1. Condom failure during sexual intercourse can lead to unplanned pregnancy and can facilitate the transmission of sexually transmitted diseases. The article "Fatigue testing in condoms" (*Polymer Testing*, **2009**; 567-571) examined the breaking strength of a random sample of $n = 20$ "Brand G" condoms. Individual condoms were placed on an apparatus that, when in motion, mimics sexual intercourse. Condom breaking strength was determined by measuring the number of cycles until breakage (the larger the number of cycles, the stronger the condom). Here were the data:

2226	2283	875	733	1390	1744	1174	1468	1229	1386
1843	914	1518	1288	1507	1321	819	1370	1543	2986

Here are the sample mean and sample standard deviation for this sample of condoms:

```
> mean(cycles)
[1] 1481
> sd(cycles)
[1] 543
```

(a) Describe what you think the population is. Report point estimates for the population mean μ and the population variance σ^2. State the units attached to your estimates.
(b) Calculate an estimate of the standard error of the sample mean. Describe in words what the standard error measures.
(c) An investigator constructed a qq plot for the Brand G condom data under a normal population distribution assumption (see next page). After looking at the plot, the investigator proclaims,

> *"These data are not normally distributed. Confidence intervals for the population mean will not be useful."*

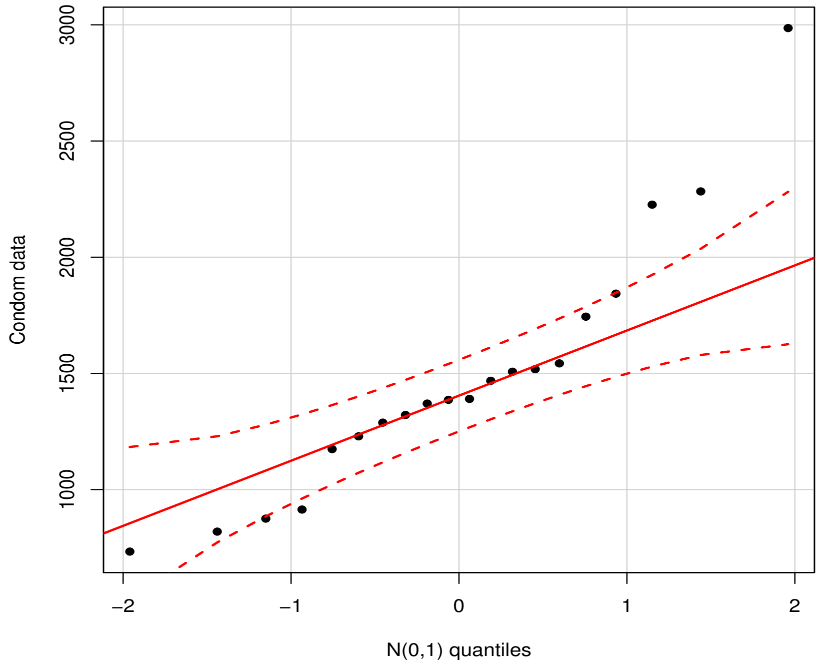
How would you respond to the investigator?

2. A random sample of female workers from a large oil company was observed; among the 541 workers sampled, 120 were classified as "obese." The results from this study were published in a 2008 article in *Annals of Epidemiology.*

(a) Based on the information in the sample, calculate a 95 percent confidence interval for the population proportion of obese female workers. Use

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

and $z_{\alpha/2} = z_{0.05/2} \approx 1.96$. Interpret precisely what your interval means.
(b) The article also summarized the results of an independent random sample of male



workers; among the 3612 workers sampled, 1084 were classified as obese. I calculated a 95 confidence interval for the population difference $p_1 - p_2$ (female minus male) using

$$(\widehat{p}_1 - \widehat{p}_2) \pm 1.96\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{541} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{3612}},$$

and I obtained $(-0.116, -0.040)$.

```
> proportion.diff.interval(120,541,1084,3612)
[1] -0.116 -0.040
```

Interpret precisely what this interval means. Then answer the following question: "Is there a difference between the population-level female and male obesity proportions?"
(c) Another characteristic recorded for each individual in the study was

$$Y = \text{the number of missed work days annually due to health reasons.}$$

I calculated the sample variances for each gender:

```
> var(female)
[1] 5.62
> var(male)
[1] 4.37
```

That is, the sample variance for the 541 female workers is $s_1^2 = 5.62$, and the sample variance for the 3612 male workers is $s_2^2 = 4.37$. With regard to the two genders, how do you think the population variances of Y compare? Describe an analysis you could do to answer this question. Provide as many details as possible, including concerns you might have with the analysis.

3. Hexavalent chromium has been identified as an inhalation carcinogen and an air toxin linked to various cancers. An article published recently in the *Journal of Air and Waste Management Association* gave the data below on both indoor (I) and outdoor (O) concentrations (nanograms/m^3) for a random sample of $n = 33$ houses in southwestern Ontario province.

House	1	2	3	4	5	6	7	8	9	10	11
Indoor	0.07	0.08	0.09	0.12	0.12	0.12	0.13	0.14	0.15	0.15	0.17
Outdoor	0.29	0.68	0.47	0.54	0.97	0.35	0.49	0.84	0.86	0.28	0.32
House	12	13	14	15	16	17	18	19	20	21	22
Indoor	0.17	0.18	0.18	0.18	0.18	0.19	0.20	0.22	0.22	0.23	0.23
Outdoor	0.32	1.55	0.66	0.29	0.21	1.02	1.59	0.90	0.52	0.12	0.54
House	23	24	25	26	27	28	29	30	31	32	33
Indoor	0.25	0.26	0.28	0.28	0.29	0.34	0.39	0.40	0.45	0.54	0.62
Outdoor	0.88	0.49	1.24	0.48	0.27	0.37	1.26	0.70	0.76	0.99	0.36

(a) Explain why this is a matched pairs study. Are the indoor and outdoor samples independent or dependent? Explain.
(b) I asked R for the data differences (indoor minus outdoor) and a 95 percent confidence interval for the population mean difference in concentrations $\mu_D = \mu_I - \mu_O$ acknowledging the matched pairs structure:

```
> diff = indoor-outdoor
> t.test(diff,conf.level=0.95)$conf.int
[1] -0.561 -0.286
```

Interpret precisely what this interval means. Then state what the interval suggests about a larger population of houses in this area.
(c) When I incorrectly analyzed the hexavalent chromium data as data from a two-independent sample design (Sample 1: Indoor; Sample 2: Outdoor), I got almost identical results when compared to the matched pairs analysis:

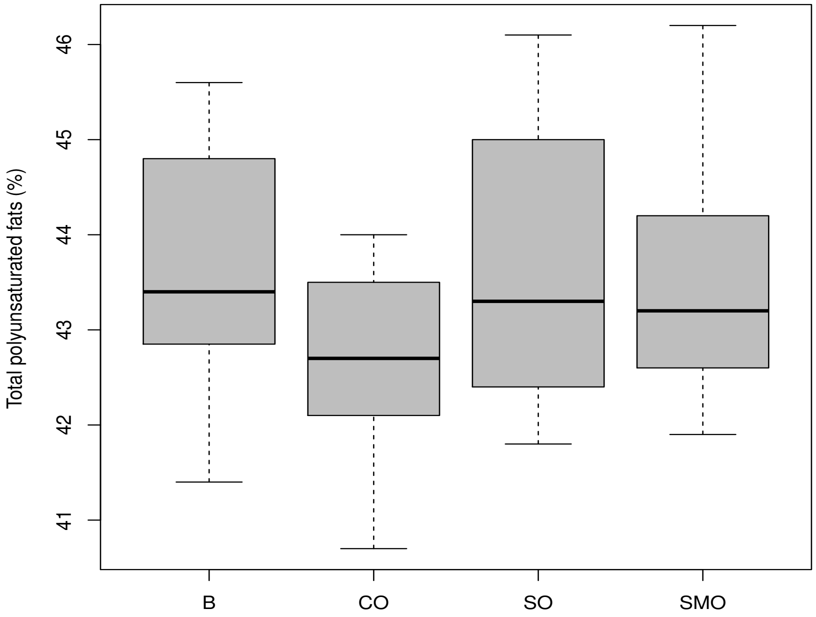
```
> t.test(indoor,outdoor,conf.level=0.95,var.equal=TRUE)$conf.int
[1] -0.563 -0.284
```

What do you think is going on here?

4. A recent article in *American Journal of Clinical Nutrition* describes an investigation where 52 preterm infants were randomly assigned to different nutrition regimens:

1. B: Breast milk ($n_1 = 8$ infants)
2. CO: Corn-oil-based formula ($n_2 = 13$ infants)
3. SO: Soy-oil-based formula ($n_3 = 17$ infants)
4. SMO: Soy-and-marine-oil-based formula ($n_4 = 14$ infants).

The response measured on each infant was Y, the total polyunsatruated fat (PUSF) percentage. The goal of the investigation was to determine if there were differences in the population mean PUSF percentage among the four groups. Here are side-by-side boxplots of the data:



(a) Looking at the boxplots only, an investigator exclaims,

> *"I knew the corn-oil-based formula was best—see, the CO group produces the smallest population median/mean PUSF percentage."*

How would you respond to the investigator?

(b) Another investigator performs an analysis of variance (ANOVA) with the data; here is the R output from the analysis (see next page):

```
> anova(lm(PUSF.percentage ~ nutrition.regimen))

                  Df Sum Sq Mean Sq F value Pr(>F)
nutrition.regimen  3  8.120  2.7068  1.6034 0.2009
Residuals         48 81.034  1.6882
```

Interpret these results. Note that the overall F statistic is ≈ 1.60 (p-value ≈ 0.20).
(c) What statistical assumptions are needed for the analysis in part (b)?
(d) If you did a follow-up analysis using Tukey confidence intervals for the $\binom{4}{2} = 6$ pairwise population mean differences, what would you expect to see?