Connections between methods:

power		$\delta = (t_{1-\alpha/2} + t_{1-\beta})\, se$	
ci width	$= 2\, t_{1-\alpha/2} \times se$	$\delta = (t_{1-\alpha/2} + t_{1-\alpha/2})\, se$	97.5% power, when $\alpha = 0.05$
half width	$= t_{1-\alpha/2} \times se$	$\delta = (t_{1-\alpha/2} + 0)\, se$	50% power
se	$= 1 \times se$	$t_{1-\beta} = 1 - t_{1-\alpha/2}$	$\approx$ 16% power

How good is the shifted-t approximation?
consider $\delta = 0.5$, $s = 0.9$, what is the power?

	n per group	
	$n = 20$	$n = 50$
shifted-t	24.2%	78.4%
non-central t	25.5%	78.5%

Application: 2 sample t-test, using pooled sd
need to know how se depends on n
gives huge flexibility in design problems
software mostly for specific (but commonly occurring) design questions
2 sample t-test: se $= s\sqrt{2/n}$

$$\delta = \left(t_{1-\alpha/2,\, df} + t_{power,\, df}\right) s\sqrt{2/n}$$

How big a true difference is needed to get 80% power when n = 20 per group and $s = 5$
df = (20 + 20 -2) = 38. $t_{0.975,\,38} = 2.024$, $t_{0.8,\,38} = 0.851$, se $= 5\sqrt{2/20} = 1.58$.
$\delta = (2.024 + 0.851) \times 1.58 = 4.54$
What n is needed for 80% power to detect a difference of 2 when $s = 14.8$ (WWE study)

$$n = 2\left(t_{1-\alpha/2,\, df} + t_{1-\beta,\, df}\right)^2 (s/\delta)^2$$

harder, because both df and se depend on n. Iterate to a solution.
I start with df = 60 ($n = 31$ per group) and find new n
$t_{0.975,60} = 2.000$, $t_{0.8,60} = 0.847$, $s/\delta = 7.4$,
$n = 2 \times 2.847^2 \times 7.4^2 = 888$ per group
new df = 1774, $t_{0.975,1774} = 1.961$, $t_{0.8,1774} = 0.842$,
$n = 2 \times 2.803^2 \times 7.4^2 = 860$ per group
This is similar to previous n, so stop. If not, continue.
Note: when df large, T quantiles are almost the same across a wide range of df
Can also ask what is the power if use $n = 400$, when $\delta = 2$, $s = 14.8$

$$t_{1-\beta} = \sqrt{n/2}\,(\delta/s) - t_{1-\alpha/2}$$

df = 798, $t_{0.975,798} = 1.963$, $t_{0.8,798} = 0.842$
$t_{1-\beta} = \sqrt{200}/7.4 - 1.963 = 1.911 - 1.963 = -0.052$
need to find P[$T_{798df} < -0.052$] = 48%

Notes:
power and sample size calculations depend on δ/s or s/δ
don't actually have to know each one separately
standardized effect sizes, δ/s, are common in social sciences
If you want to use a smaller type I error rate, e.g., $\alpha = 0.01$,
$t_{1-\alpha/2}$ goes up, sample size goes up, power goes down
Comparison for df = 1774:
$\alpha = 0.05$, $t_{1-\alpha/2} = t_{0.975} = 1.961$
$\alpha = 0.01$, $t_{1-\alpha^*/2} = t_{0.995} = 2.579$

Implementing a sample size determination:
What power value to use?

what difference matters, i.e., what is δ?

must have an estimate of σ, i.e. s - where to get this?

What can you do if n is too large?

Pairing:
Improve precision by reducing unwanted variability

"stat 587" analysis of paired data (we'll see more later)
compute difference between pairs

then sd of **differences**: here $s_d = 17.8$
se $= s_d/\sqrt{n}$, df $= n - 1$
Comparison: 2 sample difference $= 14.8\sqrt{2/n} = 20.9/\sqrt{n}$
paired se $(= 17.8/\sqrt{n}) <$ 2 sample se
All 3 statistical sample size methods depend on se
Smaller se for the same $n \Rightarrow$ higher precision, narrower ci, more power
Or, smaller n to acheive the same precision, same ci width, or same power

More than 2 groups: review of ANOVA, F tests
Example (will come back later): how tasty are 3 protein supplements?
old (current formulation), new/liquid, new/solid
Individuals will taste one supplement

response is a score: horrible to yummy
Q 1: any difference?
H0: all groups have the same mean
Ha: at least one group has a different mean
Answered by an F test
compares fit of H0 model to fit of Ha model
under usual ANOVA assumptions, fit measured by sum-of-squared errors
Can do power / sample size computations for the F test
Requires specifying all group means

"After the F test" aka "After the ANOVA" methods, multiple comparisons:
Q2: which pairs of groups are different?
multiple comparisons issues:
10 groups: 45 pairs.
if each pair tested at $\alpha = 0.05$, expect $45 \times 0.05 = 2.25$ "significant" results
make it "harder" to declare significance
all pairs: Tukey-Kramer adjusted p-values and simultaneous confidence intervals
Bonferroni: use $\alpha^* = \alpha/k$ instead of α
Simple to do sample size computations with Bonferroni

$$\delta = (t_{1-\alpha^*/2} + t_{1-\beta})\, se$$

Example: WWE study with 4 treatments, df $= 4(n-1)$
$\alpha^* = 0.05/4 = 0.0125$, $1 - \alpha^*/2 = 0.99375$
start with $n = 860$ per group, df = 3,436, $s = 14.8$, $\delta = 2$
$t_{0.99375} = 2.498$, $t_{0.8,3436} = 0.842$,
$n = 2\,(2.498 + 0.842)^2(14.8/2)^2 = 1,222$ per group!

"After the F test", linear contrasts
Q3: Before the data are collected, I have 2 specific questions:
a) What is the difference between new/liquid and new/solid?
b) What is the difference between old and new (average of new/liquid and new/solid)?
These are example of *a-priori* linear contrasts of means

$$\sum c_i\mu_i = c_o\mu_o + c_l\mu_l + c_s\mu_s, \text{ estimated by } \sum c_i\overline{Y}_i$$

Question	old	new/liquid	new/solid
a)	0	1	-1
b)	1	-0.5	-0.5

Distinguishing features of linear contrasts
1. Coefficients (the c's) sum to 0
2. Question(s) posed before seeing the data
3. Small number of questions(contrasts). Ideally no more than $k-1$

se of a linear contrast of k groups, assuming equal variances

$$\text{se} = s\sqrt{\sum_{i=1}^{k}\left(\frac{c_i^2}{n_i}\right)}$$

Examples:

a) new/liquid - new/solid: $\text{se} = s\sqrt{\frac{1^2}{n} + \frac{(-1)^2}{n}} = s\sqrt{2/n}$

b) old - new average: $\text{se} = s\sqrt{\frac{1^2}{n} + \frac{(-0.5)^2}{n} + \frac{(-0.5)^2}{n}} = s\sqrt{(1 + 0.25 + 0.25)/n} = s\sqrt{1.5/n}$

Orthogonal contrasts

property of a **pair** of contrasts

When two contrasts are orthogonal ⇒ estimates are independent

= Two unrelated quantities

Contrast 1: $\sum k_i\mu_i$, Contrast 2: $\sum l_i\mu_i$

are orthogonal when $\sum k_i\, l_i/n_i = 0$, where n_i is sample size for i'th group

when sample sizes are equal, i.e., all $n_i = n$, condition is $\sum k_i\, l_i = 0$

Example: contrast a and contrast b for the food supplement study

Contrast	old	liquid	solid
l - s	0	1	-1
o - (l+s)/2	1	-0.5	-0.5
product	0	-0.5	0.5

Sum of products = 0 + (-0.5) + 0.5 = 0

These are orthogonal

There are at most $k - 1$ orthogonal contrasts among k means

General practice is to **not** adjust for multiple comparisons

if a small (ca $k - 1$ contrasts) set of contrasts

especially if they are orthogonal

Implementing a sample size calculation for > 2 groups:

I ask:

What treatments?

Do the treatments suggest important questions (i.e., contrasts)?

What "after the F test" method?

Some commonly asked questions I get:

Is it appropriate to construct contrasts for all pairwise differences?

consultee wants to use contrast methodology to examine all pairs

because I said don't need to adjust contrasts

Question	old	new/liquid	new/solid
l-s	0	1	-1
o-l	1	-1	0
o-s	1	0	-1

No. Not a small number of contrasts, Not an orthogonal set
Need to use a multiple comparisons adjustment

Do sample sizes need to be equal?
Answer: No! Equal gives most precise difference, but close is almost as good
Consider a study that can use at most a total of $n_1 + n_2 = 20$ units, sd = 1
Compute se of the difference $= s\sqrt{1/n_1 + 1/n_2}$ for different choices of n_1 and n_2

n_1	n_2	se_{diff}	
10	10	0.402	
2	18	0.67	much larger
			similar se to $n_1 = 4$, $n_2 = 4$
9	11	0.404	essential the same as when equal
8	12	0.41	only slightly larger than equal

Exceptions to "recommend equal sample sizes":

1) When comparing many treatments back to a single reference (or control)
i.e., comparisons are Ref - Trt_1, Ref - Trt_2, $\cdots$, Ref - Trt_k
Reference mean gets used in many comparisons.
Treatment means only used once each. Increase sample size for the Reference, even if it means decreasing Treatment sample sizes slightly
e.g., Instead of $n_i = 10$ for Reference + 7 treatments, se $= 0.45\sigma$
Use $n_i = 17$ for Reference and 9 for each treatment, se $= 0.41\sigma$
Or $n_i = 24$ for Reference and 8 for each treatment, se $= 0.408\sigma$
Bigger improvement if more than 7 active treatments
Usually consider 2x - 4x sample size in the reference treatment
More than 4x in the reference has little impact on se_{diff}.
Then, se_{diff} depends mostly on se_{trt}

2) When Treatment and Control have different costs
Example: Evaluation of a potential remediation technique if another Chernobyl spreads radioactive isotopes across the landscape
Treating very large areas of the landscape (multiple square miles per plot) very expensive, ca \$100,000 per treated plot
Control plots quite cheap - only the cost of measurement, ca \$2,000 per plot

Original proposal (not mine) was 2 treated and 2 control plots
because "sample sizes need to be equal".
My proposal: 2 treated plots (all that could be afforded) and
8 control plots (cheap)
If you assume equal variances,
replication of the control plots informs variability between treated plots
much narrower confidence intervals for the difference