Catherine M. Crespi

Exercises for Power and Sample Size in R Version 1.0 July 29, 2025

2

Exercises

2.1 A systolic blood pressure reading of 140 mm Hg or higher is considered high blood pressure (hypertension). Suppose that systolic blood pressure measurements in a population are normally distributed with a mean of 145 and a standard deviation that is known to be 10 mm Hg. You can use an appropriate R function for the calculations in this problem.

(a) What is the difference between 145 and 140 in units of standard deviation? Is this considered a small, medium or large effect size?

(b) If we have a sample of 25 individuals from the population, what is the power to reject the null hypothesis $H_0\colon \mu \leq 140$ and conclude $H_A\colon \mu > 140$ when specifying a significance level of 0.025?

(c) With a sample of 25 individuals, what is the lowest mean systolic blood pressure that could be detected with 80% power, when testing $H_0\colon \mu \leq 140$ versus $H_A\colon \mu > 140$ at significance level 0.025?

(d) What sample size is needed to reject $H_0\colon \mu \leq 140$ and conclude $H_A\colon \mu > 140$ with 80% power when specifying a significance level of 0.025?

(e) Continuing part (d), suppose that we expect that only 90% of the individuals that we recruit will provide a valid measurement of systolic blood pressure. How many individuals should we recruit to achieve the desired level of power? Round your answer up to the next highest whole number.

2.2 A depressive symptoms score of 16 or higher on a questionnaire is defined as "severe depression". Suppose that investigators plan to test whether the mean depressive symptoms score in a population is different from 16. The standard deviation in the population is known to be 6.

(a) Suppose that the true mean in a population is 15. Create a figure showing N (x-axis) versus power (y-axis) for a two-sided z test with $\alpha = 0.05$ for a range of sample sizes.

(b) Suppose that the true mean in a population is 12. Create a figure showing N (x-axis) versus power (y-axis) for a two-sided z test with $\alpha = 0.05$ for a range of sample sizes.

(c) In general, the sample sizes for the scenario in part (b) are smaller than those in part (a). Explain why.

2.3 Simulating power for a one-sample z test: z tests (and t tests) assume that the data are normally distributed. In this exercise, we simulate power for a one-sample z test when the data follow a uniform distribution.

Suppose that a sample of size 30 is taken from a population that follows a continuous uniform distribution on the interval $(0, 1)$.

(a) For a Uniform(a, b) distribution, the mean is $(b-a)/2$ and the variance is $(b-a)^2/12$. What are the mean and variance of a Uniform$(0, 1)$ distribution?

(b) Use simulation to estimate the power of a one-sample z test of $H_0\colon \mu = 0.4$ versus $H_A\colon \mu \neq 0.4$. Use at least 100,000 replications.

(c) What would be the power if the sample was drawn from a normal distribution with the same mean and variance? Note that the power for sampling from a uniform distribution should be only slightly lower because z and t tests are robust to violations of the normality assumption.

2.4 For a one sample z test, the sample size needed to achieve power of $1-\beta$ for a one-sided test with significance level α is $N \geq \frac{(z_{1-\beta}+z_{1-\alpha})^2\sigma^2}{(\mu_A-\mu_0)^2}$. Define the standardized mean effect size as $d = |\mu_A - \mu_0|/\sigma$.

(a) Solve the equation for d.

(b) Suppose that the standardized effect size d is doubled. What is the change in N?

(c) What is the smallest d that can be detected with 80% power, using one-sided α of 0.05, for N of 16, 36, 64 or 100?

2.5 For an upper-tailed, one-sample z test of $H_0\colon \mu \leq \mu_0$ versus $H_A\colon \mu > \mu_0$ with normally distributed data $Y_1, \ldots, Y_N$, power is the probability that the test statistic, $T = \frac{\bar{Y}-\mu_0}{\sigma/\sqrt{N}}$, falls in the rejection region given that the true mean is equal to μ_A, $P\left[T > z_{1-\alpha} \,\middle|\, Y_i \sim \mathcal{N}(\mu_A, \sigma^2)\right]$.

(a) Show that this probability expression is equal to the cumulative normal probability $\Phi\left(z_\alpha + \frac{\sqrt{N}(\mu_A-\mu_0)}{\sigma}\right)$.

(b) What is the probability of rejecting the null hypothesis when $\mu_A = \mu_0$? Which type of error is this?

(c) Solve the inequality $1-\beta \leq \Phi\left(z_\alpha + \frac{\sqrt{N}(\mu_A-\mu_0 A)}{\sigma}\right)$ for N, and thereby show that the sample size needed to achieve power of $1-\beta$ is $N \geq \frac{(z_{1-\beta}+z_{1-\alpha})^2\sigma^2}{(\mu_A-\mu_0)^2}$.

3

Exercises

3.1 A systolic blood pressure reading of 140 mm Hg or higher is considered high blood pressure (hypertension). Suppose that systolic blood pressure measurements in a population are normally distributed.

(a) We expect that the mean systolic blood pressure in a population is 145 and that the standard deviation is 10 mm Hg. If we take a sample of 25 individuals from the population, what is the power for a one-sample t test to reject the null hypothesis $H_0\colon \mu \leq 140$ and conclude $H_A\colon \mu > 140$ when specifying a significance level of 0.025?

(b) What sample size is needed to reject $H_0\colon \mu \leq 140$ and conclude $H_A\colon \mu > 140$ with 80% power when specifying a significance level of 0.025?

(c) Because the standard deviation is not known in advance and has a strong effect on power, it is a good parameter to target for a sensitivity analysis. Produce a table showing how the sample size requirement changes as the true standard deviation varies from 6 to 20 mm Hg.

(d) Suppose that in part (b), we expect that only 75% of the individuals that we recruit will provide a valid measurement of systolic blood pressure. How many individuals should we recruit to achieve the desired level of power? Round your answer up to the next highest whole number.

3.2 A randomized trial will use a two-sided, two-sample t test with significance level of 0.05 to compare systolic blood pressure outcomes in a treatment and a usual care condition.

(a) When using equal allocation with a total N of 100, what is the power to detect standardized effect sizes d ranging from 0.4 to 0.6?

(b) The investigators are considering a 3:1 allocation to the treatment and usual care conditions. If the allocation ratio $r = n_T/n_{UC}$ is equal to 3, what proportion of the sample will be allocated to each condition?

(c) Find the power to detect $d = 0.6$ when total N is 100 and allocation ratio $r = n_T/n_{UC} = 3$. How does this compare to the power for $d = 0.6$ in part (a)?

(d) Suppose that the variance of the outcome is expected to be twice as

high in the treatment group compared to the usual care group. The investigators prefer the simplicity of equal allocation, but they are concerned that equal allocation could result in a loss of power. What allocation ratio will provide the highest power? What would be the proportions of the sample allocated to each group? How does power for this allocation compare to power for equal allocation? What would you advise in this situation?

3.3 The POINTER trial was a two-arm randomized controlled trial comparing immediate to postponed catheter drainage in patients with infected necrotizing pancreatitis (Grinsven et al. Trials 20, 239 (2019). `https://doi.org/10.1186/s13063-019-3315-6`). The primary endpoint was the Comprehensive Complications Index (CCI), including all complications other than pre-existent complications occurring after randomization until 6 months after randomization. Based on this excerpt describing the sample size calculation, replicate the calculation:

"The sample size was calculated based on the primary endpoint, the CCI. A mean CCI score of 40 (with standard deviation of 27) for postponed catheter drainage is based on the number of complications identified in the step-up arm of the PANTER trial [3] and TENSION trial [9]. Analysis by Student's t test will have 80% power to detect a clinically relevant reduction of 15 to a CCI score of 25 [21] at a significance level of 0.05; for a sample size that equals 2 x 51, this will result in 102 evaluable patients. Assuming a dropout rate of about 2%, then 104 patients need to be included."

3.4 Investigators are planning a single-arm pre-post study in which participants will be assessed for the outcome variable (Y_1), an intervention will be applied, and then the outcome variable will be assessed again (Y_2). They will conduct a one-sided test $H_0 \colon \mu_d \leq 0$ versus $H_A \colon \mu_d > 0$, where $\mu_d = \mu_1 - \mu_2$ with $\alpha = 0.025$. Suppose that the true mean difference is 4 and the variance of the outcome variable is expected to be 100. The correlation between measurements within participants is expected to be in the range of 0.4-0.6.

(a) How many participants are needed to have 80% power to reject H_0? Compute the required sample size assuming values of 0.4, 0.5 and 0.6 for the correlation. Comment on how the value of the correlation affects the sample size required.

(b) Suppose that a total of 40 participants are available for the study. What is the smallest mean difference that can be detected with 80% power?

3.5 For t tests, sample size impacts power through two routes. Explain these two routes.

3.6 For a two-sample t test with equal group variances, show that power for a $k:1$ allocation to groups 1 and 2 is equal to power for a $1:k$ allocation to groups 1 and 2. (Hint: Show that the noncentrality parameters are equal and the degrees of freedom for the test statistic are equal.)

3.7 Sometimes, publications do not directly provide information that we would like, but we can calculate or estimate the quantity we are interested in using the information that is provided. Robinson et al (2016) `https://doi.org/10.1111/eip.12137` report results of a pilot study to assess the efficacy of a suicide prevention program among secondary school students. The study had a single-arm, pre-post design. The analysis of responses on the Suicide Ideation Questionnaire (SIQ) used a paired t test. The paper reports a T test statistic of 6.2 with 20 df. Using this information, obtain an estimate of the correlation between the pre and post test measurements. (Hint: What is the formula for the T statistic for a paired t test?)

3.8 Show that for a two independent sample t test with equal variances, maximal power occurs when we have equal sample sizes in each group, $n_1 = n_2$, i.e., allocation ratio $r = 1$. (Hint: Power is maximized when the noncentrality parameter is maximized.)

3.9 In a pharmacokinetic study, investigators want to compare two formulations of a drug with respect to an outcome variable called the area under the curve (AUC). They have 60 participants available and will use equal allocation, with half of the participants getting formulation 1 and half getting formulation 2. AUC is highly skewed and the data will be log-transformed prior to analysis, which will be conducted using a two-sample t test.

Let γ_1 and γ_2 represent the medians for the two formulations on the original scale. The study plans to test $H_0\colon \frac{\gamma_1}{\gamma_2} \leq 1$ versus $H_A\colon \frac{\gamma_1}{\gamma_2} > 1$. Suppose that the true median for formulation 1 is 3 and the true ratio of medians is 1.5.

(a) What is the true median for formulation 2? Assuming lognormal data, what are the means for formulations 1 and 2 on the log-transformed scale?

(b) Suppose that the common CV is 1.8. What is the common σ on the log-transformed scale?

(c) What is the power to reject the null with one-sided $\alpha = 0.025$?

3.10 A small pilot study is planned that will compare the survival times of two groups of insects. Survival times are expected to follow an exponential distribution. The median survival times in Groups 1 and 2 are expected to be 4 and 10 days. There will be 15 insects in each group.

(a) What are the rate parameters for the exponential distributions in the two groups?

(b) What is the power to reject the null hypothesis that the survival time distributions in the two groups are equal using a Wilcoxon-Mann-Whitney rank-sum test, with two-sided α of 0.05? Use a Monte Carlo simulation approach to estimate power.

(c) Could power be improved by using a log-transformation of the data and then comparing the two groups using a two-sample t test? Find the expected means and standard deviations of the log-transformed data in each group and compute power for a two-sample t test.

4

Exercises

4.1 Suppose that you are helping to design a study that seeks to show that a new medication is superior by a margin to a placebo for improving cognitive abilities among individuals with mild dementia. For the outcome, Cognitive Abilities Scale (CAS), a higher score corresponds to better cognitive abilities. The mean CAS in the two conditions will be compared using a two-sample t test. Participants will be randomized 3:1 to the new medication:placebo conditions. The margin of superiority is 2 and the common standard deviation is assumed to be 10.

(a) State the null and alternative hypotheses.

(b) If the means are expected to be 30 and 36 in the placebo and medication groups, respectively, compute the sample sizes required to achieve 80% power, when using a one-sided test with α of 0.025. Compute using both a normal approximation and in software that uses the more precise t-distribution based calculation.

4.2 You are helping to design a study that aims to demonstrate that an app-based intervention (Intervention 1) is noninferior to an in-person intervention (Intervention 2) for controlling hypertension in individuals with high blood pressure. The primary outcome variable is systolic blood pressure in mmHg; for this outcome, a **lower** mean is considered a better outcome. We will compare the means in the two groups using a two-sample t test assuming equal variance and using equal allocation. We will use one-sided α of 0.025. The common standard deviation is assumed to be $\sigma = 15$ mmHg. The noninferiority margin is 3 mmHg.

(a) State the null and alternative hypotheses.

(b) Although noninferiority studies are based on one-sided tests, inference for noninferiority studies often uses two-sided confidence intervals. What confidence coefficient should be used for a two-sided CI such that the inference will correspond to one-sided α of 0.025 (meaning, for a X% CI, what should be the value of X)?

(c) Assuming that $\mu_1 = \mu_2$, compute the sample size required for 80% power.

(d) Compute the sample size requirement again but assuming that outcomes for Intervention 1 are slightly worse (higher) than for Intervention 2, with $\mu_1 = \mu_2 + 1$. Is the required N higher or lower? Why?

4.3 A study is planned to test whether an internet-delivered psychological treatment is equivalence to a treatment as usual (TAU) face-to-face treatment in terms of reducing anxiety and depressive symptoms, as measured by change in the Hospital Anxiety and Depression Scale (HADS). A pilot study found that the mean change score for TAU was 3.9 with a standard deviation of 6.0. The equivalence limit is specified as 2. The study plans to use a parallel group design with equal allocation.

(a) State the null and alternative hypotheses.

(b) Assuming no true difference in mean change scores in the two groups, what sample size is needed for 80% power, when specifying a significance level of 0.05? Compute the sample size using both the normal approximation formula and using the "exact" Owen's Q method (in R).

(c) What is the confidence coefficient for a two-sided confidence interval that will correspond to the significance level of 0.05, that is, for an X% confidence interval, what is X?

(d) Suppose that the internet-delivered treatment is actually less effective than TAU, resulting in a lower mean change. Conduct a sensitivity analysis showing how the sample size requirement changes as the true difference in mean change increases.

4.4 Zegels et al (2013) `http://dx.doi.org/10.1016/j.joca.2012.09.017` reports a three-arm randomized trial in patients with knee osteoarthritis that compared single dose chondroitin 46 sulfate (CS), 3 daily dose CS and placebo. You are helping to design a new three-arm trial that will compare single-dose CS, a new experimental drug BS and placebo. The primary hypothesis is that CS and BS are equivalent; secondary hypotheses are that CS and BS are each superior by a margin to placebo. Using the information reported in the article to help provide parameter estimates, compute the following. You can assume equal variances in the three groups. The outcome measure is change in the total score of algofunctional Lequesne index (LI). For this exercise, you can rely mainly on the information in Table II of the paper, which reports results for change scores.

(a) When equal variances of change scores is assumed, a 95% confidence interval for a difference in mean change scores between two groups is formed as $\bar{d}_1 - \bar{d}_2 \pm t_{0.025,n_1+n_2-2}\hat{\sigma}_d\sqrt{1/n_1 + 1/n_2}$. Using this formula, solve for the standard deviation of the change scores.

(b) What sample size per group is needed to achieve 80% power to conclude that CS and BS are equivalent in an intent to treat analysis, using a significance level of 0.05? Include a sensitivity analysis. Use an equivalence margin of 1.0.

(c) The plan calls for doing a per protocol analysis to compare CS and BS. Assuming 15% of patient do not adhere to their assigned treatment, what sample size per group would you need for a per protocol analysis of the equivalence hypothesis?

(d) What sample size per group is needed to achieve 80% power to conclude that CS and BS are each superior to placebo, using a one-sided test of superiority with $\alpha = 0.025$, using an intent to treat analysis? Use a superiority margin of 1.0.

(e) Will the sample sizes found for (b) be adequate for (c)? What sample size per group is needed to provide adequate power for all three analyses?

(f) Write a succinct paragraph describing your sample size analysis, such as you would for it to be included in a trial protocol.

4.5 In a pharmacokinetic study, investigators want to compare two formulations of a drug with respect to an outcome variable called the area under the curve (AUC). They plan to use a parallel group design with equal allocation. AUC is highly skewed and the data will be log-transformed prior to analysis. (Note: It is more common to use a crossover design for pharmacokinetic studies. Crossover designs are discussed later in the course.)

(a) State the null and alternative hypotheses.

(b) What is the required sample size when specifying a CV of 30%, true geometric mean ratio of 1, 90% power, significance level of 0.05 and equivalence bounds of (80%, 125%)?

5

Exercises

5.1 Investigators are planning a study that will be analyzed using a one-way ANOVA. The factor variable is dose of medication and the response variable Y is serum concentration. They plan to evaluate four different doses: 0, 10, 20 and 30 milligrams. The expect the means will be 0, 20, 25 and 35, respectively. They assume that $\sigma = 30$.

(a) They will use the one-way ANOVA model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ where $\sum_i \alpha_i = 0$. What are the values of $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ for this study?

(b) What is the standard deviation of the effects, σ_e? What is the f effect size? Is the effect size small, medium or large?

(c) For a balanced design, what sample size is needed per group in order to obtain 90% power for an omnibus F test, for $\alpha = 0.05$?

(d) The investigators plan to conduct to test whether the mean of the means for the groups assigned to 10, 20 and 30 milligrams is equal to the mean for the group assigned to 0 milligrams. Express the contrast coefficient vector for this contrast. For groups of size 10, compute the power for a two-sided contrast test with α of 0.05.

5.2 A study seeks to investigate the effects of training time on hand grip strength, which is measured in kilograms. Men aged 60-75 will be randomized to 4 groups, with 25 in each group. The 4 groups will receive 0, 30, 60 and 90 minutes of training per week for 6 weeks. The standard deviation of hand grip strength in this population is expected to be 8.5 kg. It is planned that the data will be analyzed using a one-way ANOVA. The investigators expect that the group means will be 40, 42, 44 and 46.

(a) Compute the factors effects, the standard deviation of the effects, the ANOVA standardized effect sizes and the f effect size.

(b) Compute power for the omnibus F test for the one-way ANOVA, using α of 0.05.

(c) If you are helping to plan this study, would you recommend that the study be powered based on the omnibus F test? Why or why not?

(d) We can use a linear contrast to test whether there is a linear trend among the group means. A linear trend implies that for each change

in group level, the mean increases by a fixed amount. For a four-group ANOVA, the contrast coefficients for a linear contrast are $(c_1, c_2, c_3, c_4) = (-3, -1, 1, 3)$. Compute power for this linear contrast test at significance level 0.05.

5.3 An experiment is being designed to test whether the time to complete a math problem varies by the font type. Participants will be randomized 1:1:1 to three different font types: Arial, Bauhaus and Courier.

(a) What sample size per group is needed to detect a medium f effect size ($f = 0.25$) for an omnibus test with 90% power when using a significance level of 0.05?

(b) The standard deviation of time-to-complete is expected to be about 20 seconds. Provide two different sets of means (μ_A, μ_B, μ_C) that would correspond to an f effect size of 0.25.

(c) The investigators plan to compare all means pairwise. How many comparisons will there be?

(d) Suppose that the means of time-to-complete are expected to be $(\mu_A, \mu_B, \mu_C) = (60, 65, 75)$ with SD of 20 seconds. Assuming equal allocation to groups, find the smallest group size that will provide at least 80% power for testing each pairwise contrast, using two-sided tests and a Bonferroni correction to control the familywise error rate at 0.05.

(e) The investigators are considering whether to adjust for participant's score on a quantitative aptitude test when fitting the ANOVA model. What are the potential advantages and disadvantages of including such a control variable in the model?

(f) The expected correlation between quantitative aptitude score and time to complete math problems is 0.4. If you use an ANCOVA, how does this affect your answer to part (d)?

5.4 Investigator are planning a study with a 2×2 factorial design. They expect the following means:

	B1	B2
A1	15	21
A2	19	25

The population SD is assumed to be 12.

(a) These means assume there are no interactions. Calculate the values of the parameters $\mu, \alpha_1, \alpha_2, \beta_1, \beta_2$. The parameters should satisfy the zero-sum constraints.

(b) Calculate the f effect sizes for factors A and B. Are they small, medium or large?

(c) Determine the smallest sample size per cell that is needed to ensure at least 80% power for both factors. Assume equal sample sizes per cell.

Suppose that the investigators expect that for subjects getting B2, the level of factor A will have no effect. The postulated cell means are:

	B1	B2
A1	15	23
A2	19	23

(d) Calculate the values of the parameters $\mu, \alpha_1, \alpha_2, \beta_1, \beta_2$ and all of the $(\alpha\beta_{ij})$. The parameters should satisfy the zero-sum constraints.

(e) Calculate the f effect sizes for factors A and B and for the interaction. Are they small, medium or large?

(f) Determine the sample size per cell that is needed to ensure at least 80% power for both factors and the interaction effect. Assume equal sample sizes per cell.

5.5 Investigators are planning a 2x2 factorial ANOVA study to test the separate and joint effects of two treatments, A and B. They postulate the following means at follow-up:

	B	No B
A	10	20
No A	15	30

The within-group SD is assumed to be 20.

(a) Derive a table of effects for this study.

(b) Does the table of means imply an interaction between Factors A and B? If so, is the interaction effect small, medium or large?

(c) Suppose that the investigators are only interested in testing for the main effects of A and B and want 80% power for both. What n per group and total N are needed? Use two-sided tests at alpha of 0.05 for all tests.

(d) Suppose the investigators are interested in testing for the main effects of A and B and the interaction effect AB and want 80% power for all 3 tests. What n per group and total N are needed? Use two-sided tests at alpha of 0.05 for all tests.

(e) Plot power for each effect versus sample size for a reasonable range of sample sizes around the n your found in part (d). For example, if you found $n = 100$, plot power for n of 80-120.

(f) What is the power to detect the simple effect of factor A in the absence of B?

5.6 One-way ANOVA can be regarded as an extension of the two-sample t test with equal variances to more than 2 means. In this problem, we demonstrate the equivalence of a one-way ANOVA with two groups and an equal variance two-sample t test.

Suppose that $Y_{1i} \sim N(\mu_1, \sigma^2)$ and $Y_{2i} \sim N(\mu_2, \sigma^2)$, with all Y_{1i} and Y_{2i} independent. We collect samples of size n from each group and we wish to test the hypotheses $H_0 \colon \mu_1 = \mu_2$ versus $H_A \colon \mu_1 \neq \mu_2$ at two-sided $\alpha = 0.05$.

(a) Provide the test statistic T for testing these hypotheses using a two-sample t test and give its distribution when the null is true and when some alternative $\mu_1 \neq \mu_2$ is true. Express sample size in terms of group size n rather than total sample size N.

(b) Provide the test statistic F for testing these hypotheses using a one-way ANOVA approach and give its distribution when the null is true and when some alternative $\mu_1 \neq \mu_2$ is true. Express sample size in terms of group size n rather than total sample size N.

(c) In the one-way ANOVA model, we parametrize the group means as $\mu_i = \mu + \alpha_i$, where μ is the grand mean (mean of the group means) and $\sum_i \alpha_i = 0$. When there are two groups, the grand mean is $\mu = (\mu_1 + \mu_2)/2$. Show that when there are two groups, $\sum_i \alpha_i^2 = \frac{(\mu_1 - \mu_2)^2}{2}$.

(d) Show that the squared noncentrality parameter for T is equal to the noncentrality parameter for F.

(e) The Cohen standardized effect size is $d = \frac{\mu_1 - \mu_2}{\sigma}$. The f effect size for one-way ANOVA is $f = \frac{\sigma_e}{\sigma}$ where $\sigma_e = \sqrt{\frac{\sum_i \alpha_i^2}{a}}$. Show that $f = d/2$.

(f) Cohen defined small, medium and large effect sizes for comparing two means as d of 0.2, 0.5 and 0.8. He also defined small, medium and large effects in an ANOVA as f of 0.1, 0.25 and 0.4. Do these effect sizes imply the same difference between two means?

6

Exercises

6.1 A study is planned to assess vaccine efficacy. Participants will be randomized 2:1 to vaccine and placebo. The study aims to conclude that vaccine efficacy exceeds 40%. The true vaccine efficacy is assumed to be 65%.

(a) State the hypotheses.

(b) What proportion of the total number of disease events is expected to occur in the vaccinated group under the null hypothesis? What is this proportion when the vaccine efficacy is 65%?

(c) Calculate the target number of events needed to provide 90% power with $\alpha = 0.025$.

(d) If disease incidence in the unvaccinated group is expected to be 6 per 1000, how many evaluable participants are needed in each group?

6.2 Investigators want to assess an intervention designed to promote the receipt of mammograms to screen for breast cancer among high-risk women. Participants will be randomized 1:1 to an intervention group and a control group. The primary outcome is receipt of a mammogram within 12 months. They expect that in the control group, about 20% of the participants will obtain a mammogram within 12 months.

(a) The intervention is expected to increase mammogram receipt by 10 percentage points. How many participants are needed for a one-sided test at $\alpha = 0.025$ to achieve 80% power?

(b) Suppose that in the control group, 30% of participants are expected to obtain a mammogram within 12 months, and all other assumptions are the same. How many participants are needed for 80% power?

(c) Briefly explain why the sample size is different under these two scenarios, even though in both scenarios, the difference in proportions is 0.10.

(d) Compute the h effect sizes for the scenarios in parts (a) and (b). Are the effect sizes small, medium or large?

6.3 We are designing a study that aims to demonstrate that a new treatment is superior to a standard treatment by a margin. The outcome is disease remission in 30 days, and thus experiencing the outcome is "good". The investigators decide that the margin of clinical superiority is 0.1.

(a) What are the null and alternative hypotheses for this study?

(b) What sample size is needed to have 80% power to conclude that the new treatment is superior to the standard treatment, when the true outcome proportions are 0.6 (standard treatment) and 0.85 (new treatment)? Use one-sided α of 0.05 and assume equal allocation.

6.4 In a single-arm trial with a binary outcome variable measured at pretest and posttest, we expect that 70% of the participants will have a positive response at pretest. We further expect that 30% of the total number of participants will shift from a positive to a negative response and 10% will shift from negative to positive.

(a) What are the expected proportions in each of the four cells, p_{00}, p_{01}, p_{10} and p_{11}? What are the marginal proportions, p_{pre} and p_{post}?

(b) What is the value of ϕ, the phi coefficient? What is maximum phi coefficient?

(c) What are the expected discordant proportion ratio and proportion of observations that are discordant?

(d) What is the total number of patients that is required for power of 80% at a two-sided significance level of 0.05?

(e) (Challenging) Suppose that the trial will use 3:1 allocation to AB and BA rather than 1:1 allocation. If the total number of patients will be the same, which allocation scheme will provide higher power? Justify your answer.

7

Exercises

7.1 The endpoint of a phase I, single-arm study of dostarlimab monotherapy in patients with advanced and recurrent solid tumors is the objective response rate (ORR), a binary outcome. The null hypothesis is that the true ORR is $\leq 20\%$. The study is using a single-stage, fixed sample design.

(a) What is the smallest sample size required to achieve power of at least 90% with type I error rate of 2.5% (one-sided) when the true ORR is 40%?

(b) If the sample size is 65 patients, what is the power to reject the null hypothesis when the true ORR is 40% with type I error rate (one-sided) of not more than 2.5%? What is the decision rule for rejecting the null hypothesis?

7.2 Investigators have a new technique for performing gallbladder surgery. They need your help designing a single-arm study to assess the new technique. If its success rate is 45% or less, the new technique would not be acceptable for use. They plan a single-arm study to test $H_0 \colon p \leq 0.45$ versus $H_A \colon p > 0.45$. They want a type I error rate of not higher than 0.05 and type II error rate not higher than 0.2.

(a) If they expect that the true success rate will be 95%, would a sample of 4 patients be sufficiently large to achieve their objectives using a single-stage design? Makes a table showing the probability of $X = 0, 1, 2, 3, 4$ under the null hypothesis and under the alternative and explain how the table enables you to answer this question.

(b) For the previous question, find the optimal and minimax Simon two-stage designs (use software). Explain how to calculate the probability of early termination under the null (PET(p_0)) and the expected sample size under the null (EN(p_0)).

(c) Suppose that they expect the true success rate to be 65%. If they use a single-stage design, what is the minimum sample size required and the critical value? If they use a Simon two-stage design, what are the optimal and minimax designs? What are the PET and the expected sample sizes when the null is true? Briefly discuss how the three designs compare to each other.

7.3 Researchers plan a study to compare antimicrobial susceptibility, categorized as susceptible or non-susceptible, associated with extended oral antibiotic prophylaxis (EOAP) and with standard antibiotic prophylaxis (Std). Data analysis will use the Fisher exact test. Suppose there will be 15 samples in each antibiotic prophylaxis group. Susceptibility in the Std group is expected to range from 10-20%. For two-sided type I error rate of 0.05, what are the minimum detectable difference in proportions that can be detected with 80% power?

7.4 Repeat Exercise 4 using an exact test approach.

8

Exercises

8.1 In a randomized trial, patients will be randomized to an enhanced care intervention condition and a usual care condition with 2:1 allocation. The outcome variable is patient satisfaction with care, with categories of satisfied, neutral and unsatisfied. The expected proportions with these outcomes in the control condition are 0.4, 0.3 and 0.3.

(a) Suppose that the odds ratio for being at or below category k for the intervention group compared to the control group is 3. What are the group sample sizes needed for 80% power to test the hypothesis of no association between condition and patient satisfaction with a significance level of 0.05, under a proportional odds assumption?

(b) Under a proportional odds assumption, what are the expected proportions in the satisfied, neutral and unsatisfied categories?

(c) Assuming these expected proportions, what are the group sample sizes needed for 80% power to test the hypothesis of no association between condition and patient satisfaction using a chi-square test of independence, using a significance level of 0.05?

(d) What is the w effect size? Is this effect size considered small, medium or large?

(e) Which testing approach would you recommend and why?

9

Exercises

9.1 Investigators are planning to conduct a pilot study to gather information to support the planning of a larger study. The study will involve a total of 50 participants randomized 1:1 to two conditions. They are interested in how precisely they will be able to estimate various quantities.

(a) They will use the pilot data to estimate the difference in means between the two conditions using a 95% confidence interval. What is the probability that with this sample size, the halfwidth of a 95% confidence interval for the standardized difference between two means would not exceed 0.6?

(b) They expect that 80% of the participants will be retained, that is, will complete the study. If they compute a 95% confidence interval for the proportion out of 50 who are retained, what is the expected width of the confidence interval? (Hint: You can use the `prec_prop` function from the `presize` package with the "wilson" option.)

(c) Suppose that the investigators change their expectations and now expect that 70% of the participants will be retained. What is the expected width of a 95% confidence interval for this proportion? Is the expected width greater or less than the expected width in part (b)? Why?

10

Exercises

10.1 One of the objectives of a planned study is to estimate the correlation between two inflammatory markers, CRP and IL-6, in blood serum. Each participant will provide a blood sample and both CRP and IL-6 will be measured. Suppose that the true correlation between CRP and IL-6 is expected to be 0.4.

(a) What sample size is required to achieve 80% power to reject the null that $\rho = 0$, two-sided α of 0.05?

(b) What sample size is required to achieve 80% power to reject the null that $\rho < 0.2$, one-sided α of 0.025?

10.2 Investigators hypothesize that a certain genotype is associated with insulin resistance in older adults. They are planning a study that will collect cross-sectional data from a sample of adults ages 50 years and older. To test their hypothesis, they will use linear regression to regress a measure of insulin resistance (insulinres) on genotype (1=yes, 0=no) and the potential confounders age, race category, and education category. To help them plan their sample size requirement, they have a data set ("insulinresist", available in the powertools package) with similar variables collected from 60 individuals. They wish to know what sample size is needed for the planned study in order to have 80% probability of rejecting the null hypothesis that there is no association between the genotype and insulin resistance.

Conduct a sample size calculation for the planned study. Use the dataset to estimate the quantities that you need for the calculation. Include some sensitivity analysis to see how your required sample size varies as you change important assumptions. Write a short description of your calculation.

10.3 The formula for the noncentrality parameter (NCP) for the two-sample t test of $H_0\colon \mu_T = \mu_C$ versus $H_A\colon \mu_T \neq \mu_C$ with equal allocation and assuming equal variances is $\lambda = \sqrt{\frac{n}{2}}\frac{(\mu_T - \mu_C)}{\sigma}$, where n is the sample size per group and σ is the standard deviation of the outcome variable. Suppose that you plan to analyze the data from a planned two-group trial with equal allocation using a linear regression model. The model will include a dummy variable X_i equal to 1 for treatment group and 0

for control group. Recall that for a simple linear regression model, the NCP for the test of $H_0: \beta_1 = \beta_{10}$ is

$$\frac{(\beta_{1A} - \beta_{10})\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}}{\sigma_\epsilon}$$

where β_{1A} is the assumed true value of the regression coefficient, N is the total sample size, and σ_ϵ is the error standard deviation. Show that this quantity is mathematically equivalent to the NCP for the two-sample t test.

10.4 (a) In the insulinresist dataset, what is the correlation between insulin resistance (insulinres) and genotype (1=yes, 0=no)? Is this considered a small, medium or large effect size?

(b) What is the partial correlation between these variables after controlling for age? Is this considered a small, medium or large effect size?

(c) Use the partial correlation estimate to determine the sample size needed for a planned study in which you want to have 80% probability of rejecting the null hypothesis that there is no association between insulin resistance (Y variable) and genotype after controlling for age. Include some sensitivity analysis showing the sample size required for higher and lower values of the partial correlation.

10.5 Investigators would like to develop a model to predict the level of General Fatigue in breast cancer survivors. They need to determine a reasonable minimum sample size for this project. They will use information reported in Bower et al (2019) to provide likely values of parameters for their calculation. They will use the methods described in Riley et al. (2018) and implemented in the `pmsampsize` R package for their sample size calculations.

(a) Table 4 of Bower et al (2019) gives an R^2 value for a model for General Fatigue. What is the adjusted R^2 for this model?

(b) What is the mean and SD of General Fatigue reported in the Bower et al paper?

(c) Suppose the investigators have identified a set of predictors of interest that will require 14 predictor parameters (i.e., regression coefficients) in order to model. What is the smallest sample size that meets all four of the suggested criteria in Riley et al.?

(d) Assuming 14 predictor parameters, what is the smallest sample size that meets the criteria of apparent $R^2_{app} - R^2_{adj} \leq 0.05$?

(e) Suppose that the available development dataset has 400 observations. What is the maximum number of predictor parameters that could be used that would still allow all four of the suggested criteria in Riley et al. to be met?

11

Exercises

11.1 An observational study is planned to investigate the association between overweight/obesity (X) and diabetes (yes/no) (Y). Study participants will be classified as overweight or obese (BMI $\geq$ 25) or normal weight (BMI $<$ 25). The investigators expect that about 30% of the participants will be classified as overweight or obese. Among participants with normal weight, the proportion with diabetes is expected to be 0.05.

(a) If the study will collect data from 300 participants, what is the lowest odds ratio for diabetes that can be detected with 80% power, two-sided α of 0.05? Assume that being overweight/obese increases the odds of diabetes.

(b) Suppose the outcome analysis will use a logistic regression model adjusting for age. The correlation between age and overweight/obesity is expected to be 0.3. How does this change the lowest detectable odds ratio?

11.2 A study proposes to examine the association between the risk of a serious fall (binary outcome) and number of emergency department visits in the previous year among an over-65 population. The average number of emergency department visits in the previous year is 1.5.

(a) Assuming number of visits in past year follows a Poisson distribution, what is the probability of no visits in the past year? What is the probability of 1 visit?

(b) What is the odds ratio associated with one additional visit in the previous year?

(c) What sample size is needed to detect the association with at least 90% power?

11.3 Researchers are interested in developing a predictive model for the presence of a specific genotype. They have pilot data which are collected into the insulinresist dataset. Their candidate predictor pool includes the variables in this dataset plus an additional 5 dichotomous predictors.

(a) Describe the three criteria for good predictive modeling performance that are discussed in the Riley et al. (2019) Statistics in Medicine paper.

(b) What minimum sample size would you recommend for this modeling project?

13

Exercises

13.1 Answer the following with regard to multisite trials.

(a) What is meant by heterogeneity of the treatment effect in a multisite trial?

(b) Under what conditions could a multisite trial require a higher total number of participants as a trial with independent subjects to achieve the same power? Under what conditions could it require a lower total number?

13.2 Investigators are planning a multisite trial with a continuous outcome. They will randomize patients to intervention and control conditions 1:1 within site. They want 80% power to detect a standardized effect size of $d = 0.3$. They assume that $\rho_0 = 0$ and $\rho_1 = 0.05$.

(a) Provide an interpretation of the parameters ρ_0 and ρ_1.

(b) What is the total sample size that would be needed for this study if all of the observations were independent?

(c) The number of patients per site is uncertain and may range from 10 to 20. Compute the design effect associated with the multisite design when the number of patients per site equal to 10, 15 or 20.

(d) Using your answers to the previous question, compute the number of sites needed for the trial when the number of patients per site is equal to 10, 15 or 20.

(e) The efficiency of a multisite trial is reduced when there is variation in the number of participants per site. Suppose that we expect the number of participants per site to vary uniformly from 10 to 100. What would be the impact on the sample size requirement?

13.3 What is the variance ratio in a multisite trial? How do the variance ratio, number of sites, and number of subjects per site affect power for a test of heterogeneity of the treatment effect?

13.4 Investigators are planning a multisite trial with a binary outcome. They will randomize patients to intervention and control conditions 1:1 within site. The expected probabilities of the outcome are 0.5 in the control condition and 0.36 in the treatment condition. There are a total of 10 sites available to enroll patients in the study.

(a) What is the expected odds ratio? What are plausible values for σ_{u1} in this study?

(b) How many patients per site are needed to provide at least 80% power, assuming one-sided α of 0.025? Include a sensitivity analysis using a plausible range of values for σ_{u1}.

14

Exercises

14.1 A cluster randomized trial is planned that will measure a continuous outcome. The clusters will be allocated 1:1 to treatment and control conditions. The study will enroll 30 participants per cluster. The expected standardized effect size of $d = 0.2$ and the ICC value is expected to be between 0.01 and 0.03.

(a) What is the expected range of the design effect for this study? Briefly explain the interpretation of the design effect.

(b) Find the number of clusters and corresponding total number of participants needed to achieve 80% power, two-sided $\alpha = 0.05$, using both the normal approximation formula and the noncentral t distribution, for an ICC of 0.02.

(c) Suppose that there are only 10 sites available to be included in the study, with equal allocation of sites to conditions. For an ICC of 0.02, will it be possible to achieve 80% power using only 10 sites but increasing the number of participants per sites? If not, explain briefly why not.

(d) The number of participants per site is expected to range from 5 to 100, with an average of 30. How can this be expected to affect the standard error of the estimate of the treatment effect, compared to having equal-sized clusters? What is the expected impact on the required sample size?

14.2 A program to prevent re-arrest among inmates after release from jail will be evaluated using a cluster randomized trial design. An equal number of jails will be randomized to intervention and control conditions and at each jail, 30 inmates will participate in the trial. The expected re-arrest proportions are 0.10 and 0.20 in the intervention and control conditions.

(a) What is the expected odds ratio?

(b) What would be the number of inmates required to participate in the study to provide 80% power, with two-sided α of 0.05, if the outcomes were independent (i.e., no clustering)?

(c) To account for clustering on jail, we need to an estimate of σ_u. Use the `crt.varexplore` function to explore a range of values for σ_u. In

recent history for this jail system, the maximum observed re-arrest rate for a jail is 35%. Given this information, what are plausible values of σ_u?

(d) Calculate the number of jails required to provide 80% power, with two-sided α of 0.05, using the multilevel logistic modeling approach. Use the maximum plausible value for σ_u. How many inmates does this correspond to?