DATA2002

Sample size calculations and power

Garth Tarr

THE UNIVERSITY OF SYDNEY

Power and sample size

Errors in hypothesis testing

	H_0 **true (innocent)**	H_0 **false (guilty)**
Don't reject H_0 (acquit)	Correct decision	Type II error (β)
Reject H_0 (guilty)	Type I error (α)	Correct decision ($1-\beta$)

- Type I errors: level of significance, $\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$
- Type II errors: call it β
- Power: $1-\beta = P(\text{reject } H_0 \mid H_1 \text{ true})$

General testing setup

- Suppose we are interesting in inference concerning an unknown population mean μ.
- We are considering a fixed value μ_0 ("**hypothesised value**").
- We then observe the data $x_1, \ldots, x_n$, obtaining
- the sample mean $\bar{x}$
- the sample sd s and thus the *estimated standard error* (se) $s/\sqrt{n}$.
- We decide to perform a (say, two-sided) t-test, that is to say if the *observed discrepancy* $\bar{x} - \mu_0$ is *large* compared to the se, we will "reject" the value μ_0 as "implausible":

$$\text{Reject if } |\bar{x} - \mu_0| > c\frac{s}{\sqrt{n}}\ ,$$

where c is chosen so that the **false alarm rate** is some fixed, small value α (e.g. 0.05, 0.01).

Model assumptions

- The **false alarm rate** determination can only be made if a suitable statistical model is assumed for the data.
- If we model the data $x_1, \ldots, x_n$ as values taken by iid $N\left(\mu, \sigma^2\right)$ random variables $X_1, \ldots, X_n$ (with μ and σ^2 *both unknown*), then whatever be the true value μ, the ratio

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \,.$$

- The **false alarm rate** is

$$P_{\mu_0}\left(|\bar{X} - \mu_0| > c\frac{S}{\sqrt{n}}\right) = P_{\mu_0}\left(\frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} > c\right) = P(|t_{n-1}| > c) = 2P(t_{n-1} > c)$$

  by symmetry.
- The $P_{\mu_0}(\cdot)$ indicates probability when the true value is μ_0, i.e. the value we specified in the null hypothesis.

Beer example

- Suppose we have $n = 6$ and choose a **false alarm rate** of $\alpha = 0.05$.
- Then the constant c needs to satisfy

$$2P(t_5 > c) = \alpha$$

so

$$P(t_5 > c) = \alpha/2 = 0.025$$

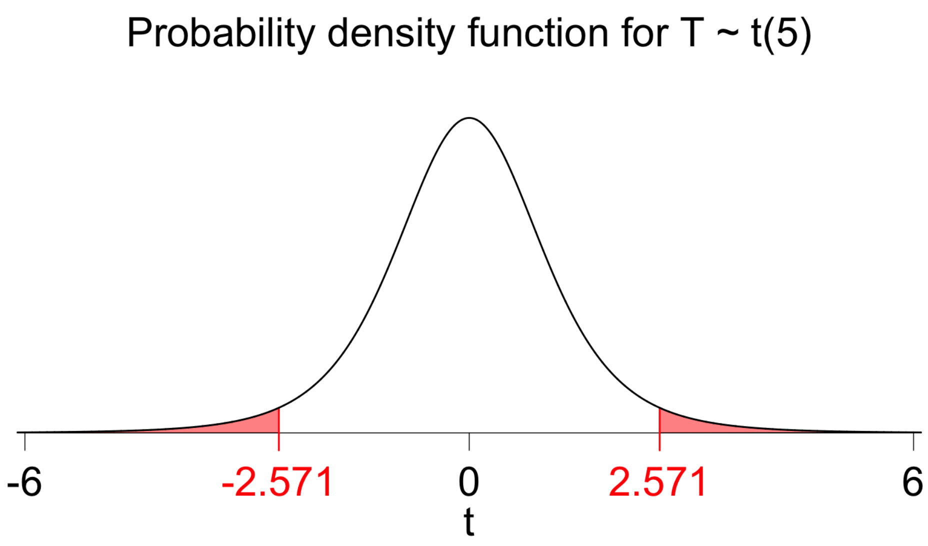
and thus

$$P(t_5 \leq c) = 0.975\,.$$

```
c_05 = qt(0.975,5)
c_05
```

```
## [1] 2.570582
```

The total area in the tails is 0.05, so the area in each tail is 0.05/2 = 0.025.



Why allow false alarms at all?

- A fair question is: *why would you set things up to have 5% **false alarm rate**?*
- Why not make it *really small*, like 10^{-6}?
- *Answer:* because then you would never reject anything, even if you should!
- The technical reason: because then the test would have **no power**.

The power of a test is the probability that the test rejects the null hypothesis, H_0 when a **specific** alternative hypothesis H_1 is true.

$$\text{Power} = P(\text{reject } H_0 \mid H_1 \text{ is true}).$$

Statistical power in one sample t-test

- Consider the probability of "rejecting" as a function of the true population mean μ:

$$P_{\mu}\left(\text{reject } H_0\right) = P_{\mu}\left(\left|\bar{X} - \mu_0\right| > c\frac{S}{\sqrt{n}}\right) = P_{\mu}\left(\frac{\left|\bar{X} - \mu_0\right|}{S/\sqrt{n}} > c\right).$$

- This is the **statistical power function** of the test.
- To determine this we need to know the *distribution* of the t-statistic for testing μ_0:

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

when the **true population mean** μ is *not necessarily equal to* μ_0 (the hypothesised population mean)!

We won't go into details, with the HTML slides, press p if you'd like to see the details.

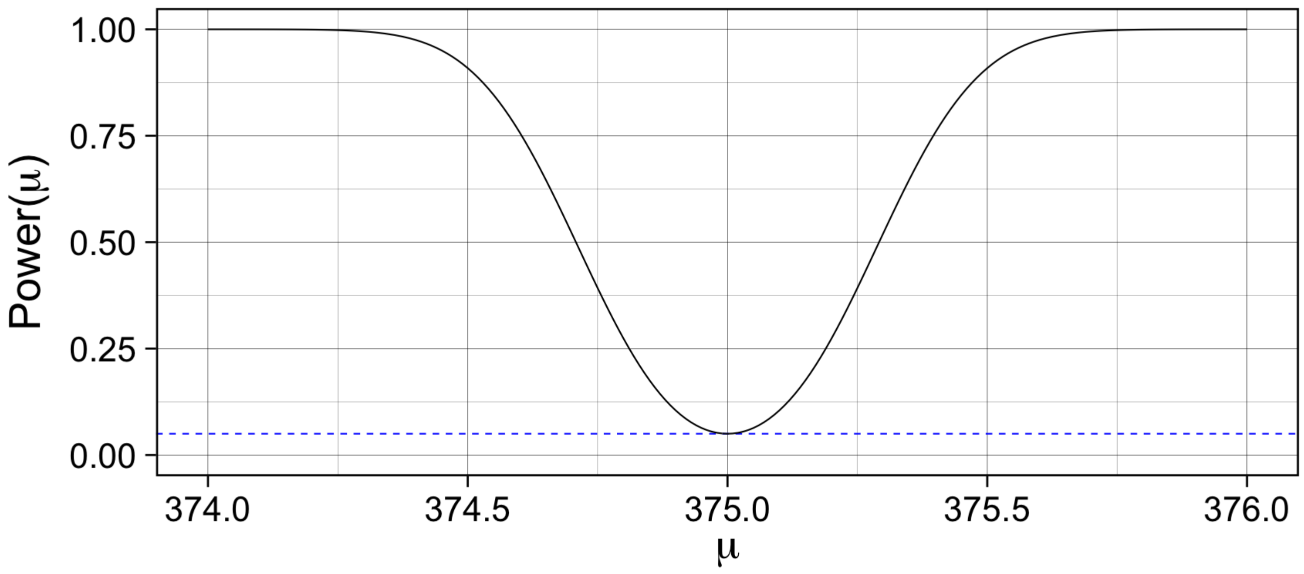
Beer example: power calculations

- Suppose the sample sd is *indicative* of the "true" population sd. We want to plot the power function of the test as a function of μ.
- First let us assume the "true" σ is equal to the sample value

```
x = c(374.8, 375.0, 375.3, 374.8, 374.4, 374.9)
sig = sd(x)
sig
```

```
## [1] 0.294392
```

Power as a function of μ



Note: this *supposes* the "true" σ is equal to the estimate 0.294; it is all a guess, but is still useful as an "estimated" power function.

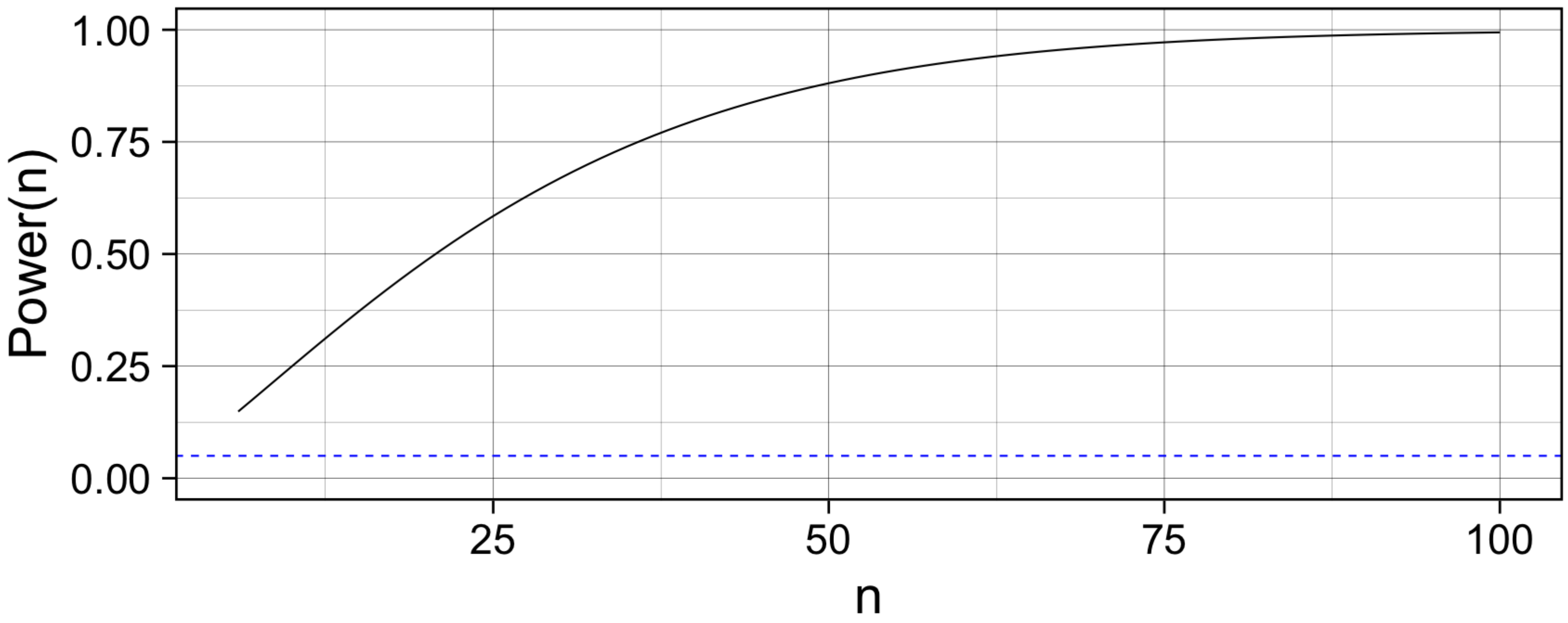
Power as a function of n

- Now let us suppose that *both* the sample mean and sample sd are indicative of the "true" values μ and σ:

```
xbar = mean(x)
c(xbar, sig)
```

```
## [1] 374.866667   0.294392
```

- Can we see how the power ought to behave as a function of n?
- *Note* the degrees of freedom, and thus the critical value, change with n.



Again, this is assuming the "true" values μ and σ equal the sample values $\bar{x}$ and s. But it is still useful!

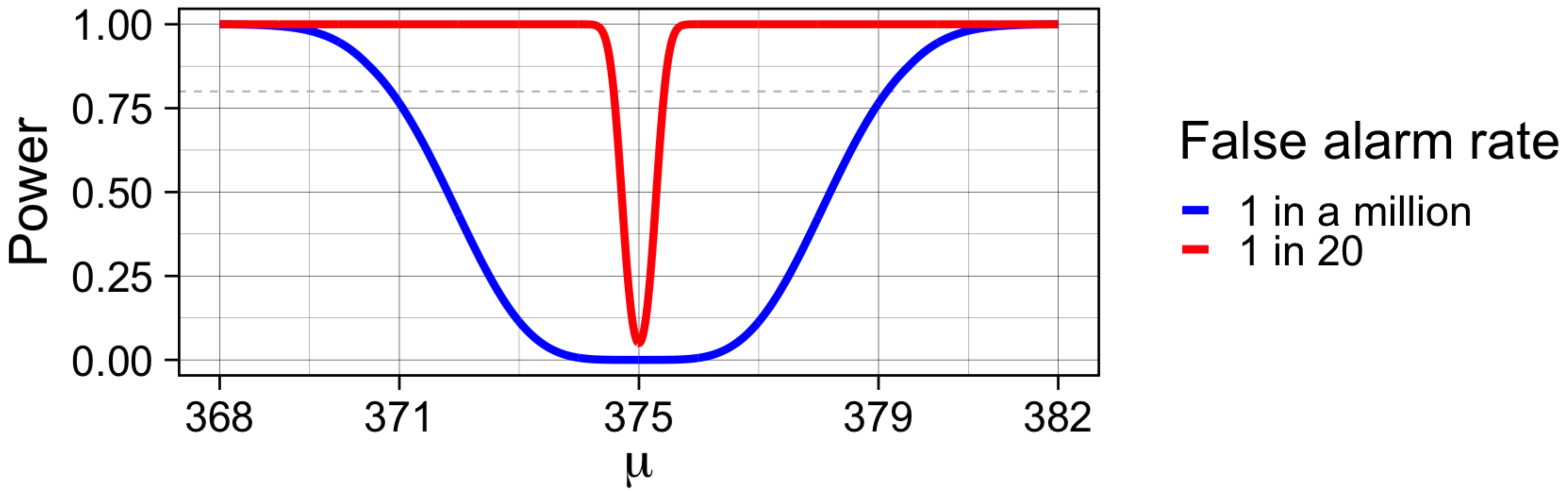
Comparing to false alarm rate of 10^{-6}.

- For a test with $n = 6$ and a **false alarm rate** of $\alpha = 10^{-6}$, we need a critical value of

```
c.million = qt(1-(1e-6)/2, df = 5)
c.million
```

```
## [1] 28.47847
```

So we would need a discrepancy equal to more than 28 standard errors before we would reject 375!



Power of about 80% would require a true μ lower than 371 or more than 379!

Just tell me how to do it easily in R

There's a package for that

The **pwr** package.

```r
# install.packages("pwr")
library(pwr)
```

The key functions:

- `pwr.t.test()` t-tests (one sample, 2 sample, paired)
- `pwr.t2n.test()` t-test (two samples with unequal n)

```
pwr.t.test(n = NULL, d = NULL, sig.level = 0.05, power = NULL,
type = c("two.sample", "one.sample", "paired"),
alternative = c("two.sided", "less", "greater"))
```

```
pwr.t2n.test(n1 = NULL, n2= NULL, d = NULL, sig.level = 0.05, power = NULL,
alternative = c("two.sided",
"less","greater"))
```

Cohen's d

Rather than specifying a null mean and an alternative mean and standard deviation, the **pwr** functions take as an input "Cohen's d":

$$d = \frac{|\mu_1 - \mu_2|}{\sigma}$$

Cohen suggests that d values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes respectively.

Beer example

- Supposing the population sd $\sigma = 0.294$, with a sample size of $n = 6$ how much lower than 375 does μ need to be for us to be 80% sure of "detecting" that $\mu \neq 375$ with a *two-sided* test which has **false alarm rate** 0.05?

```
res = pwr.t.test(n = 6, d = NULL, sig.level = 0.05, power = 0.8,
                 type = "one.sample", alternative = "two.sided")
res
```

```
##
##      One-sample t test power calculation
##
##               n = 6
##               d = 1.434538
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

```
res$d*0.294 # d * sigma gives the difference between means
```

```
## [1] 0.4217541
```

Beer example

Suppose that $\mu = 374.87$ and $\sigma = 0.294$, what sample size n would be needed to be 80% sure of detecting that $\mu \neq 375$ with a two-sided test which has **false alarm rate** 0.05?

```
res = pwr.t.test(n = NULL, d = (374.87-375)/0.294, sig.level = 0.05, power = 0.8,
                 type = "one.sample", alternative = "two.sided")
res
```

```
##
##      One-sample t test power calculation
##
##               n = 42.10456
##               d = 0.4421769
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Further reading

See chapter 11 of Nordmann and McAleer (2021), explore this web app and read section 6.4 of Larsen and Marx (2012)

Champely, S. (2020). *pwr: Basic Functions for Power Analysis*. R package version 1.3-0. URL: https://CRAN.R-project.org/package=pwr.

Larsen, R. J. and M. L. Marx (2012). *An Introduction to Mathematical Statistics and its Applications*. 5th ed. Boston, MA: Prentice Hall. ISBN: 978-0-321-69394-5.

Nordmann, E. and P. McAleer (2021). *Fundamentals of Quantitative Analysis*. URL: https://psyteachr.github.io/quant-fun-v2/power-and-error.html.