6 Sample Size Calculations

A **major** responsibility of a statistician: sample size calculation.

Hypothesis Testing: compare treatment 1 (new treatment) to treatment 2 (standard treatment); Assume continuous endpoints.

- $\Delta =$ treatment effect, parameter of interest. For example $\Delta = \mu_1 - \mu_2$.
- nuisance parameters: $\theta = (\mu_2, \sigma^2)$
- $H_0 : \Delta \leq 0$: stay with the standard treatment
- $H_A : \Delta > 0$: switch to the new treatment
- Data $(z_1, \ldots, z_n)$, $z_i =$ realization of $Z_i = (Y_i, A_i)$, where

$$Y_i | A_i = 1 \sim N(\mu_1, \sigma^2), \quad Y_i | A_i = 2 \sim N(\mu_2, \sigma^2)$$

- Construct a test statistic

$$T = T_n(z_1, ..., z_n).$$

For example, two-sample t-test statistic:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{n_1^{-1} + n_2^{-1}}}.$$

The statistic T should be constructed in such a way that

1. Larger values of T are evidence against H_0 (in favor of H_A).
2. The distribution of T can be (approximately) evaluated at the **border** between H_0 and H_A; i.e. at $\Delta = 0$.

- Given observed test statistic T_{obs}, p-value for testing $H_0 : \Delta \leq 0$ vs. $H_A : \Delta > 0$ is calculated as

$$P_{\Delta=0}(T \geq T_{obs}).$$

For given type I error prob α (usually 0.025 or 0.05), reject H_0 if

p-value $< \alpha$.

- **Note**:

  1. Most often, $P_\Delta(T \geq x)$ increases as Δ increases, for all x. So

  $$P_{\Delta=0}[T \geq T_{obs}] \leq \alpha \Longrightarrow P_\Delta[T \geq T_{obs}] \leq \alpha \text{ for all } \Delta \in H_0.$$

  2. The distribution of T at $\Delta = 0$ is known and

  $$T \overset{(\Delta=0)}{\sim} N(0,1).$$

  $$\text{P} - \text{value} \leq \alpha \Longleftrightarrow T_{obs} \geq z_\alpha.$$

  3. For the two-sample t-test statistic, we have

  $$T \overset{(\Delta=0)}{\sim} t_{n-2} \approx N(0,1).$$

- **Remark on two-sided tests**:

$$H_0 : \Delta = 0 \quad vs. \quad \Delta \neq 0$$

- ⋆ Reject H_0 if $|T|$ is large
- ⋆ P-value

$$P_{\Delta=0}(|T| \geq |T_{obs}|) = P_{\Delta=0}[T \geq |T_{obs}|] + P_{\Delta=0}[T \leq -|T_{obs}|].$$

- ⋆ For given α, reject H_0 if $|T| \geq z_{\alpha/2}$.

- **rejection region**

  For one-sided level α tests, the rejection region is

$$\{(z_1, \ldots, z_n) : T_n(z_1, \ldots, z_n) \geq \mathcal{Z}_\alpha\},$$

  and for two-sided level α tests, the rejection region is

$$\{(z_1, \ldots, z_n) : |T_n(z_1, \ldots, z_n)| \geq \mathcal{Z}_{\alpha/2}\}.$$

- **Power**

  For one-sides tests:

$$P_{\Delta=\Delta_A}[T \geq z_\alpha], \quad \text{for } \Delta_A \in H_A.$$

  Usually we would like to have high power (0.9) to detect a clinically

important Δ_A.

- Often time, T has (approximate) normal distribution under $\Delta = \Delta_A$:

$$T \overset{H_A=(\Delta_A,\theta)}{\sim} N(\phi(n, \Delta_A, \theta), \sigma_*^2(\Delta_A, \theta)).$$

Usually, $\sigma_*^2(\Delta_A, \theta) = 1$. In this case, $\phi(n, \Delta_A, \theta)$ is called the **non-centrality parameter**.

For example, the two-sample t-test statistic:

$$T_n = \frac{\bar{Y}_1 - \bar{Y}_2}{s_Y \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}} \approx \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma_Y \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}}$$

Then

$$\phi(n, \Delta_A, \theta) = \frac{\Delta_A}{\sigma_Y \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}}, \quad \sigma_*^2(\Delta_A, \theta) = 1.$$

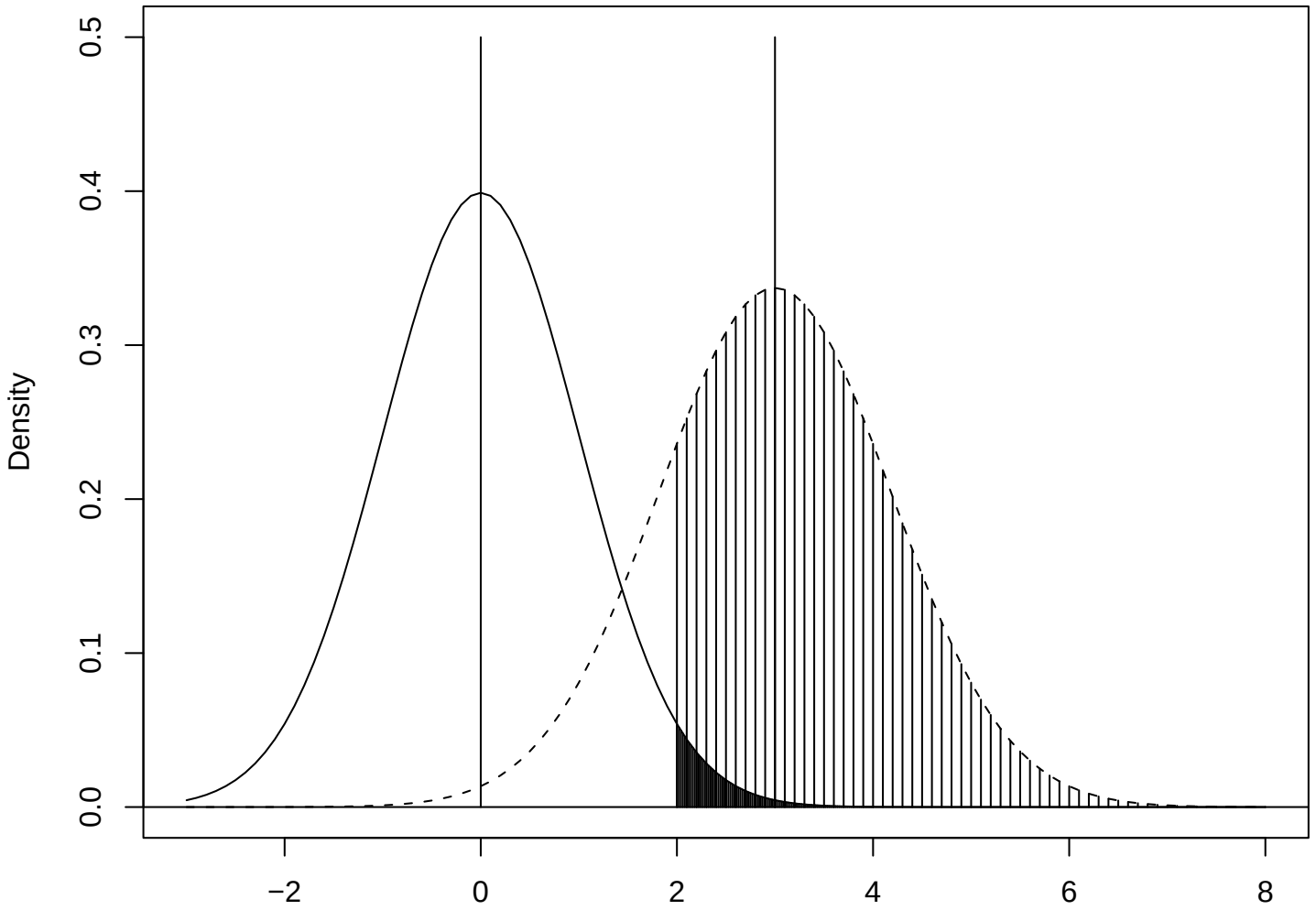
Deriving sample size to achieve desired power

Design characteristics:

- Use the above testing procedure
- Type I error probability α.
- Power $1 - \beta$ to detect clinically important treatment effect Δ_A
- Nuisance parameter θ is known

How to find sample size?

Figure 1: *Distributions of T under H_0 and H_A*



- The figure indicates the equation:

$$\phi(n, \Delta_A, \theta) = \{z_\alpha + z_\beta \sigma_*(\Delta_A, \theta)\}. \tag{6.1}$$

For the two-sample t-test, if we do equal allocation ($n_1 = n_2 = n/2$), then

$$\frac{\Delta_A}{\sigma_Y \left(\frac{2}{n} + \frac{2}{n}\right)^{1/2}} = z_\alpha + z_\beta$$

$$\Longrightarrow$$

$$n = \left\{ \frac{(z_\alpha + z_\beta)^2 \sigma_Y^2 \times 4}{\Delta_A^2} \right\}.$$

- **Note**: For two-sided tests we replace z_α by $z_{\alpha/2}$.

- **Example**: find the sample size necessary to detect a difference in mean response of 20 units between two treatments with 90% power using a t-test (two-sided) at the .05 level of significance. We assume population standard deviation of response σ_Y is expected to be about 60 units.

$$z_{\alpha/2} = z_{.025} = 1.96,\ z_\beta = z_{0.1} = 1.28,\ \Delta_A = 20,\ \sigma_Y = 60,$$

$$n = \frac{(1.96 + 1.28)^2 (60)^2 \times 4}{(20)^2} \approx 378 \text{ (rounding up)},$$

or about 189 per each treatment.

- **How large** should $\widehat{\Delta}$ be so that we will have a significant p-value (p-value=0.05) for the calculated sample size?

$$P_{\Delta=0}[T \geq |T_{obs}|] + P_{\Delta=0}[T \leq -|T_{obs}|] = 0.05$$

$$\Longleftrightarrow$$

$$P_{\Delta=0}[T \geq |T_{obs}|] = 0.025$$

$$\Longleftrightarrow$$

$$T_{obs} = z_{0.025} = 1.96$$

$$\Longleftrightarrow$$

$$\frac{\widehat{\Delta}}{\sigma_Y \left(\frac{2}{n} + \frac{2}{n}\right)^{1/2}} = z_{0.025} = 1.96$$

$$\Longrightarrow$$

$$\widehat{\Delta} = 1.96 \times \sigma_Y \left(\frac{4}{n}\right)^{1/2} = 1.96 \times 60 \times 2/\sqrt{378} = 12.1 < 20$$

- If the study result turns out to be what we expected, what P-value will be expected?

$$T_{obs} = \frac{\widehat{\Delta}}{\sigma_Y \left(\frac{2}{n} + \frac{2}{n}\right)^{1/2}} = \frac{20}{60 \times 2/\sqrt{378}} = 3.24$$

$$\mathrm{P-value} = 2P_{\Delta=0}[T > |T_{obs}|] = 2P_{\Delta=0}[T > 3.24] = 0.001.$$

Comparing two response rates

- $\pi_1 =$ response rate of treatment 1, $\pi_2 =$ response rate of treatment 2
- Treatment effect $\Delta = \pi_1 - \pi_2$
- n_1 patients are to be assigned to treatment 1, n_2 patients are to be assigned to treatment 2 (usually, $n_1 = n_2$)
- Wish to test $H_0 : \Delta \leq 0$ $(\pi_1 \leq \pi_2)$ versus $H_A : \Delta > 0$ $(\pi_1 > \pi_2)$.
- Data from each treatment:

$$X_1 \sim bin(n_1, \pi_1), \quad X_2 \sim bin(n_2, \pi_2)$$

- $p_1 = X_1/n_1$, $p_2 = X_2/n_2$ best estimates of π_1 and π_2

$$E(p_1) = \pi_1, \ var(p_1) = \frac{\pi_1(1-\pi_1)}{n_1},$$

$$E(p_2) = \pi_2, \ var(p_2) = \frac{\pi_2(1-\pi_2)}{n_2}.$$

- Test statistic for testing H_0:

$$T = \frac{p_1 - p_2}{\left\{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}^{1/2}},$$

where $\bar{p} = (X_1 + X_2)/(n_1 + n_2)$, best estimate of $\pi_1(\pi_2)$ under $\pi_1 = \pi_2$.

Note: The T^2 is the usual chi-square test used to test equality of proportions.

- We can write

$$\bar{p} = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = p_1 \left(\frac{n_1}{n_1 + n_2} \right) + p_2 \left(\frac{n_2}{n_1 + n_2} \right).$$

So

$$\bar{p} \approx \pi_1 \left(\frac{n_1}{n_1 + n_2} \right) + \pi_2 \left(\frac{n_2}{n_1 + n_2} \right) = \bar{\pi},$$

$\bar{\pi}$ is a weighted average of π_1 and π_2.

Therefore,

$$T \approx \frac{p_1 - p_2}{\left\{ \bar{\pi}(1 - \bar{\pi}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{1/2}}.$$

- The mean and variance of T under $\Delta = 0$:

$$\begin{aligned}
\mathrm{E}_{\Delta=0}(T) &\approx \mathrm{E}_{\Delta=0}\left\{\frac{p_1 - p_2}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}^{1/2}}\right\} \\
&= \frac{\mathrm{E}_{\Delta=0}(p_1 - p_2)}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}^{1/2}} = 0 \\
\mathrm{var}_{\Delta=0}(T_n) &\approx \frac{\{\mathrm{var}_{\Delta=0}(p_1) + \mathrm{var}_{\Delta=0}(p_2)\}}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}} \\
&= \frac{\left\{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right\}}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}} = 1.
\end{aligned}$$

So under $\Delta = 0$,

$$T \overset{(\Delta=0)}{\sim} \mathrm{N}(0,1)$$

- Under $H_A : \Delta = \Delta_A$:

$$T \overset{(\Delta=\Delta_A)}{\sim} \mathrm{N}(\phi(n, \Delta_A, \theta), \sigma_*^2)$$

where

$$\begin{aligned}
\phi(n, \Delta_A, \theta) &= \mathrm{E}_{H_A}(T) \approx \frac{(\pi_1 - \pi_2)}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}^{1/2}} \\
&= \frac{\Delta_A}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}^{1/2}}, \\
\sigma_*^2 &= \mathrm{var}_{H_A}(T) \approx \frac{\left\{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right\}}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}.
\end{aligned}$$

- With equal allocation of treatment, $n_1 = n_2 = n/2$, then

$$\phi(n, \Delta_A, \theta) = \frac{\Delta_A}{\left\{\bar{\pi}(1-\bar{\pi})\frac{4}{n}\right\}^{1/2}},$$

and

$$\sigma_*^2 = \frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{2\bar{\pi}(1-\bar{\pi})},$$

where $\pi_1 = \pi_2 + \Delta_A$.

- If we want to have power $1-\beta$ to detect an increase of Δ_A with significance level α using one-sided test (and equal allocation), the sample size n have to satisfy

$$\frac{n^{1/2}\Delta_A}{\{4\bar{\pi}(1-\bar{\pi})\}^{1/2}} = \mathcal{Z}_\alpha + \mathcal{Z}_\beta \left\{\frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{2\bar{\pi}(1-\bar{\pi})}\right\}^{1/2}.$$

So the sample size is given by

$$n = \frac{\left\{ \mathcal{Z}_\alpha + \mathcal{Z}_\beta \left\{ \frac{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}{2\bar{\pi}(1-\bar{\pi})} \right\}^{1/2} \right\}^2 4\bar{\pi}(1-\bar{\pi})}{\Delta_A^2}. \qquad (6.2)$$

Note: For two-sided tests we replace $\mathcal{Z}_\alpha$ by $\mathcal{Z}_{\alpha/2}$.

Example: Suppose the standard treatment of care (treatment 2) has a response rate of about .35 (best guess). After collaborations with your clinical colleagues, it is determined that a clinically important difference for a new treatment is an increase in .10 in the response rate. That is, a response rate of .45 or larger. If we are to conduct a clinical trial where we will randomize patients with equal allocation to either the new treatment (treatment 1) or the standard treatment, then how large a sample size is necessary to detect a clinically important difference with 90% power using a one-sided test at the .025 level of significance?

Note for this problem

- $\alpha = .025$, $\mathcal{Z}_\alpha = 1.96$
- $\beta = .10$ (power $= .9$), $\mathcal{Z}_\beta = 1.28$
- $\Delta_A = .10$
- $\pi_2 = .35$, $\pi_1 = .45$, $\bar{\pi} = .40$

Substituting these values into (6.2) we get

$$n = \frac{\left\{1.96 + 1.28\left\{\frac{.45\times.55+.35\times.65}{2\times.40\times.60}\right\}^{1/2}\right\}^2 4\times.40\times.60}{(.10)^2} \approx 1,004,$$

or about 502 patients on each treatment arm.

Arcsine square root transformation

- One **problem** with the above test statistic is that it does not have equal variance under $\Delta = 0$ and $\Delta = \Delta_A$, because p_i's variance depends on π_i.
- Variance stabilization transformation: $p = X/n$, $\mathrm{E}(p) = \pi$, $\mathrm{var}(p) = \pi(1-\pi)/n$.

  Want to find a monotone function $g(x)$ such that $\mathrm{var}(g(p))$ is a constant.
- Using delta method,

$$\mathrm{var}(g(p)) \approx [g'(\pi)]^2 \frac{\pi(1-\pi)}{n}$$

- If $g(x)$ satisfies

$$g'(x) = \frac{c}{\sqrt{x(1-x)}},$$

then $\mathrm{var}(g(p)) \approx \mathrm{aconstant}$

- It can be shown that one such $g(x)$ is given by

$$g(x) = sin^{-1}\sqrt{x}$$

such that $\mathrm{var}(g(p)) \approx 1/(4n)$.

- This $g(x) = sin^{-1}\sqrt{x}$ is a monotone function. Therefore,

$$H_0 : \pi_1 = \pi_2 \Longleftrightarrow H_0 : sin^{-1}(\sqrt{\pi_1}) = sin^{-1}(\sqrt{\pi_2}),$$

and

$$H_A : \pi_1 > \pi_2 \Longleftrightarrow H_A : sin^{-1}(\sqrt{\pi_1}) > sin^{-1}(\sqrt{\pi_2})$$

- The test statistic testing H_0 would be:

$$T = \frac{sin^{-1}(\sqrt{p_1}) - sin^{-1}(\sqrt{p_2})}{\left(\frac{1}{4n_1} + \frac{1}{4n_2}\right)^{1/2}},$$

- By what we derived,

$$T \overset{\Delta=0}{\sim} \mathrm{N}(0,1),$$

and

$$T \overset{\Delta=\Delta_A}{\sim} \mathrm{N}(\phi(n, \Delta_A, \theta), 1),$$

where

$$\phi(n, \Delta_A, \theta) = \mathrm{E}_{\Delta=\Delta_A}(T) = \frac{sin^{-1}(\sqrt{\pi_1}) - sin^{-1}(\sqrt{\pi_2})}{\left(\frac{1}{4n_1} + \frac{1}{4n_2}\right)^{1/2}}$$

and

$$\Delta_A = sin^{-1}(\pi_1)^{1/2} - sin^{-1}(\pi_2)^{1/2}.$$

- With equal allocation, $n_1 = n_2 = n/2$, the non-centrality parameter is

$$\phi(n, \Delta_A, \theta) = \mathrm{E}_{\Delta=\Delta_A}(T) = \sqrt{n}(sin^{-1}(\sqrt{\pi_1}) - sin^{-1}(\sqrt{\pi_2}))$$

In this case, the sample size n has to satisfy

$$n^{1/2}(sin^{-1}(\sqrt{\pi_1}) - sin^{-1}(\sqrt{\pi_2})) = (\mathcal{Z}_\alpha + \mathcal{Z}_\beta),$$

That is,

$$n = \frac{(\mathcal{Z}_\alpha + \mathcal{Z}_\beta)^2}{\Delta_A^2},$$

where

$$\Delta_A = sin^{-1}(\pi_1)^{1/2} - sin^{-1}(\pi_2)^{1/2}.$$

Note: Replace $\mathcal{Z}_\alpha$ by $\mathcal{Z}_{\alpha/2}$ for a two-sided test.

- Going back to our previous example,

$$n = \frac{(1.96 + 1.28)^2}{\{sin^{-1}(.45)^{1/2} - sin^{-1}(.35)^{1/2}\}^2} = \frac{(1.96 + 1.28)^2}{(.7353 - .6331)^2} = 1004,$$

the same result. This is because when sample size is this big, normal approximation is pretty good so it does not matter whether or not we do variance stabilization transformation.

Non-inferiority Trials

- Standard treatment in the market (treatment 2, response rate π_2)
- Want to show the new treatment (treatment 1) may be little bit worse than the standard treatment but within our tolerance limit Δ_A.
- Therefore, our hypothesis testing problem is:

$$H_0 : \pi_1 \leq \pi_2 - \Delta_A \text{ versus } H_A : \pi_1 > \pi_2 - \Delta_A.$$

- The test statistic:

$$T_n = \frac{p_1 - p_2 + \Delta_A}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}},$$

where n_1 and n_2 denote the number of patients allocated to treatments 1 and 2 respectively.

- On $H_0 \cap H_A$, i.e., $\pi_1 = \pi_2 - \Delta_A$,

$$T_n \overset{(\pi_1=\pi_2-\Delta_A)}{\sim} N(0,1).$$

So for the given level α, we reject H_0 if

$$T_n \geq \mathcal{Z}_\alpha.$$

Using this strategy, the new drug will not be approved with high probability ($\geq 1-\alpha$) when in fact it is worse than the standard treatment by at least Δ_A.

Remark: we didn't use the arcsin square-root transformation here. Because the arcsin square-root is non-linear; thus, a fixed difference of Δ_A in response probabilities between two treatments (hypothesis of interest) does not correspond to a fixed difference on the arcsin square-root scale.

Sample size calculations for non-inferiority trials

- We usually want to have high power to detect if the new drug is at least as good as the standard treatment. That is, the power of our test is calculated at $\pi_1 = \pi_2$.
- Under $\pi_1 = \pi_2$, our test statistic

$$T_n \overset{a}{\sim} \mathrm{N}(\phi, 1),$$

where

$$\phi = E(T_n) \overset{(\pi_1=\pi_2=\pi)}{\approx} \frac{\Delta_A}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

- If we do equal allocation, $n_1 = n_2 = n/2$, then

$$\phi = \frac{\Delta_A}{\sqrt{\pi(1-\pi)\left(\frac{4}{n}\right)}}.$$

- In order to have power $1-\beta$ to detect that the new drug is at least as good as the standard treatment, the non-centrality parameter ϕ has to satisfy

$$\frac{\Delta_A}{\sqrt{\pi(1-\pi)\left(\frac{4}{n}\right)}} = \mathcal{Z}_\alpha + \mathcal{Z}_\beta$$

or

$$n = \frac{(\mathcal{Z}_\alpha + \mathcal{Z}_\beta)^2 \times 4\pi(1-\pi)}{\Delta_A^2}. \tag{7.14}$$

Note: Since usually Δ_A is very small, the sample size will be much larger than the ones from superiority trials.

- **Example**: Suppose the response rate of the standard treatment is about 30% and a 95% CI of the treatment effect (difference of response rates between the standard treatment and placebo) in a clinical trial is [0.1, 0.25], and we want to show a new treatment is not inferior to the standard one.
  - ⋆ $\Delta_A = 0.1/2 = 0.05$
  - ⋆ $\alpha = 0.05$, so $\mathcal{Z}_{0.05} = 1.64$
  - ⋆ Good power: $1 - \beta = 0.9$, $\beta = 0.1$, $\mathcal{Z}_{0.1} = 1.28$
  - ⋆ $\pi_1 = 0.3$.
- The total sample size n:

$$n = \frac{(1.64 + 1.28)^2 \times 4 \times .3 \times .7}{(.05)^2} = 2864,$$

or 1432 patients per treatment arm.