Planning an experiment

Overall considerations

- What treatments do you wish to compare?
  - To each other? To a control?
- What population should experimental units, e.g. patients, mice, or cells, represent?
- What is the primary outcome measure?
- What possible effects of treatment could be expected?
- What other factors might influence the outcome?
  E.g. time or other experimental conditions.
  How can these be controlled for?

Statistical considerations:

- Choice of design: Paired or unpaired? Single or multi-factorial?
- Choice of model to be used for analysis.
  - Can we assume a (log) normal distribution?
- Choice of adjustment for multiple comparisons - if any.
- Choice of sample size(s): Do power calculations!

And don’t forget: Use proper randomisation!
Outline

Basics of randomization

Basics of statistical power

How to do power calculations
  Power in 2x2 tables
  Two sample t-test
  Paired t-test
  Linear regression and correlation

Statistical design of experiments

Why randomize?

If treatment is properly randomized.
  ▶ No systematic difference between groups should occur other than due to treatment.
  ▶ The difference between outcomes can rightfully be interpreted as the causal effect of treatment.

In lack of randomisation (silly example).
A researcher does not know how to create a randomisation list and decides to give active treatment to all female mice and use untreated male mice as controls.
  ▶ We cannot interpret the difference as a treatment effect; it might as well be a gender difference!

Proper randomization

How do we randomize, e.g. 10 cell lines to two treatment groups?
  ▶ Assign treatment numbers 1, ..., 10
    (First half gets A and second half gets B).
  ▶ Do a random permutation: 2, 4, 8, 9, 7, 1, 10, 6, 3, 5.
    (Number 1, ..., 10 gets A, A, B, B, B, A, B, B, A, A.)

SAS program on slide XX.

Like an old fashion ballot:
  ▶ Draw treatments for each cell line
  ▶ from a bowl with five A’s and five B’s
  ▶ without replacement

Blocked randomization

In a paired design, randomisation must be performed "within".

Example: Randomize four treatments within 8 litters of ratpups.

<table>
<thead>
<tr>
<th>litter no.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>pup no. 1</td>
<td>D</td>
<td>C</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>pup no. 2</td>
<td>B</td>
<td>A</td>
<td>C</td>
<td>D</td>
<td>C</td>
<td>C</td>
<td>D</td>
<td>C</td>
</tr>
<tr>
<td>pup no. 3</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>D</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>pup no. 4</td>
<td>A</td>
<td>D</td>
<td>D</td>
<td>C</td>
<td>D</td>
<td>B</td>
<td>B</td>
<td>D</td>
</tr>
</tbody>
</table>

SAS program on slide XX.

Here pup number should be a completely random labeling not largest to smallest or firstborn to lastborn.
Simple randomization

Two treatments are compared in a randomized experiment.

▶ The experiment was planned for a power of 80%, which should be attained with 10 independent replications of each treatment.
▶ The experiment is carried out over several days.
▶ Every time a new subexperiment is started treatment is randomly selected with 50-50 probability.

When the study is complete, \( n_A = 14 \) and \( n_B = 6 \).

▶ Not quite as planned . . .
▶ Treatment is still randomized.
▶ But power may be reduced due to unequal sample sizes.

Pitfall: Unfortunate randomization

Two treatments are compared in a randomized experiment.

▶ The experiment was planned for a power of 80%, which should be attained with 10 independent replications of each treatment.
▶ The experiment is carried out over several days.
▶ Every time a new subexperiment is started treatment is randomly selected with 50-50 probability.
▶ When one of the treatments reaches target size of 10, the remaining subexperiments are carried out with the other treatment.

When the study is complete, \( n_A = 10 \) and \( n_B = 10 \).

▶ Exercise: So what is the problem?

Pitfall: uncontrollable circumstances

Realistic example (less silly, but same mistake):
In a method comparison study 12 operators measured the same 250 specimen in randomized order with three different methods.

Session 1: Measurements with method 1.
Session 2: Measurements with method 2.
Session 3: Measurements with method 3.

Best accuracy was found for method 1 followed by 2 and 3.
▶ BUT: What if there is a time effect?
▶ Are operators becoming sloppy because they get bored?

Could have controlled the ordering in a cross-over design:

<table>
<thead>
<tr>
<th>Order</th>
<th>Operators</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2-3</td>
<td>9, 12</td>
</tr>
<tr>
<td>1-3-2</td>
<td>3, 8</td>
</tr>
<tr>
<td>2-1-3</td>
<td>10, 7</td>
</tr>
<tr>
<td>2-3-1</td>
<td>6, 4</td>
</tr>
<tr>
<td>3-1-2</td>
<td>5, 2</td>
</tr>
<tr>
<td>3-2-1</td>
<td>11, 1</td>
</tr>
</tbody>
</table>

Making a randomization lists in SAS

Example: Randomize 12 cell lines to two treatments with six in each group.

```sas
data numlist;
do cellline=1 to 12;
  if (cellline <= 6) then treatment = 1;
  else treatment = 2;
output;
end;
run;
```

/* Fix random seed so that results can be reproduced */
```sas
proc plan seed=12345;
factors cellline=12;
output data=numlist out=randomlist;
run;
```
Making blocked randomization in SAS

Example: Randomize 12 litters of four to four different treatments.

```sas
data blocknum;
  do litter=1 to 12;
    do ratpup=1 to 4;
      treatment=ratpup;
      output;
    end;
  end;
run;
```

```sas
proc plan seed=12345;
  factors litter=12 ordered treatment=4;
  output data=blocknum out=randomblock;
run;
```

Outline

Basics of randomization

Basics of statistical power

How to do power calculations
  - Power in 2x2 tables
  - Two sample t-test
  - Paired t-test
  - Linear regression and correlation

Statistical design of experiments
Why do power/sample size calculations?

- In all experiments expenses increase with sample size. Even when using existing data to address a new research question, it is important to assess whether the researcher’s time could be spent better.
- It is unethical to treat more animals than necessary.
- You are required to do it for your protocol or grant application.
- Many journals demand that the power and sample size calculations are reported in the publication.

**Most importantly:** The sample size must be sufficiently large to draw conclusions on the scientific question of interest. I.e. sufficiently large not to overlook a clinically relevant difference.

Minimum relevant difference

Power calculations target a prespecified effect size, $\delta$, called the MInimum RElevant DIFference.

How do I determine the MIREDIF?

**Principled choice:** The smallest possible effect size such that treatment would be considered relevant in subject matter context.

**Practical choice:** The smallest possible effect size you think it would be a mistake to overlook.

**Common choice:** Though mostly wrong; The estimate I found in my pilot study / the literature.

**Note:** Choosing a sensible MIREDIF requires a careful subject matter discussion. Arguments should be stated when presenting the results of the power calculation.

Visualize the difference

A two sample t-test is being planned.

- What does the treatment effect of $\delta_0$ look like?

For $\delta_0 = 0.25, 0.50, 1, \text{ and } 2$ SD.

Statistical hypothesis testing

Hypothesis: no difference between the two treatments.

<table>
<thead>
<tr>
<th>Truth</th>
<th>Statement/Conclusion</th>
<th>$1-\alpha$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>Accept $H_0$</td>
<td>$\alpha$</td>
<td></td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>Reject $H_0$</td>
<td>$1-\alpha$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$1-\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>error of type II</td>
<td>power</td>
</tr>
</tbody>
</table>

**Table:** Classification of conclusions versus truth.

Usually $\alpha = 5\%$ and $1-\beta$ planned to be no less than 80%.
Clinical vs statistical significance

Clinical relevance depends on:
- The size of the treatment effect (i.e. the difference between the treatments).

Statistical significance depends on:
- The size of the treatment effect (i.e. the difference between the treatments).
- The random variation in the data.
- The experimental design.
- The sample size(s).
- (The significance level but this is usually fixed at $\alpha = 0.05$).

Reporting power and sample size calculations.

It is highly desirable for researchers to indicate how they obtained the sample sizes that they used in their studies because various assumptions must be made including parameter values which will almost always be unknown.

Moher, Dulberg, and Wells (JAMA 1994) found that sample size calculations were given only in 32% of studies that did not result in statistical significance. This could have occurred in some studies because the study was underpowered and/or unrealistic assumptions were made such as bad inputs for parameter values.

The reader of research papers needs enough information to be able to determine if the study was not well designed, so both the sample size and the manner in which it was determined should be provided in research articles.


Outline

Basics of randomization

Basics of statistical power

How to do power calculations
- Power in 2x2 tables
- Two sample t-test
- Paired t-test
- Linear regression and correlation

Statistical design of experiments

How to do a power calculation

Usually:
- Specify desired power $1 - \beta$, MIREDIF $\delta_0$, and model parameters. **Find the sample size** $N$.

Sometimes:
- Specify sample size $N$, MIREDIF $\delta_0$, and model parameters. **Find the power** $1 - \beta$.
- Specify sample size $N$, power $1 - \beta$, and model parameters. **Find the least detectable difference** $\delta_0$. 
Power in simple designs

Power and sample size using **pocket calculator formulae**:
- Two-sample comparisons for quantitative, binary, or survival outcome.
- Paired-sample comparison for quantitative or binary outcome.

Usually **only valid if sample sizes are sufficiently large**.

PROC POWER in SAS:
- Exact tests for two-sample comparisons and paired data.
- One-way ANOVA and select factorial designs.
- Linear regression and correlation.

★ E.g. formulae for t-test rely on the normal approximation, which is considered valid for degrees of freedom $\geq 60$.

Power in complex designs

**Use computer simulation**:
- Simulate a large number of experiments (datasets) with the desired structure.
- Analyze the data one at a time and count the number of times the test becomes significant.

**Note**: No explicit solution for sample size.
- Use **trial and error** to find sample size(s) that match target power.

Requirements:
- Good computing skills or money to hire a statistician
- Full specification of the complete data structure, i.e. joint distribution of outcome and covariates with reasonable guesses of all model parameters.

Power for testing a risk difference

**Use**: Compare a binary outcome between two groups.

**Model assumptions**:
- Observations are independent.
- Binary outcome with event probabilities $p_0$ and $p_1$.

**Specification for power calculation**:
- Type of test: Chi-square or Fisher’s exact test
- Event probabilities $p_0$ and $p_1$
- Desired power $1 - \beta$
- Significance level $\alpha$ (with $\alpha = 0.05$ as default).
- Ratio between group sizes (with $R = 1$ as default).

This could be e.g. $R = 1/2$ for a 1:2 distribution.

**Method 1: Explicit formula**

The test of $H_0: p_0 = p_1$ can be based on the Z-test statistic

$$z = \frac{\hat{p}_1 - \hat{p}_0}{s \cdot \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}$$

where $s = \sqrt{\frac{\hat{p}_0 + \hat{p}_1}{2} (1 - \hat{p}_0 + \hat{p}_1)}$.

The required total sample size to detect a risk difference of $\delta_0$ with power $1 - \beta$ at significance level $\alpha$ is:

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \bar{p}(1-\bar{p})(R+1)^2/R}{\delta_0^2}$$

where $\bar{p} = (p_0 + p_1)/2$ is the average event probability and the $z$’s are quantiles of the normal distribution, see NEXT SLIDE.
Quantiles of the normal distribution

Quantiles $z_{1-\beta}$ for the most common choices of desired power:

<table>
<thead>
<tr>
<th>$1 - \beta$</th>
<th>0.80</th>
<th>0.85</th>
<th>0.90</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_{1-\beta}$</td>
<td>0.8416</td>
<td>1.0364</td>
<td>1.2816</td>
<td>1.6449</td>
</tr>
</tbody>
</table>

The quantile $z_{1-\alpha/2}$ is usually 1.96 since $\alpha = 0.05$.

If multiple comparisons are planned, then $\alpha$ must be discounted according to the number of comparisons. E.g. with Bonferroni correction $\alpha$ should be replaced with $\alpha/k$ where $k$ is the number of planned tests to be performed.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0.05</th>
<th>0.05/2</th>
<th>0.05/3</th>
<th>0.05/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_{1-\alpha/2}$</td>
<td>1.9600</td>
<td>2.2414</td>
<td>2.3940</td>
<td>2.5758</td>
</tr>
</tbody>
</table>

Motivating example from introduction lecture

A new drug is tested in a small pilot experiment. At end of follow up we find 60% survivors with new drug vs 40% with old drug.

- A new study is planned to confirm the efficacy of the drug.

We can compute the sample sizes needed to detect a 20% risk reduction with 90% power:

- $\bar{p} = 0.50$ and $s^2 = \bar{p}(1 - \bar{p}) = 0.25$
- $N = \frac{(1.96+1.28)^2 \cdot 0.25 \cdot (1+1)^2}{(0.20)^2} = 196$

**BUT:** Is this what we are looking for?

Method 2: PROC POWER in SAS

A more adequate power calculation:

```
proc power;
twosamplefreq test=lrchi
groupproportions=(0.50 0.55)
ntotal=.
power=0.80
;run;
```

- `twosamplefreq` means test in a 2x2-table.
- `test=lrchi` is the likelihood ratio chi-square test.
- When `ntotal=` is missing and `power=0.80` is specified the calculation is solved for sample size.
- When `power=` and `ntotal` is specified the calculation is solved for power.

```
The POWER Procedure
Likelihood Ratio Chi-square Test for Two Proportions

Fixed Scenario Elements
Distribution Asymptotic normal
Method Normal approximation
Group 1 Proportion 0.5
Group 2 Proportion 0.55
Nominal Power 0.8
Number of Sides 2
Alpha 0.05
Group 1 Weight 1
Group 2 Weight 1

Computed N Total
Actual N
Power Total 0.800 3132

Output
```

Note: Binary data are not very informative. A huge sample size is needed to detect a small but substantial difference in risk!
What if we can’t get that large a sample?

Start a collaboration.
- Include more centres/research groups.

Be content with less than needed.
- What is the power for the attainable sample size?
- What MIREDIF can be detected with this and decent power?
- (Here the sometimes-calculations would be used).

Choose a different outcome measure or a different design, e.g.
- Continuous data are often more informative than binary.
- If possible do a paired comparison, since this has higher power.

Give up on the investigation!
- Instead of wasting time and money.

Skinny tables

If one or both of the probabilities \( p_0 \) or \( p_1 \) is small, then it likely that the results of the experiment will be a skinny table with expected values \( \leq 5 \) or even zero cell counts.

Check that your power calculation yield valid results by checking:
- \( np_i \geq 5 \) and \( n(1 - p_i) \geq 5 \) for \( i = 0, 1 \)

If this is not the case, plan an analysis based on Fisher’s exact test:

```plaintext
proc power;
twosamplefreq test=fisher
groupproportions=(0.20 0.02)
ntotal=.
power=0.80
;
run;
```

Output

The POWER Procedure
Fisher’s Exact Conditional Test for Two Proportions

<table>
<thead>
<tr>
<th>Fixed Scenario Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution</td>
</tr>
<tr>
<td>Method</td>
</tr>
<tr>
<td>Group 1 Proportion</td>
</tr>
<tr>
<td>Group 2 Proportion</td>
</tr>
<tr>
<td>Nominal Power</td>
</tr>
<tr>
<td>Number of Sides</td>
</tr>
<tr>
<td>Alpha</td>
</tr>
<tr>
<td>Group 1 Weight</td>
</tr>
<tr>
<td>Group 2 Weight</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Computed N Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual N</td>
</tr>
<tr>
<td>Power Total</td>
</tr>
<tr>
<td>0.801</td>
</tr>
<tr>
<td>102</td>
</tr>
</tbody>
</table>

The two-sample t-test

**Use:** Compare a quantitative outcome between two groups.

**Model assumptions:**
- The two samples are independent.
- The outcome is normally distributed within each group, \( N(\mu_0, \sigma_0^2) \) and \( N(\mu_1, \sigma_1^2) \), respectively.

**Specification for power calculation:**
- MIREDIF \( \delta_0 = \mu_1 - \mu_0 \)
- Variances \( \sigma_1^2 \) and \( \sigma_2^2 \)
- Desired power \( 1 - \beta \)
- (Significance level \( \alpha \))
- (Ratio between group sizes \( R \)).
Method 1: Explicit formula

The estimated effect size is \( \hat{\delta} = \bar{y}_1 - \bar{y}_0 \) and the t-test statistic:

\[
t = \frac{\bar{y}_1 - \bar{y}_0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}}
\]

Assuming identical variances \( \sigma_0^2 = \sigma_1^2 = \sigma^2 \), the required total sample size to detect a difference of \( \delta_0 \) with power \( 1 - \beta \) at significance level \( \alpha \) is given by:

\[
N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2(R + 1)^2}{\delta^2}.
\]

where the \( z \)'s are quantiles of the standard normal distribution, e.g. \( z_{1-\alpha/2} = 1.96 \) for \( \alpha = 0.05 \) and \( z_{1-\beta} = 0.8416 \) for \( 1 - \beta = 0.80 \) (see slide XX).

Example

We plan an experiment with a treatment and a control group of equal size (i.e. with \( R = 1 \)).

- The primary outcome \( y \) is presumably normally distributed
- We make a qualified guess that SD(\( y \)) \( \simeq 40 \)
  (using the literature and our clinical experience).

To detect a treatment effect of \( \delta = 15 \) with power \( 1 - \beta = 0.80 \) at least

\[
N = \frac{(1.96 + 0.84)^24(40)^2}{15^2} = 223
\]

subjects in total are needed, or \( n = 112 \) in each group.

Method 2: PROC POWER in SAS

proc power;
twosamplemeans test=diff
  meandiff=15
  stddev=40
  ntotal=.
  power=0.80
run;

The POWER Procedure
Two-Sample t Test for Mean Difference

---

Computed N Total

Actual Power Total
0.801 226

Power of the one-sample or paired t-test

Use: Compare a the mean of a quantitative outcome to a specific value, e.g. test that the mean difference in observation pairs is zero (this is the paired t-test).

Model assumptions:

- The observations are normally distributed \( N(\mu, \sigma^2) \) or the differences are normally distributed \( N(\mu_d, \sigma^2_d) \).

Specification for power calculation:

- MIREDIF \( \delta_0 = \mu - \mu_0 \) or \( \delta_0 = \mu_d \)
- Variance \( \sigma^2 \) or \( \sigma^2_d \)
- Desired power \( 1 - \beta \)
- Significance level \( \alpha \).
Method 1: Explicit formula

The test of $H_0 : \mu = \mu_0$ is based on the $t$-test statistic:

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

The required number of observations (or pairs) to detect a difference of $\delta_0$ with power $1 - \beta$ at significance level $\alpha$ is:

$$N = \left( z_{1-\alpha/2} + z_{1-\beta} \right)^2 \frac{\sigma^2}{\delta_0^2}$$

where the $z$'s are quantiles of the standard normal distribution (slide XX).

Method 2: PROC POWER one-sample $t$-test

```
proc power;
onesamplemeans test=t
mean=-0.1054
stddev=0.1548
ntotal=.
power=0.85;
run;
```

Output: Exercise.

Method 3: PROC POWER paired $t$-test

Let SAS compute the SD of the difference:

```
proc power;
pairedmeans test=diff
meandiff=-0.1054
pairedstddevs=(0.173087 0.173087)
corr=0.80
npairs=.
power=0.85;
run;
```

Output: Exercise.

What is the SD of the difference?

Usually we have an idea of what the standard deviation of a single measurement, $Y_0$ or $Y_1$ is.

- But what about the difference $d = Y_1 - Y_0$?

It depends on the correlation $\rho = \text{Corr}(Y_0, Y_1)$.

- If $\sigma_0 \neq \sigma_1$ then $\sigma_d^2 = \sigma_0^2 + \sigma_1^2 - 2\rho \sigma_0 \sigma_1$.
- If $\sigma_0 = \sigma_1$ then $\sigma_d^2 = 2 \cdot (1 - \rho) \sigma_0^2$

Use this to make a qualified guess of $\sigma_d$.

- A very conservative guess would be $\sqrt{2} \cdot \sigma_0$ (a positive correlation is expected, but nothing more is known).
Power in linear regression

Interest is the linear model

\[ Y_i = a + bX_i + \varepsilon, \quad i = 1, \ldots, n \]

- Observations are independent
- Error terms are normal \( \varepsilon \sim N(0, \sigma^2) \) with homogeneous variances.

We want to test \( H_0 : b = 0 \).

Two different situations:
- \( X \)s are planned (e.g. doses).
- \( X \)s are sampled together with \( Y \)s (e.g. two biomarkers)

Power calculation

The test for \( H : b = 0 \) is an approximate t-test based on

\[ t = \frac{\hat{b}}{s_{\varepsilon} \sqrt{\frac{1}{N-1}}} = \frac{\hat{b} \cdot S_X}{s_{\varepsilon} \sqrt{\frac{1}{N-1}}} \]

on \( N - 2 \) degrees of freedom where \( s_{\varepsilon} \) is the estimated residual standard deviation and \( N = nk \) is the total sample size.

Hence, do a power calculation for a one-sample t-test
- \( \text{MIREDIF} \delta_0 = b_0 \cdot S_X \).
- Make a qualified guess of the residual standard deviation.

Finally add one to the suggested sample size (since the one-sample t-test has a \( \sqrt{N} \) in the denominator and \( N - 1 \) degrees of freedom) and choose \( N' \geq N \) such that \( N' = nk \).

Linear regression with planned \( x \)'s

Use: Test if there is a linear relation between \( X \) and \( Y \).

Model assumptions:
- Planned \( X_1, \ldots, X_k \) with \( n \) observations each.
- Observations are independent and normally distributed \( Y_{ij} \sim N(a + b \cdot x_i, \sigma^2) \), for \( i = 1, \ldots, k \), \( j = 1, \ldots, n \).

Must specify:
- \( \text{MIREDIF} \delta_0 = b_0 \cdot S_X \), where \( S_X = \sqrt{\frac{1}{k} \sum_{i=1}^{k} (x_j - \bar{x})^2} \)
- residual variance \( \sigma^2_{\varepsilon} \)
- desired power \( 1 - \beta \)
- significance level \( \alpha \).

Linear regression with sampled \( x \)'s

Use: Test if there is a linear relation between \( X \) and \( Y \).

Model assumptions:
- Observations are independent
- Observations follow a 2D normal distribution

\[
\begin{pmatrix}
X \\
Y
\end{pmatrix} \sim N\left(\begin{pmatrix}
\mu_X \\
\mu_Y
\end{pmatrix}, \begin{pmatrix}
\sigma^2_X & \rho \sigma_X \sigma_Y \\
\rho \sigma_X \sigma_Y & \sigma^2_Y
\end{pmatrix}\right)
\]

Must specify:
- \( \text{MIREDIF} \delta_0 = \frac{\sigma_X}{\sigma_Y} \cdot b_0 \).
- desired power \( 1 - \beta \)
- significance level \( \alpha \).
Regression and correlation

The 2D normal model implies a linear relation $Y = a + bX + \varepsilon$.

$\varepsilon$ is independent of $x$ and $\sigma^2_{\varepsilon} = (1 - \rho^2)\sigma^2_Y$.

Power calculation

Test of $H : b = 0$ is equivalent to the test of $H : \rho = 0$.

The MIREDIF for $b$ translates to a MIREDIF for $\rho$ via

$$\rho = \frac{\sigma_X}{\sigma_Y} \cdot b$$

using qualified guesses of $\sigma_X$ and $\sigma_Y$.

```plaintext
proc power;
onecorr test=pearson
corr = 0.35
ntotal = .
power = 0.85
;
run;
```

Outline

Basics of randomization

Basics of statistical power

How to do power calculations

- Power in 2x2 tables
- Two sample t-test
- Paired t-test
- Linear regression and correlation

Statistical design of experiments

Some statistical designs

- Multiple comparisons with a control
- Two-way ANOVA (the stratified design)
- Matched samples ANOVA (the randomized block design)
Multiple comparisons with a control

Idea: We do a \( k \)-in-one experiment re-using the control group. This saves time and money.

Optimal design:
- \( n_0 \) in control group, \( n_1 \) in each of \( k \) treatment groups,
- Estimated effect of treatment \( j \) is \( \bar{Y}_j - \bar{Y}_0 \) with
  \[
  \text{SE}(\bar{Y}_j - \bar{Y}_0) = \frac{s}{\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}}
  \]
- For a fixed total \( N (= n_0 + kn_1) \) this is minimized by
  \[
  n_0 = \frac{N}{1 + \sqrt{k}}, \quad n_1 = \frac{N}{k(1 + \sqrt{k})}
  \]
- The optimal ratio is \( R = n_1/n_0 = 1/\sqrt{k} \).

A note on unequal group sizes

If just a single two-sample t-test is planned.
- Assuming that variances are equal, then optimal power is obtained with equal sample sizes (since \( (R+1)^2/R \) is smallest when \( R = 1 \)).
- Otherwise it is more efficient to have a larger sample in the group which has the larger variance.

Other reasons for choosing an uneven distribution on groups:
- One treatment is much more expensive than the other.
- We plan multiple comparisons with the same control group.
- (It is unethical to treat many patients with placebo if prior studies indicate a treatment effect.)

Power for multiple comparisons

We plan an experiment comparing 4 treatments with a control.
- Power as for the two-sample t-test
- Need to specify MIREDIF (15) and SD of outcome (30).
- Optimal ratio of treated to controls: \( R = 1/\sqrt{4} = 1/2 \).
- Bonferroni adjustment: \( \alpha = 0.05/4 = 0.0125 \).

```plaintext
proc power;
twosamplemeans test=diff
meandiff=15
stddev=30
groupweights=(1 2)
ntotal=.
alpha=0.0125
power=0.85;
run;
```

Output

The POWER Procedure
Two-Sample t Test for Mean Difference

Fixed Scenario Elements

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Exact</td>
</tr>
<tr>
<td>Alpha</td>
<td>0.0125</td>
</tr>
<tr>
<td>Mean Difference</td>
<td>15</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>30</td>
</tr>
<tr>
<td>Group 1 Weight</td>
<td>1</td>
</tr>
<tr>
<td>Group 2 Weight</td>
<td>2</td>
</tr>
<tr>
<td>Nominal Power</td>
<td>0.85</td>
</tr>
<tr>
<td>Number of Sides</td>
<td>2</td>
</tr>
<tr>
<td>Null Difference</td>
<td>0</td>
</tr>
</tbody>
</table>

Computed N Total

<table>
<thead>
<tr>
<th>Actual N</th>
<th>Power Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.850</td>
<td>228</td>
</tr>
</tbody>
</table>

This implies: \( n_0 = 152 \) and \( n_1 = 76 \). We thus need \( N = 456 \) in total.
Power for independent t-tests

If we had equal size groups:

```plaintext
proc power;
   twosamplemeans test=diff
      meandiff=15
      stddev=30
      ntotal=.
      alpha=0.0125
      power=0.85
   ;
run;
```

The POWER Procedure
Two-Sample t Test for Mean Difference
Computed N Total
Actual N  Power Total
0.852 204

We need $N = 510$ in total to get the same power with equal size groups.

Stratification

**Dilemma:** A higher power is obtained when the test population is very homogeneous, i.e. has a small variance.

- Use rats or mice of similar breed and gender.
- Use cells from similar tissue or tumor type.

**BUT:** The conclusion from the experiment are only valid for the population that was sampled from, e.g. only males.

**Solution:** Plan an experiment with e.g. equal numbers of males and females randomized to treatment.

- We say that treatment is randomized within *gender-strata*.

Two-way ANOVA

Plan a **balanced design** with:

- Assume $m$ strata and $k$ treatments.
- Assume $n_j$ observations with treatment $j$ in each stratum.
- Then we have $N = m \cdot \sum_{j=1}^{k} n_j$ observations in total.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strata</td>
<td>$m-1$</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Treatment</td>
<td>$k-1$</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Error</td>
<td>$N-m-k+1$</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Total</td>
<td>$N-1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How precise is our estimate of the residual variance?

- $s^2 = \text{MS(Err)}$ on $N-m-k+1$ degrees of freedom.

Test of a treatment effect

Tests of treatment effect, e.g. 1 vs 2, is based on treatment averages:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

that has a t-distribution on $N-m-k+1$ degrees of freedom.

Do the power calculation as for a two-sample t-test.

- **MIREDIF:** $\delta_0$ for the treatment effect
- **Qualified guess of residual standard deviation.**
- **$R = n_1 / n_2$.**
- Find required sample size $N$ for a given power.
- Add $m-3$ to **compensate for degrees of freedom lost.**
- Choose smallest possible $N' = m \cdot (n_1 + n_2)$. 
What about interaction?

We are interested in finding an overall treatment effect assuming this to be the same in all strata.

Hence the sample size is planned for a two-way ANOVA without interaction.

With replicates in each cell, we can still test the interaction, but the test is typically performed as a goodness of fit test rather than planned as a primary aim of the investigation.

What factors should be controlled for?

Only factors that have a marked effect on the outcome should be controlled for. Other factors with less impact can be thought of as part of the random variation in the population.

We have a trade-off:

Many strata: Small residual variance but many extra observations needed to estimate the strata-differences.

Few strata: Larger residual variance but only few extra observations needed to estimate the strata-differences.

Matched samples

(aka The randomized block design in the statistical literature).

Idea: Apply all of the treatments on each experimental unit (cell line or litter). Then treatment effects are evaluated on paired data which gives a higher power with fewer experimental units since variation is reduced when units act as stratas.

We have a variance component model, e.g.

\[ Y_{ij} = \mu_i + a_j + \epsilon_{ij} \]

- \( \mu_i \) mean with i’th treatment
- \( a_j \sim N(0, \omega^2_B) \) deviations between litters.
- \( \epsilon_{ij} \sim N(0, \sigma^2_W) \) deviations within litters.

Matched samples ANOVA

Assume \( m \) litters and \( k \) treatments.

Then we have \( N = mk \) observations in total.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>m-1</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Treatment</td>
<td>k-1</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Residual</td>
<td>N-m-k+1</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Total</td>
<td>N-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Estimated variance components:

- within litters: \( \sigma^2_W = \text{MS(Residual)} \)
- between litters: \( \omega^2_B = \text{MS(Subject)} - \frac{\text{MS(Residual)}}{k} \)
Test of a treatment effect

Tests of a treatment effects, e.g. 1 vs 2, are based on the t-test:

\[ t = \frac{\bar{d}_{1 vs 2}}{\text{SD}(d_{1 vs 2})} = \frac{\bar{Y}_1 - \bar{Y}_2}{\hat{\sigma}_W \cdot \frac{2}{\sqrt{m}}} \]

on \( N - m - k + 1 \) degrees of freedom.

Explanation for the standard error:

- \( \bar{Y}_1 - \bar{Y}_2 = \bar{\varepsilon}_1 - \bar{\varepsilon}_2 \), which has variance \( 2 \cdot \frac{\sigma^2_W}{m} \)

**NOTE:** This is just a paired t-test.

- Total variance \( \text{Var}(Y) = \omega^2_B + \sigma^2_W \).
- Correlation between replicate measurements is the intra class correlation (ICC) \( \rho = \omega^2_B / (\omega^2_B + \sigma^2_W) \).
- Hence \( \sigma^2_W = (1 - \rho) \cdot \text{Var}(Y) \).

---

Test of a treatment effect

Do the power calculation as for a paired t-test.

- **MIREDIF:** \( \delta_0 \) for the treatment effect
- **Qualified guess of the standard deviation** within (or of the total variance and the ICC).
- Can adjust \( \alpha \) if multiple comparisons are planned.
- Find required **number of litters** \( m \) for a given power.

**Note:** No need to adjust for degrees of freedom lost when estimating the effect of several treatments at a time.

- Degrees of freedom \( N - m - k + 1 = (m - 1)(k - 1) \) are always larger than \( m - 1 \) when \( k > 2 \).