
- *scientific conventions are adopted uncritically.*
…unless scientists are scrupulously careful to be objective…..scientific studies used for estimating risk to humans health for regulatory purposes can be considerably more controversial and political than most people think'

'The moral and policy judgements are forced by the statistical equations themselves..'

What does he mean by this?

Background

Relative risk = incidence among exposed / incidence among non-exposed, e.g.

Lung cancer incidence in non-exposed (non-smokers) is 7/100,000
Lung cancer incidence in exposed (heavy smokers) is 166/100,000

relative risk = 166 / 7 = 23.7

- At what point does the relative risk become large enough that society should be concerned and control measures introduced (e.g. chemicals banned)?
- What relative risks ($\equiv$ effect size) can we detect in realistic studies?

Significance level (alpha) = risk of a Type I error (saying there is an effect when there isn't).

Typically set at 0.05 because we want to be reasonably certain that observed effects are real (therefore will be predictable and repeatable) and not just due to chance.

You will struggle to gain scientific acceptability with $\alpha > 0.05$.

Beta = risk of Type II error (not finding an effect when it exists), $1 - \beta$ = Power.

Typically β should be a <= 0.2 (giving minimum power of 0.8).

Think of α, β and power as 'measures of risk' or 'standards of proof'.
α is the risk of a false positive result.
β is the risk of a false negative result.

What chance of error should you adopt when undertaking an epidemiological (or any) study?

Is it is worth a 20% gamble β = 0.2, 80% power) if a study shows that workers or the public are not contracting cancer from exposure to chemical when (unbeknown to all) they are?

If α, β and power are 'standards of proof' then how much proof do you need and does it change with circumstances?

"Must researchers be 51% sure that benzene is a carcinogen presenting a risk to employees before regulating it in the workplace or should scientists in agencies be permitted to take a 49% chance (β = 0.49, power = 0.51) that substances are not high risk carcinogens to the populace, when in fact they might be?"

There are 4 important, *and interlinked*, variables in an epidemiological study.
- alpha
- beta
- relative risk
- sample size

If any 3 are known the fourth can be found. If α and β are fixed then the magnitude of relative risk that can be detected is related to the sample size, larger samples can identify smaller relative risks.

Cranor's example

Suppose
- A particular cancer has a normal incidence of 8 / 10,000.
- We suspect that a particular chemical may be a carcinogen for this cancer.
- A relative risk of 3 is thought to be a serious risk and worth investigating for public health purposes.

How does the design of our study impinge on our ability to detect the relative risk associated with this chemical. Two groups of study are considered, A & B.

Group A studies – we can alter the sample size.

A1 Set false+ and false- risks equal, i.e. $\alpha = \beta = 0.05$ (95% power).

This means that we have 95% chance of detecting a relative risk of 3 or more if it exists and only a 5% chance of detecting a non-existent risk.

Sample size required for this study is 13,495 (almost 27,000 in total).

Suppose we cannot afford these sample sizes.

A2 Keep α at 0.05 (for scientific acceptability), increase β to 0.2 (80% power).

Now we have a 1 in 5 chance of type II error – failing to detect a relative risk of 3 when it exists.

Sample size required for this study is 7,695 (almost 15,500 in total).

Even with reduced power these sample sizes are likely to cause problems.
- Can we find sufficient people in both populations?
- Can we study such large groups – do we have sufficient staff?

The impracticalities will probably mean that we have to study a smaller sample size.

Group B studies – sample size is fixed at 2150 (e.g. no more people exposed or insufficient funds for a larger study).

B1 Keep α at 0.05 (for scientific acceptability), and β at 0.2 (80% power).

Since the sample size is fixed the question is now what relative risk can we identify with α = 0.05 and 80% power?

The answer is 6. Thus, even with a 1 in 5 chance of failing to detect it, the smallest risk that we can reliably detect is twice as large as the figure we consider to be a serious risk.

If our study has p value >0.05, suggesting 'no risk' the best that we could really claim for the study is that the relative risk for the exposed group is not as high as 3.

B2. Keep α at 0.05 (for scientific acceptability), increase β to 0.49 (51% power).

Since the sample size is fixed the question is now what relative risk can we identify with α = 0.05 and 80% power?

The answer is 3.8.

In addition there is now a 49% chance of mistaking a toxic substance as benign. We might as well toss a con. Would you like to see chemicals regulated on the toss of coin?

B3 Increase α to 0.33 (which would be scientifically unacceptable) and set β to 0.2.

This would give us an 80% chance of detecting a relative risk of 3, if it existed. There is also a 1 in 3 chance of identifying a relative risk of 3 when it is lower than that.

Would we risk scientific unacceptability for a greater chance of detecting a toxic substance?

"It is not immediately obvious which is the most attractive"
A1 & A2 Most accurate, but too costly or impractical.
B1 Cannot detect the risk of concern
B2 We might as well toss a coin
B3 Would be considered scientifically unacceptable.

If the incidence was less, e.g. 8/100,000 rather than 8/10,000 then:

A1 & A2 need sample sizes of 270,000 and 150,000 respectively!
B1 – smallest detectable risk is 39 instead of 3.
B2 – false negative rate now > 60% (better off with a coin toss).
B3 – smallest detectable risk is 12.

As long as $\alpha < \beta$ we are perceived to be doing "better" science (because of fewer spurious effects) but we may be offering greater protection to harmless chemicals rather than the general population (i.e. we will fail to identify toxic chemicals and they will remain in use).

Suppose we test 2,400 substances and find:
960 (40%) are carcinogens
840 (36%) are benign
600 (24%) are uncertain

Epidemiological studies with α = 0.05, β = 0.20 and sample sizes are large enough we would produce the following errors:
192 (20%) of carcinogens would not be detected, and their use would continue;
43 (5%) of benign chemicals would be labelled toxic and subjected to unnecessary control

Presumably we consider that the two errors are not equivalent. One affects human health, the other affects company profits. There will be other circumstances in which the rank of the errors is reversed, e.g. if a potential life-saver is subject to a false negative result this much more serious than identifying a 'neutral' compound as a 'life-saver'.

The Alar example

In the late 1980s the EPA issued an alert on Alar (see http://www.cjr.org/year/96/5/alar.asp for a review of the way in which this was covered in the media and http://www.ewg.org/pub/home/reports/alar/alar.html for a review of the response of the Chemical Industry).

Alar is a systemic chemical used on fruits such as apples. Its primary benefit is to prevent apples falling off the tree too early and to delay the onset of rotting. Thus, its benefits are entirely financial. The EPA thought that Alar was one of the most potent carcinogens tested, long term exposure to Alar would increase the risk of cancer.

Note that much of this material is in the Power Notes – what is covered in more detail here are the relative consequences of the various mistakes.