Power analysis and sample size calculation

Eirik Skogvoll, MD, PhD
Professor, Faculty of Medicine
Consultant, Dept. of Anaesthesiology and Emergency medicine
The hypothesis test framework

### Rosner tbl. 7.1

<table>
<thead>
<tr>
<th>Conclusion</th>
<th>$H_0$ correct</th>
<th>$H_0$ wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ accepted</td>
<td>$1 - \alpha$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>$H_0$ rejected</td>
<td>$\alpha$</td>
<td>$1 - \beta$</td>
</tr>
</tbody>
</table>

- **Type I error**: rejecting $H_0$ if $H_0$ is true
- **Type II error**: accepting $H_0$ if $H_0$ is wrong

- **P (type I error)**
- **P (type II error)**
- **Power**

Truth

- **Power**
Hypothesis

A statement about one or more population parameters, e.g.

- $\mu \leq 3$ (Expected value, or “true mean”)
- $\mu_1 = \mu_2$ (Two expected values)
- $\sigma^2 = 14$ (variance)
- $\pi = 0.4$ (probability of a “success”, in a binomial trial)
- $\pi_1 = \pi_2$ (two probabilities)
- $\psi = 1$ (odds ratio, from a contingency table)
- $\beta_1 = 0$ (regression coefficient)
- $\rho > 0$ (correlation coefficient)
The sample size is a compromise…

• Significance level
• Required or optimal power
• Available resources
  – Money
  – Time
  – Practical issues of recruiting patients or subjects
• The problem under investigation
Power

• The probability that some test will detect a difference, if it exists
• With low power, even large differences may go unnoticed:

Absence of evidence is not evidence of absence!

• Power is usually set to 0.8 – 0.9 (80-90 %)
• An infinitely small difference is impossible to detect
• The researcher must specify the smallest difference of (clinical) interest
• Power calculations requires knowledge of the background variability (usually the standard deviation)
Power or significance?

- We wish to reduce the probability of both a type I error ($\alpha$) and a type II error ($\beta$), because...
  - A small $\alpha$ means that it is difficult to reject $H_0$
  - A small $\beta$ means that it is easier to reject $H_0$ (and accept $H_1$)
  - But minimizing $\alpha$ and $\beta$ at the same time is complicated, because $\alpha$ increases as $\beta$ decreases, and vice versa

- Strategy:
  - Keep $\alpha$ at a comfortable level (0.10, 0.05, 0.01); the maximum "acceptable" probability of making a type I error; i.e. rejecting $H_0$ when it is true
  - Then choose a test that minimizes $\beta$, i.e. maximizes power ($=1 - \beta$).
    Beware of $H_1$: tests may have different properties!
Power (Rosner 7.5)

One-sided alternative hypothesis

H₀: μ = μ₀ \hspace{1cm} \text{vs.} \hspace{1cm} H₁: μ = μ₁ < μ₀

H₀ is rejected if \( z_{obs} < z_α \) and accepted if \( z_{obs} ≥ z_α \).

The decision does not depend on \( μ₁ \) as long \( μ₁ < μ₀ \). \hspace{1cm} \text{(Rosner Fig. 7.5)}

Power, Pr (reject \( H₀ \)|\( H₀ \) wrong) = 1 – Pr(type II error) = 1 – \( β \), however, depends on \( μ₁ \).
Rosner
Fig. 7.5

Distribution of \( \bar{X} \)
under \( H_1 \)
\( N(\mu_1, \sigma^2/n) \)

Distribution of \( \bar{X} \)
under \( H_0 \)
\( N(\mu_0, \sigma^2/n) \)

\[
H_1: \mu = \mu_1 < \mu_0 \quad \text{vs.} \quad H_0: \mu = \mu_0
\]
Under $H_1$, $\bar{X} \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)$

We reject $H_0$ if the observed statistic $z$ is less than critical value $z_{\alpha}$, i.e. if $z_{obs} < z_{\alpha}$.

Using the corresponding value for $\bar{X}$:

$$
\Pr(Z < z_{\alpha}) = \Pr\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha}\right) = \Pr\left(\bar{X} < \mu_0 + z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}\right)
$$

$(z_{\alpha} < 0)$
... remember that \( Z = \frac{\bar{X} - \mu_1}{\sigma} \leftrightarrow \bar{X} = Z \cdot \frac{\sigma}{\sqrt{n}} + \mu_1 \)

thus, \( \text{Power} = P\left(\bar{X} < \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = P\left(Z \frac{\sigma}{\sqrt{n}} + \mu_1 < \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) \)
\[
\begin{align*}
&= P \left( Z + \frac{\mu_1}{\sigma} < \frac{\mu_0}{\sigma} + \frac{z_\alpha}{\sqrt{n}} \right) = P \left( Z < \frac{\mu_0}{\sigma} - \frac{\mu_1}{\sigma} + \frac{z_\alpha}{\sqrt{n}} \right) \\
&= P \left( Z < \frac{\mu_0 - \mu_1}{\sigma} + z_\alpha \right) = P \left( Z < z_\alpha + \frac{\mu_0 - \mu_1}{\sigma} \right) = P \left( Z < z_\alpha + \frac{(\mu_0 - \mu_1)}{\sigma} \cdot \sqrt{n} \right) \\
&= \Phi \left( z_\alpha + \frac{(\mu_0 - \mu_1)}{\sigma} \cdot \sqrt{n} \right) \quad \text{... because } P(Z < z) = \Phi(z) \quad \text{... table 3a}
\end{align*}
\]
One-sided alternatives, in general:

\[ H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu = \mu_1 \]

\[
\text{Power} = \Phi\left(z_\alpha + \frac{|\mu_0 - \mu_1|}{\sigma} \cdot \sqrt{n}\right) = \Phi\left(-z_{1-\alpha} + \frac{|\mu_0 - \mu_1|}{\sigma} \cdot \sqrt{n}\right)
\]

(Rosner, Eq. 7.19) (remember that \(z_\alpha = -z_{1-\alpha}\))
Factors influencing power:

- Power is reduced if $\alpha$ is reduced ($z_\alpha$ becomes more negative)
- Power increases with increasing $|\mu_0 - \mu_1|$
- Power is reduced with increasing $\sigma$
- Power is increased with increasing $n$
Example (Rosner expl. 7.26)
Birth weight.

H₀: μ₀ = 120 oz, vs. H₁: μ₁ = 115 oz, 
α = 0.05, σ = 24, n = 100.

Power = Pr(rejecting H₀|H₀ wrong) =
Φ\left( z_{0.05} + \frac{120 - 115}{24} \cdot \sqrt{100} \right) = \Phi(-1.645 + \frac{5 \cdot 10}{24}) = \Phi(0.438)
= Pr(Z < 0.438) = 0.66
→ Pr(type II error) = Pr(accept H₀|H₀ wrong) = β = 1 - 0.669 = 0.331
Example (Rosner expl. 7.29)
Birth weight.

\[ H_0: \mu_0 = 120 \text{ oz}, \text{ vs. } H_1: \mu_1 = 110 \text{ oz}, \]
\[ \alpha = 0.05, \sigma = 24, n = 100. \]

\[ \Phi \left( z_{0.05} + \frac{120 - 110}{24} \cdot \sqrt{100} \right) = \Phi \left( -1.645 + \frac{10 \cdot 10}{24} \right) = \Phi(2.52) \]

= \Pr(Z < 2.52) = 0.99
Power for a one sample test, two sided alternative (in general):

\( H_0: \mu = \mu_0 \) vs. \( H_1: \mu = \mu_1 \neq \mu_0 \)

\[
\text{Power} \approx \Phi \left( -z_{\frac{1-\alpha}{2}} + \frac{|\mu_0 - \mu_1|}{\sigma} \cdot \sqrt{n} \right) \quad \text{(Rosner Eq. 7.21)}
\]
From Power to Sample size…

From Rosner Eq. 7.21:

\[
\text{Power} = \Phi \left( -z_{1-\alpha/2} + \frac{|\mu_0 - \mu_1| \cdot \sqrt{n}}{\sigma} \right) = 1 - \beta \quad \text{...now we set ("fix") } 1 - \beta
\]

\[
\Phi^{-1} \left( \Phi \left( -z_{1-\alpha/2} + \frac{|\mu_0 - \mu_1| \cdot \sqrt{n}}{\sigma} \right) \right) = \Phi^{-1}(1 - \beta)
\]

\[-z_{1-\alpha/2} + \frac{|\mu_0 - \mu_1| \cdot \sqrt{n}}{\sigma} = z_{1-\beta} \iff \frac{|\mu_0 - \mu_1| \cdot \sqrt{n}}{\sigma} = z_{1-\beta} + z_{1-\alpha/2}
\]

\[\iff |\mu_0 - \mu_1| \cdot \sqrt{n} = \sigma \cdot (z_{1-\beta} + z_{1-\alpha/2}) \iff \sqrt{n} = \frac{\sigma \cdot (z_{1-\beta} + z_{1-\alpha/2})}{|\mu_0 - \mu_1|}\]
Sample size for one sample test, two sided alternative:

\[ H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu = \mu_1 \neq \mu_0 \]

\[ n = \frac{\sigma^2 \left( z_{1-\beta} + z_{1-\alpha/2} \right)^2}{(\mu_0 - \mu_1)^2} \quad \text{(Rosner Eq. 7.28)} \]
Two sample test: sample size (Rosner 8.10)

Given $X_1 \sim \text{N}(\mu_1, \sigma^2_1)$ and $X_2 \sim \text{N}(\mu_2, \sigma^2_2)$

$H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$ (two sided test),
significance level $\alpha$, and Power $1 - \beta$.

\[
 n = \frac{(\sigma_1^2 + \sigma_2^2)(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{(\mu_2 - \mu_1)^2} \quad \text{.... in each group} \quad (\text{Rosner Eq. 8.26})
\]
Rules of thumb

Lehr’s "quick" equations for two-sample comparison:

- 80% power, $\alpha = 0.05$ (two sided), $\Delta = $ standardized difference, $m = $ no. in each group:

  $$m = \frac{16}{\Delta^2}$$

- 90% power, $\alpha = 0.05$ (two sided), $\Delta = $ standardized difference, $m = $ no. in each group

  $$m = \frac{21}{\Delta^2}$$
The standardized difference

Altman 1991 (figure 15.2)

Need to specify the "standardized difference":
\[ \Delta = \frac{|\mu_2 - \mu_1|}{\sigma} \]

… the least, clinically important difference expressed as standard deviations. May vary 0.1 – 1.0, where \( \Delta = 0.1 - 0.2 \) constitutes a "small" difference, while \( \Delta = 0.8 \) is "large".
What is the background variability ("noise")?

- From other studies
- Pilot data
- From the range of possible values ("quick and dirty"): 

\[
SD \approx \frac{Range}{4}
\]

- This results stems from the basic properties of the Normal (Gaussian) distribution, in which approx. 95 % of the population are found within ± 2SD of the mean.
Example

- 80% power, $\alpha = 0.05$ (two sided), $\Delta = \text{standardised difference} = 0.5$

$$m = \frac{16}{\Delta^2} = \frac{16}{0.25} = 4 \times 16 = 64$$

Total sample size $= 64 \times 2 = 128$
Nomogram (Altman 1991)

Standardised difference

(total sample size)
Example

80% power, \( \alpha = 0.05 \) (two sided), \( \Delta = 0.5 \)

\( N \) (total) = 125

Standardised difference
Factors influencing power (1-β)

- Power decreases if $\alpha$ is reduced ($z_\alpha$ becomes more negative)
- Power increases with increasing $|\mu_0 - \mu_1|$
- Power decreases with increasing $\sigma$
- Power increases with increasing sample size "n"
Factors leading to increased sample size ($n$)

- Large standard deviation
- Strict requirement for $\alpha$ (low level, 0.01-0.02)
- Strict requirement for power ($1 - \beta$) (high level: 0.90-0.95)
- Low absolute difference between parameters under $H_1$ and $H_0$ ($|\mu_1 - \mu_0|$)
Confidence intervals

• Most journals want CI’s reported along with (or instead of) p-values
• The expected length of a confidence interval is determined by the sample size:

\[ l = 2 \cdot t_{n-1,1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \]

...where

\( t = \) relevant quantile of the t distribution with \( n-1 \) degrees of freedom

\( s = \) sample standard deviation

\( n = \) no. of observations
How to make efficient use of your subjects

- Re-use subjects if possible, using a repeated measurements design. Intra-subject variance will be reduced, as each subject acts as his own control.
- Enter more than one experimental factor, using a factorial design. Each subject counts twice (or thrice…)! 
- Refine your measurement methods to reduce residual error
- Never, never, never…! dichotomize a continuous outcome variable to a binary one (i.e. 1 or 0)! This has the largest variance.
Software for sample size calculations

• Machin et al. "Sample Size Software for Clinical studies" 3.ed, 2009 -- but you need to set dot (.) as decimal separator 😞
• Sample Power
• R
• StatExact (for special procedures)
References

- Bland JM. The tyranny of power: is there a better way to calculate sample size? BMJ. 2009 October 6, 2009;339(oct06_3):b3985-.