Statistical Power

PSY 300


IT'S IN THE
IT'S IN THE
SYLLABUS
IT'S IN THE
SYLLABUS

Statistical power

- Today's lecture is relevant for the midterm
- Significance level α: Probability of rejecting the null hypothesis H_0 given that the null hypothesis is true
  - Or: Probability of a "false positive" or "Type I" error

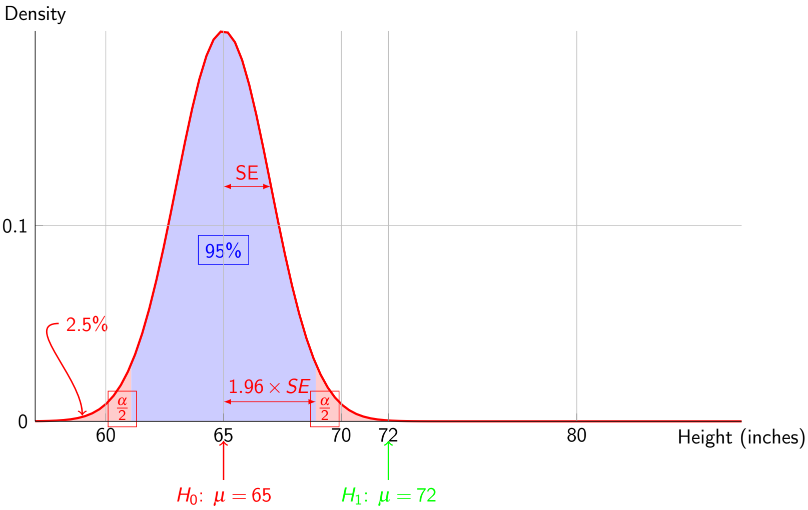
Statistical power

- Today's lecture is relevant for the midterm
- Significance level α: Probability of rejecting the null hypothesis H_0 given that the null hypothesis is true
  - Or: Probability of a "false positive" or "Type I" error
- Statistical power $1 - \beta$: Probability of rejecting the null hypothesis H_0 given that the null hypothesis is false

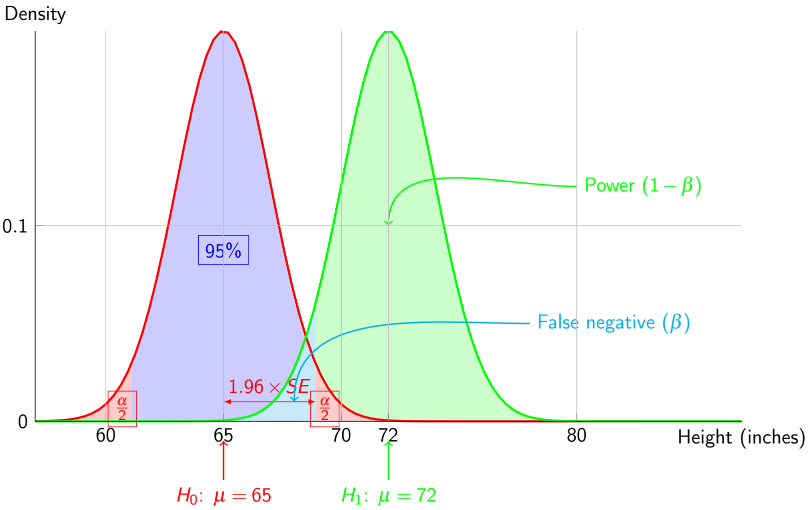
Statistical power

- Today's lecture is relevant for the midterm
- Significance level α: Probability of rejecting the null hypothesis H_0 given that the null hypothesis is true
  - Or: Probability of a "false positive" or "Type I" error
- Statistical power $1-\beta$: Probability of rejecting the null hypothesis H_0 given that the null hypothesis is false
  Alternative ways of saying this:
  - $Power = \Pr\left(\text{reject } H_0 \mid H_0 \text{ is false}\right)$.
  - Power is the probability of *correctly rejecting the null hypothesis*
  - Probability of *not* obtaining a "false negative" or "Type II" error.
    - The probability of making a Type II error is β, so the probability of *not* doing that is $1-\beta$.

"Distribution under the Null"



Distributions under the Null and an Alternative



Probability of a false negative (Type II error, β)

- Power for a specific alternative hypothesis is the probability of rejecting the null when the alternative hypothesis is true.

Probability of a false negative (Type II error, β)

- Power for a specific alternative hypothesis is the probability of rejecting the null when the alternative hypothesis is true.
- Consider an alternative hypothesis that claims that $\mu = 72$. What is the probability that we will reject the null of $\mu = 65$ when the alternative hypothesis is true?

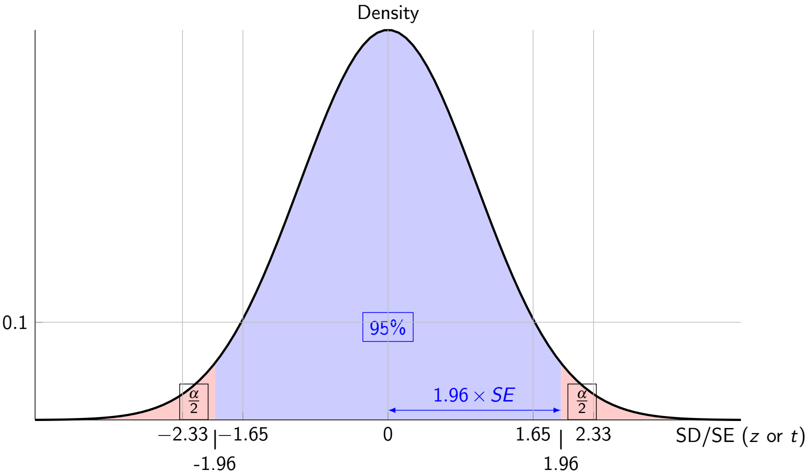
Probability of a false negative (Type II error, β)

- Power for a specific alternative hypothesis is the probability of rejecting the null when the alternative hypothesis is true.
- Consider an alternative hypothesis that claims that $\mu = 72$. What is the probability that we will reject the null of $\mu = 65$ when the alternative hypothesis is true?
- It's the probability of drawing from a sampling distribution centered at 72 a sample that lies in the left-most 2.5% of the distribution under the null, plus the probability of drawing from that sampling distribution centered at 72 a sample that lies in the right-most 2.5% of the distribution under the null.

Probability of a false negative (Type II error, β)

- Power for a specific alternative hypothesis is the probability of rejecting the null when the alternative hypothesis is true.
- Consider an alternative hypothesis that claims that $\mu = 72$. What is the probability that we will reject the null of $\mu = 65$ when the alternative hypothesis is true?
- It's the probability of drawing from a sampling distribution centered at 72 a sample that lies in the left-most 2.5% of the distribution under the null, plus the probability of drawing from that sampling distribution centered at 72 a sample that lies in the right-most 2.5% of the distribution under the null.
- So we need to figure out those areas. How do we calculate the area under a normal distribution?

Areas under the normal distribution



Areas under the normal distribution

- Areas under the normal distribution are best obtained by *standardizing* the distribution first

Areas under the normal distribution

- Areas under the normal distribution are best obtained by *standardizing* the distribution first
- For an entire variable, this is achieved by subtracting the mean of the variable, and dividing by the standard deviation: $x_{std} = \frac{x - \bar{x}}{SD}$. The resulting standardized variable will have mean 0 and standard deviation 1. It is called a "normalized" or "standardized" or "z-scored" variable.

Areas under the normal distribution

- We can also calculate what a particular value of the variable would turn into if the distribution were normalized. This amounts to simply expressing the distance between that value and the mean of the distribution in terms of standard deviations. Example: where does 72 lie in a normal distribution centered at 65 with SD=2 when that distribution is normalized?

  $$z = \frac{72 - 65}{2} = 3.5$$

  We say that "72 is 3.5 standard deviations from the mean of the distribution" (or 3.5 standard errors if the distribution is a sampling distribution).

Areas under the normal distribution

- We can also calculate what a particular value of the variable would turn into if the distribution were normalized. This amounts to simply expressing the distance between that value and the mean of the distribution in terms of standard deviations. Example: where does 72 lie in a normal distribution centered at 65 with SD=2 when that distribution is normalized?

  $$z = \frac{72 - 65}{2} = 3.5$$

  We say that "72 is 3.5 standard deviations from the mean of the distribution" (or 3.5 standard errors if the distribution is a sampling distribution).
- This is also called z-scoring. So we can say "72 has a z-score of 3.5".

Areas under the normal distribution

- We can also calculate what a particular value of the variable would turn into if the distribution were normalized. This amounts to simply expressing the distance between that value and the mean of the distribution in terms of standard deviations. Example: where does 72 lie in a normal distribution centered at 65 with SD=2 when that distribution is normalized?

$$z = \frac{72 - 65}{2} = 3.5$$

  We say that "72 is 3.5 standard deviations from the mean of the distribution" (or 3.5 standard errors if the distribution is a sampling distribution).

- This is also called z-scoring. So we can say "72 has a z-score of 3.5".
- If this looks familiar to the way we calculated t last time: Note that z is the asymptotic equivalent of t; t is for finite samples. It's usually fine to use either, and in most applications you don't have to worry about which one is correct because the computer will make that choice for you.

Areas under the normal distribution

- For a given z-score, we can use the computer to look up what area of the normal distribution is to the left of it.

Areas under the normal distribution

- For a given z-score, we can use the computer to look up what area of the normal distribution is to the left of it.
- This is the same as the probability of obtaining a statistic at least as extreme as this one on that side of the distribution. (That's half of the p-value.)

Areas under the normal distribution

- For a given z-score, we can use the computer to look up what area of the normal distribution is to the left of it.
- This is the same as the probability of obtaining a statistic at least as extreme as this one on that side of the distribution. (That's half of the p-value.)
- Notation: Area to the left of z under the standard normal distribution $= \Phi(z)$. Example: $\Phi(-1.96) \approx 0.025$

Areas under the normal distribution

- For a given z-score, we can use the computer to look up what area of the normal distribution is to the left of it.
- This is the same as the probability of obtaining a statistic at least as extreme as this one on that side of the distribution. (That's half of the p-value.)
- Notation: Area to the left of z under the standard normal distribution $= \Phi(z)$. Example: $\Phi(-1.96) \approx 0.025$
- In *R*:
  - Using the z-distribution (asymptotic):

    ```
    > pnorm(-1.96, mean = 0, sd = 1)
    [1] 0.0249979
    ```

  - Using the t-distribution (finite sample with $N = 100$):

    ```
    > pt(-1.96,99)
    [1] 0.02640356
    ```

Areas under the normal distribution

- For a given z-score, we can use the computer to look up what area of the normal distribution is to the left of it.
- This is the same as the probability of obtaining a statistic at least as extreme as this one on that side of the distribution. (That's half of the p-value.)
- Notation: Area to the left of z under the standard normal distribution $= \Phi(z)$. Example: $\Phi(-1.96) \approx 0.025$
- In *R*:
  - Using the z-distribution (asymptotic):

    ```
    > pnorm(-1.96, mean = 0, sd = 1)
    [1] 0.0249979
    ```

  - Using the t-distribution (finite sample with $N = 100$):

    ```
    > pt(-1.96,99)
    [1] 0.02640356
    ```

- When *R* does a power calculation, it uses the z-distribution.

Some important z-scores

Left + right tail area (α)	Central area $(1-\alpha)$	Corresponding z-score
0.1	0.9	1.645
0.05	0.95	1.96
0.01	0.99	2.58

Standardized effect sizes

- When we study treatment effects, we often express the treatment effect in standard deviation or z-score units. This is also called the "standardized effect size" or "Cohen's d".

Standardized effect sizes

- When we study treatment effects, we often express the treatment effect in standard deviation or z-score units. This is also called the "standardized effect size" or "Cohen's d".
  - Example: "We find that a lottery win increases happiness by 0.2 SD."

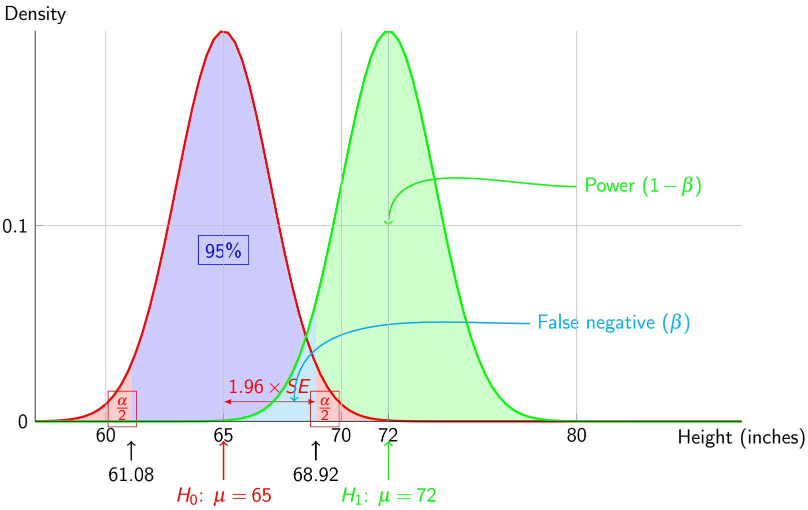
Standardized effect sizes

- When we study treatment effects, we often express the treatment effect in standard deviation or z-score units. This is also called the "standardized effect size" or "Cohen's d".
  - Example: "We find that a lottery win increases happiness by 0.2 SD."
- Cohen: 0.2 (small), 0.5 (medium), 0.8 (large)

Standardized effect sizes

- When we study treatment effects, we often express the treatment effect in standard deviation or z-score units. This is also called the "standardized effect size" or "Cohen's d".
  - Example: "We find that a lottery win increases happiness by 0.2 SD."
- Cohen: 0.2 (small), 0.5 (medium), 0.8 (large)
  - This was the olden days—now we think even 0.2 SD is a perfectly respectable effect size (Funder & Ozer, 2018)

Back to our power calculation example



Back to our power calculation example

- Power for a specific alternative hypothesis is the probability of rejecting the null when the alternative hypothesis is true.

Back to our power calculation example

- Power for a specific alternative hypothesis is the probability of rejecting the null when the alternative hypothesis is true.
- We want the probability of drawing from a sampling distribution centered at 72 with $SD = 2$ a sample that lies to the left of 61.08, or to the right of 68.92.

Back to our power calculation example

- Power for a specific alternative hypothesis is the probability of rejecting the null when the alternative hypothesis is true.
- We want the probability of drawing from a sampling distribution centered at 72 with $SD = 2$ a sample that lies to the left of 61.08, or to the right of 68.92.
- Probability of a sample to the left of 61.08:
  - z-score: $z = \frac{61.08-72}{2} = -5.46$
  - Probability: $\phi(-5.46) \approx 2.4 \times 10^{-8} \approx 0$

Back to our power calculation example

- Power for a specific alternative hypothesis is the probability of rejecting the null when the alternative hypothesis is true.
- We want the probability of drawing from a sampling distribution centered at 72 with $SD = 2$ a sample that lies to the left of 61.08, or to the right of 68.92.
- Probability of a sample to the left of 61.08:
  - z-score: $z = \frac{61.08-72}{2} = -5.46$
  - Probability: $\phi(-5.46) \approx 2.4 \times 10^{-8} \approx 0$
- Probability of a sample to the right of 68.92:
  - z-score: $z = \frac{68.92-72}{2} = -1.54$
  - Probability: $1 - \phi(-1.54) \approx 1 - 0.0617 \approx 0.94$

Back to our power calculation example

- Power for a specific alternative hypothesis is the probability of rejecting the null when the alternative hypothesis is true.
- We want the probability of drawing from a sampling distribution centered at 72 with $SD = 2$ a sample that lies to the left of 61.08, or to the right of 68.92.
- Probability of a sample to the left of 61.08:
  - z-score: $z = \frac{61.08-72}{2} = -5.46$
  - Probability: $\phi(-5.46) \approx 2.4 \times 10^{-8} \approx 0$
- Probability of a sample to the right of 68.92:
  - z-score: $z = \frac{68.92-72}{2} = -1.54$
  - Probability: $1 - \phi(-1.54) \approx 1 - 0.0617 \approx 0.94$
- Power is the sum of these: $Power = 2.4 \times 10^{-08} + 0.94 \approx 0.94$

Power calculations in *R*: One-sample *t*-test

```
> install.packages("pwr")
> library(pwr)
> pwr.t.test(d = 0.2, sig.level = 0.05,
   power = 0.8, type = c("one.sample"))
   One-sample t test power calculation
              n = 198.1508
              d = 0.2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Power calculations in *R*: Two-sample *t*-test

```
> pwr.t.test(d = 0.2, sig.level = 0.05,
     power = 0.8, type = c("two.sample"))
     Two-sample t test power calculation
              n = 393.4057
              d = 0.2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
NOTE: n is number in *each* group
```

Power calculations in *R*: Two-sample *t*-test

```
> pwr.t.test(d = 0.2, sig.level = 0.05,
     power = 0.8, type = c("two.sample"))
     Two-sample t test power calculation
              n = 393.4057
              d = 0.2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
NOTE: n is number in *each* group
```

What to write: "A power calculation determined that with 393 participants in each group (786 participants in total), the study had 80 percent power to detect effect sizes of 0.2 standard deviations at the 5 percent significance level."

Two-sample t-test: finding the detectable effect size

```
> pwr.t.test(n=100, sig.level = 0.05,
     power = 0.8, type = c("two.sample"))
    Two-sample t test power calculation
              n = 100
              d = 0.3981407
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
NOTE: n is number in *each* group
```

Two-sample *t*-test: finding the detectable effect size

```
> pwr.t.test(n=100, sig.level = 0.05,
     power = 0.8, type = c("two.sample"))
    Two-sample t test power calculation
              n = 100
              d = 0.3981407
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
NOTE: n is number in *each* group
```

What to write: “A power calculation determined that with 100 participants in each group (200 participants in total), the study had 80 percent power to detect effect sizes of 0.39 standard deviations at the 5 percent significance level.”

Two-sample t-test: finding power

```
> pwr.t.test(d=0.2, n=100, sig.level = 0.05,
     type = c("two.sample"))
     Two-sample t test power calculation
              n = 100
              d = 0.2
      sig.level = 0.05
          power = 0.2906459
    alternative = two.sided
NOTE: n is number in *each* group
```

Two-sample t-test: finding power

```
> pwr.t.test(d=0.2, n=100, sig.level = 0.05,
     type = c("two.sample"))
     Two-sample t test power calculation
              n = 100
              d = 0.2
      sig.level = 0.05
          power = 0.2906459
    alternative = two.sided
NOTE: n is number in *each* group
```

What to write: "A power calculation determined that with 100 participants in each group (200 participants in total), the study had 29 percent power to detect effect sizes of 0.2 standard deviations at the 5 percent significance level."

What does power depend on?

- Significance level α: The larger α, the greater power; but also: the greater the false positive rate!

What does power depend on?

- Significance level α: The larger α, the greater power; but also: the greater the false positive rate!
- Standard error: $SE = \frac{SD}{\sqrt{n}}$. The smaller the standard error, the greater power.

What does power depend on?

- Significance level α: The larger α, the greater power; but also: the greater the false positive rate!
- Standard error: $SE = \frac{SD}{\sqrt{n}}$. The smaller the standard error, the greater power.
  - Reduce SD: make the population more homogeneous; use control variables
  - Increase n
  - Avoid clustering (more on this later)

What does power depend on?

- Significance level α: The larger α, the greater power; but also: the greater the false positive rate!
- Standard error: $SE = \frac{SD}{\sqrt{n}}$. The smaller the standard error, the greater power.
  - Reduce SD: make the population more homogeneous; use control variables
  - Increase n
  - Avoid clustering (more on this later)
- Effect size: The larger the effect size, the more power we have to detect it with a given sample size.

What does power depend on?

- Significance level α: The larger α, the greater power; but also: the greater the false positive rate!
- Standard error: $SE = \frac{SD}{\sqrt{n}}$. The smaller the standard error, the greater power.
  - Reduce SD: make the population more homogeneous; use control variables
  - Increase n
  - Avoid clustering (more on this later)
- Effect size: The larger the effect size, the more power we have to detect it with a given sample size.
- Ratio of group sizes (e.g. treatment vs. control observations): maximized at $n_0 = n_1$

What does power depend on?

- Significance level α: The larger α, the greater power; but also: the greater the false positive rate!
- Standard error: $SE = \frac{SD}{\sqrt{n}}$. The smaller the standard error, the greater power.
  - Reduce SD: make the population more homogeneous; use control variables
  - Increase n
  - Avoid clustering (more on this later)
- Effect size: The larger the effect size, the more power we have to detect it with a given sample size.
- Ratio of group sizes (e.g. treatment vs. control observations): maximized at $n_0 = n_1$
- We have some control over all of these; mostly n.

Clustering

- Sometimes, studies are structured such that the observations are grouped:
  - Students in classrooms
  - Lab participants in experimental sessions (e.g. you test groups of 6 at the same time)
  - Households in cities

Clustering

- Sometimes, studies are structured such that the observations are grouped:
  - Students in classrooms
  - Lab participants in experimental sessions (e.g. you test groups of 6 at the same time)
  - Households in cities
- When there is correlation between units within a cluster, they are not fully independent observations. But our tests assume that they are!

Clustering

- How to fix this problem? We *adjust the standard errors for clustering* during analysis. Will talk about this more later.
  - When the outcomes of units within a cluster are positively correlated, this will increase the standard errors, i.e. make our estimates less precise.
  - Rarely, the outcomes of units in a cluster are negatively correlated. Then, standard errors decrease, i.e. our estimates become more precise.

Clustering

- How to fix this problem? We *adjust the standard errors for clustering* during analysis. Will talk about this more later.
  - When the outcomes of units within a cluster are positively correlated, this will increase the standard errors, i.e. make our estimates less precise.
  - Rarely, the outcomes of units in a cluster are negatively correlated. Then, standard errors decrease, i.e. our estimates become more precise.
- Intuitive way of thinking about it: when the outcomes of a group are correlated, the *effective number of observations* is smaller than the number of units in the group.

Clustering and power

- If there is positive within-cluster correlation, and randomization is at the cluster level (i.e. we have to cluster standard errors):
  - Standard errors will increase
  - Therefore power will decrease!

Clustering and power

- If there is positive within-cluster correlation, and randomization is at the cluster level (i.e. we have to cluster standard errors):
  - Standard errors will increase
  - Therefore power will decrease!
- We can adjust for clustering during the power calculation

Power calculation with clusters in *R*

```
> install.packages("CRTSize")
> library(CRTSize)
> n4means(delta=0.2, sigma=1, m=10, ICC=0.2,
alpha=0.05, power=0.8, AR=1, two.tailed=TRUE, digits=3)
The required sample size is a minimum of 110 clusters
of size 10 in the Experimental Group and a minimum of
110 clusters (size 10) in the Control Group.
```

- delta: effect size; sigma: standard deviation (1 if using Cohen's d); m: cluster size; ICC: intra-cluster correlation (estimated from data, or a guess); alpha: significance level; AR: allocation ratio (1 if treatment and control group sizes equal); digits: rounding.

Power calculation with clusters in *R*

```
> install.packages("CRTSize")
> library(CRTSize)
> n4means(delta=0.2, sigma=1, m=10, ICC=0.2,
alpha=0.05, power=0.8, AR=1, two.tailed=TRUE, digits=3)
The required sample size is a minimum of 110 clusters
of size 10 in the Experimental Group and a minimum of
110 clusters (size 10) in the Control Group.
```

- delta: effect size; sigma: standard deviation (1 if using Cohen's d); m: cluster size; ICC: intra-cluster correlation (estimated from data, or a guess); alpha: significance level; AR: allocation ratio (1 if treatment and control group sizes equal); digits: rounding.
- Note the sample size required here (2200) is much larger than when detecting the same effect size without clustering (786).

What to base power calculations on

- It's common to run power calculations based on effect sizes observed in previous studies. For example, we sometimes read something like this: "Previous study X found that the effect of interest is 0.2 SD. We therefore powered our study to detect effects of this magnitude with 90% power."

What to base power calculations on

- It's common to run power calculations based on effect sizes observed in previous studies. For example, we sometimes read something like this: "Previous study X found that the effect of interest is 0.2 SD. We therefore powered our study to detect effects of this magnitude with 90% power."
- Problem: the "statistical significance filter": Statistically signifiant findings have a higher likelihood of getting published.
  - This means that effect sizes in the published literature tend to be overestimated.
  - Therefore, it's good to run power calculations for smaller effect sizes than those observed in the literature

What to base power calculations on

- It's common to run power calculations based on effect sizes observed in previous studies. For example, we sometimes read something like this: "Previous study X found that the effect of interest is 0.2 SD. We therefore powered our study to detect effects of this magnitude with 90% power."
- Problem: the "statistical significance filter": Statistically signifiant findings have a higher likelihood of getting published.
  - This means that effect sizes in the published literature tend to be overestimated.
  - Therefore, it's good to run power calculations for smaller effect sizes than those observed in the literature
- So what should we do?

What to base power calculations on

- Power the study for “the smallest effect size you care about” based on theoretical or practical significance
  - E.g.: “The smallest difference in scores considered clinically important by psychiatrists is X. We therefore power the study to observe effect sizes of this magnitude.”
  - Or: “Due to budget constraints, we the study was powered to detect effect sizes of 0.2 SD with 80% power at the 5 percent significance level. In our view, this detectable effect size is small enough to be theoretically and practically interesting.”

Ex ante vs. ex post power calculations

- What we have done so far are ex ante power calculations, i.e. before running the study

Ex ante vs. ex post power calculations

- What we have done so far are ex ante power calculations, i.e. before running the study
- Often, we are interested in determining how well-powered the study was after we have run it.

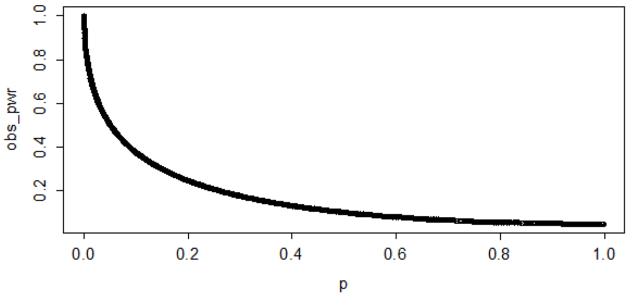
Ex ante vs. ex post power calculations

- What we have done so far are ex ante power calculations, i.e. before running the study
- Often, we are interested in determining how well-powered the study was after we have run it.
- It's tempting to run an ex-post power calculation based on your *observed* effect in the study and your actual n: "How much power did we have to detect the effect we observe?"

Ex ante vs. ex post power calculations

- What we have done so far are ex ante power calculations, i.e. before running the study
- Often, we are interested in determining how well-powered the study was after we have run it.
- It's tempting to run an ex-post power calculation based on your *observed* effect in the study and your actual n: "How much power did we have to detect the effect we observe?"
- But: This is circular; ex-post power is a direct function of your p-value.
- More information: http://daniellakens.blogspot.com/2014/12/observed-power-and-what-to-do-if-your.html

Observed power is a direct function of the *p*-value



Minimum detectable effect size (MDE)

- So: Don't run ex-post power calculations that use your observed effect size and p-value—they simply re-state the information contained in the p-value in a different form.

Minimum detectable effect size (MDE)

- So: Don't run ex-post power calculations that use your observed effect size and p-value—they simply re-state the information contained in the p-value in a different form.
- So what can we do? We can ask for the **minimum detectable effect size** (MDE): What is the effect size that we had 80% power to detect at the 5% significance level? Two ways of getting at this:

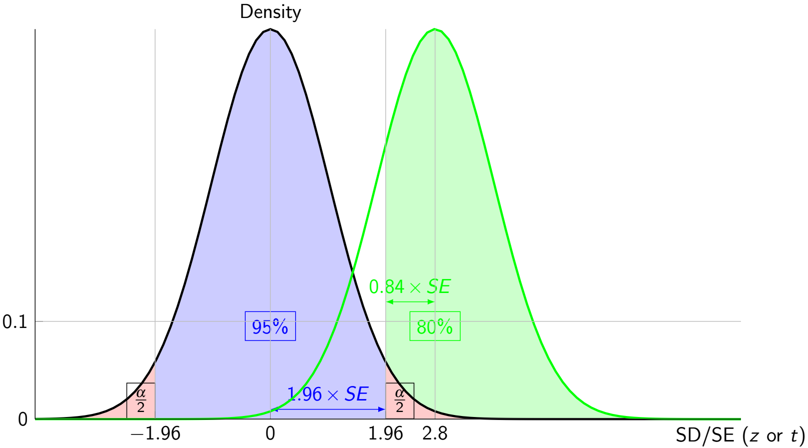
Minimum detectable effect size (MDE)

- So: Don't run ex-post power calculations that use your observed effect size and p-value—they simply re-state the information contained in the p-value in a different form.
- So what can we do? We can ask for the **minimum detectable effect size** (MDE): What is the effect size that we had 80% power to detect at the 5% significance level? Two ways of getting at this:
  - Run a power calculation in *R* using your actual n; let it give you the power of the study. Disadvantage: uses assumptions, not actual numbers, for precision gains/losses from control variables, clustering, etc.

Minimum detectable effect size (MDE)

- So: Don't run ex-post power calculations that use your observed effect size and p-value—they simply re-state the information contained in the p-value in a different form.
- So what can we do? We can ask for the **minimum detectable effect size** (MDE): What is the effect size that we had 80% power to detect at the 5% significance level? Two ways of getting at this:
  - Run a power calculation in *R* using your actual n; let it give you the power of the study. Disadvantage: uses assumptions, not actual numbers, for precision gains/losses from control variables, clustering, etc.
  - Use the observed standard error to determine what effect size we had 80% power to detect

Minimum detectable effect size (MDE)



Minimum detectable effect size (MDE)

- To declare an effect significant, the estimator has to be at least 1.96 standard errors away from zero

Minimum detectable effect size (MDE)

- To declare an effect significant, the estimator has to be at least 1.96 standard errors away from zero
- To have an 80% chance that we draw an estimator at least 1.96 standard errors away from zero, 80% of the distribution from which we draw this estimator (the sampling distribution) has to be to the right of 1.96.

Minimum detectable effect size (MDE)

- To declare an effect significant, the estimator has to be at least 1.96 standard errors away from zero
- To have an 80% chance that we draw an estimator at least 1.96 standard errors away from zero, 80% of the distribution from which we draw this estimator (the sampling distribution) has to be to the right of 1.96.
- Because $\phi(0.84) \approx 0.80$, this happens to be achieved with a sampling distribution centered at $1.96 + 0.84 = 2.8$ standard errors away from zero.

Minimum detectable effect size (MDE)

- To declare an effect significant, the estimator has to be at least 1.96 standard errors away from zero
- To have an 80% chance that we draw an estimator at least 1.96 standard errors away from zero, 80% of the distribution from which we draw this estimator (the sampling distribution) has to be to the right of 1.96.
- Because $\phi(0.84) \approx 0.80$, this happens to be achieved with a sampling distribution centered at $1.96 + 0.84 = 2.8$ standard errors away from zero.
- Example: "To determine which effect sizes we had 80% power to observe, we calculate the minimum detectable effect size (MDE) by multiplying the standard error of the estimate by 2.8. This gives an MDE of 0.05 SD."

Minimum detectable effect size (MDE)

- To declare an effect significant, the estimator has to be at least 1.96 standard errors away from zero
- To have an 80% chance that we draw an estimator at least 1.96 standard errors away from zero, 80% of the distribution from which we draw this estimator (the sampling distribution) has to be to the right of 1.96.
- Because $\phi(0.84) \approx 0.80$, this happens to be achieved with a sampling distribution centered at $1.96 + 0.84 = 2.8$ standard errors away from zero.
- Example: "To determine which effect sizes we had 80% power to observe, we calculate the minimum detectable effect size (MDE) by multiplying the standard error of the estimate by 2.8. This gives an MDE of 0.05 SD."
- More information: https://blogs.worldbank.org/impactevaluations/why-ex-post-power-using-estimated-effect-sizes-bad-ex-post-mde-not

MDEs and Null Effects

- MDEs are especially useful when you have null effects

MDEs and Null Effects

- MDEs are especially useful when you have null effects
- Null effects are often difficult to interpret: Is there no effect? Or is there an effect, but we just didn't detect it?

MDEs and Null Effects

- MDEs are especially useful when you have null effects
- Null effects are often difficult to interpret: Is there no effect? Or is there an effect, but we just didn't detect it?
- "We observe no statistically significant effect of our experimental manipulation on the outcome. Because we were powered to detect very small effect sizes (MDE = 0.05 SD), we can rule out even small treatment effects of our manipulation."