One-sample inference: Continuous data

Patrick Breheny

Introduction

- So far we've discussed how to carry out hypothesis tests and construct confidence intervals for categorical outcomes: success versus failure, life versus death
- Today we turn our attention to continuous outcomes like blood pressure, cholesterol, etc.
- We've seen how continuous data must be summarized and plotted differently, and how continuous probability distributions work very differently from discrete ones
- As we'll see today, there are also differences in how these data must be analyzed

Notation

- The usual notation for working with means is similar to that of proportions:
  - μ denote the population mean (the true, unknown population mean)
  - The observed sample mean can be denoted either $\bar{x}$ or $\hat{\mu}$, to emphasize that it estimates the population mean
  - μ_0 will denote the hypothesized value of the population mean under the null
  - H_0 is shorthand for the null hypothesis, as in $H_0 : \mu = \mu_0$

Using the central limit theorem

- We've already used the central limit theorem to construct confidence intervals and perform hypothesis tests for categorical data
- The same logic can be applied to continuous data as well, with one wrinkle
- For categorical data, the parameter we were interested in (π) also determined the standard deviation: $\sqrt{\pi(1-\pi)}$
- For continuous data, the mean tells us nothing about the standard deviation

Estimating the standard error

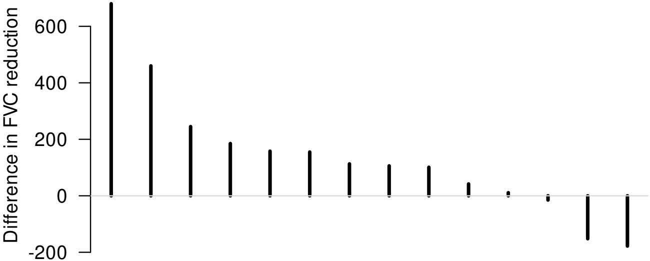
- In order to perform any inference using the CLT, we need a standard error
- We know that $\mathrm{SE} = \mathrm{SD}/\sqrt{n}$, so it seems reasonable to estimate the standard error using the sample standard deviation as a stand-in for the population standard deviation
- This turns out to work decently well for large n, but as we will see, has problems when n is small

Procedure for a z-test

- So the procedure for z-tests is:
  1. Calculate the standard error: $\mathrm{SE} = \mathrm{SD}/\sqrt{n}$
  2. Calculate the test statistic $z = (\hat{\mu} - \mu_0)/\mathrm{SE}$, where $\hat{\mu}$ is the sample mean
  3. Calculate the area under the normal curve outside $\pm z$
- This is the same procedure we had before with categorical data, except for how we estimate the SD
- One can also make z-confidence intervals based on the same idea

FVC example

- Let's calculate a p-value based on this z-test, returning to the same cystic fibrosis crossover study that we discussed back in the "One-sample categorical data" lecture
- However, instead of focusing on whether the patient did better on drug or placebo (a categorical outcome), let us now focus on *how much better* the patient did on the drug



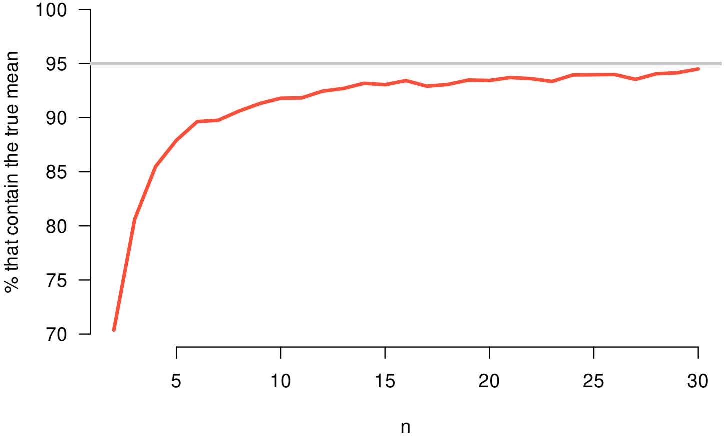
FVC example (cont'd)

- In the study, the mean difference in reduction in FVC (placebo − drug) was 136, with standard deviation 223
- Performing the z-test:
  1. $\mathrm{SE} = 223/\sqrt{14} = 60$
  2. $$\begin{aligned} z &= \frac{136-0}{60} \\ &= 2.29 \end{aligned}$$
  3. The area outside ± 2.29 is $2(0.011) = 0.022$
- This is fairly substantial evidence that the drug helps prevent deterioration in lung function

Flaws with the z-test

- However, as I mentioned before, these procedures are flawed when n is small
- This is a completely separate flaw than the issue of "how accurate is the normal approximation?"
- Indeed, this is a problem even when the sampling distribution is perfectly normal
- This flaw can be witnessed by repeatedly drawing random samples from the normal distribution, then constructing 95% confidence intervals and seeing how often they contain the true mean

Simulation results



What would a simulation involving hypothesis tests look like?

Why isn't the z-test working?

- The flaw with the z-test is that it is ignoring one of the sources of the variability in the test statistic
- We're acting as if we know the standard error, but we're really just estimating it from the data
- In doing so, we underestimate the amount of uncertainty we have about the population based on the data

Fixing the z-test

- The procedure to modify the z-test to account for this uncertainty is called the t-test, and was invented by W.S. Gossett
- Gossett's employers had him publish under the pen name "Student" because they didn't want the competition to know how useful his results could be
- Because of this, the t-test is often called "Student's t-test"

Student's curve

- Gossett showed that when the SE is estimated from the standard deviation instead of calculated exactly from the population, the statistic

  $$\frac{\hat{\mu} - \mu}{\mathrm{SE}}$$

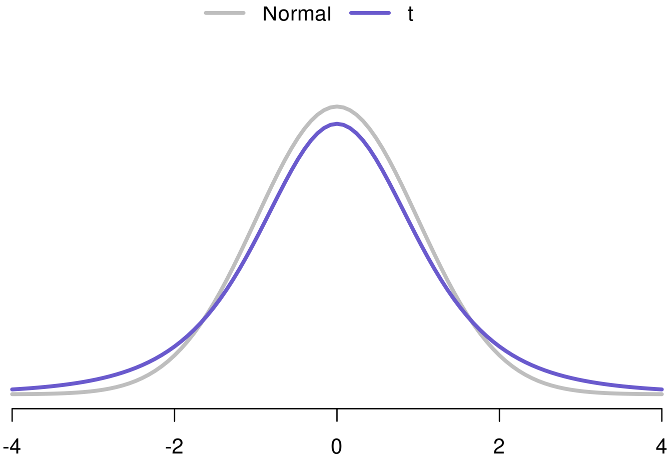
  does not follow a normal curve, but a slightly different curve instead
- This curve is often called *Student's curve*, or the t*-distribution*

Degrees of freedom

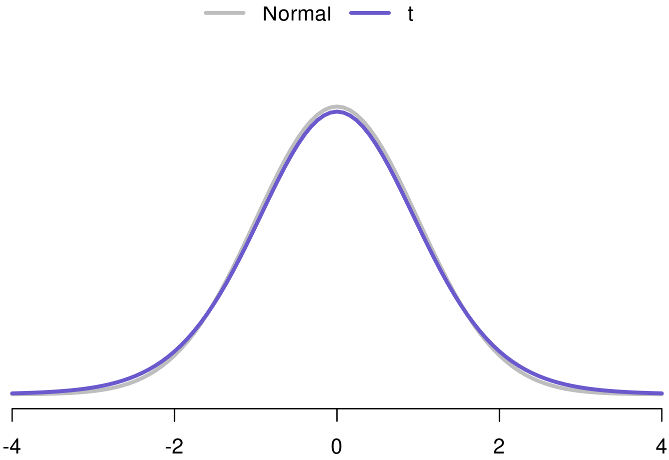
- Actually, there is a Student's curve for every number
- Just as the binomial distribution has parameters n and π, the t distribution has a parameter called the *degrees of freedom*, abbreviated df
- The term "degrees of freedom" refers to the fact that the sum of the deviations (which the SD is based on) has to add up to zero, so not all measurements can vary freely
- In the present context,

$$df = n - 1$$

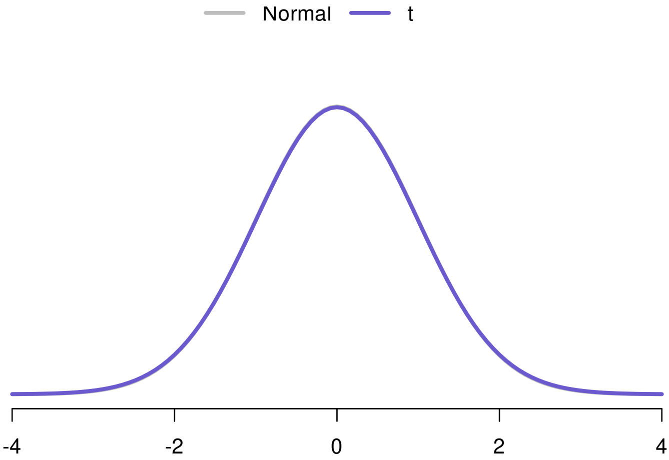
Student's curve vs. the normal curve, $df = 4$



Student's curve vs. the normal curve, $df = 14$



Student's curve vs. the normal curve, $df = 99$



Student's curve and the normal curve

- There are many similarities between the normal curve and Student's curve:
    - Both are symmetric around 0
    - The total area under the curve is equal to 1
    - As the degrees of freedom go up, Student's curve looks more and more similar in shape to the normal curve
- However, there is one very important difference:
    - The tails of Student's curve are thicker than those of the normal distribution
    - This difference can be quite pronounced for small samples

Calculating the area under Student's curve

- Just as for the normal curve, to calculate areas under Student's curve, we will need a computer or a table
- I have added a t-table to the course website
- To accommodate fitting a large number of curves onto a single table, the rows now represent degrees of freedom, and the columns represent two-tailed areas
- So suppose, for example, that we are interested in Student's curve with 10 degrees of freedom
  - If we want the t values that contain the middle 90% of the area, we look under $df = 10$ and $\alpha = 0.1$ and find that the answer is: $(-1.81, 1.81)$
  - If we want to know how much area is outside ± 2, the best we can do with the table is to say that it's between 0.05 and 0.10

Student's curve in R

- For more exact calculations, we need a computer
- In R, we have `pt` and `qt` functions that work very much like `pnorm` and `qnorm`
- So, for the Student's curve with 10 degrees of freedom, what values contain the middle 90%?

  ```
  qt(0.05, df = 10)
  # [1] -1.812461
  ```

  As before, ± 1.81

- How much area is outside ± 2?

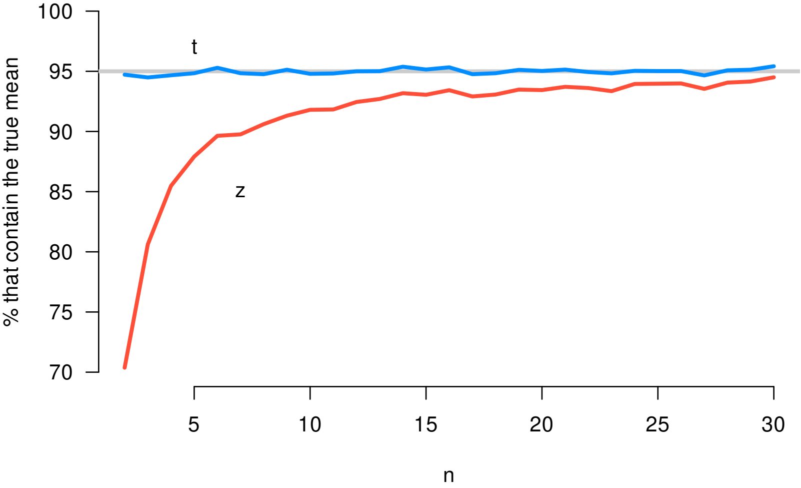
  ```
  2 * pt(-2, df = 10)
  # [1] 0.07338803
  ```

  This is indeed between 0.05 and 0.10, but is more precise

Procedure

- The procedure for carrying out a one-sample t-test is exactly the same as that for the z-test, except for the distribution to which we compare the test statistic:
  1. Calculate the standard error $\mathrm{SE} = \mathrm{SD}/\sqrt{n}$
  2. Calculate the test statistic
     $$t = \frac{\hat{\mu} - \mu_0}{\mathrm{SE}}$$
  3. Calculate the area under the Student's curve with $n - 1$ degrees of freedom curve outside $\pm t$
- As a bit of nomenclature, when applied to paired data, this test is called "the paired t-test"

Does the t-test fix the z-test's problem?



FVC example

- In the cystic fibrosis experiment, the mean difference in FVC reduction (placebo − drug) was 136, with standard deviation 223:
  1. $\mathrm{SE} = 223/\sqrt{14} = 60$
  2. $$\begin{aligned} t &= \frac{136 - 0}{60} \\ &= 2.29 \end{aligned}$$
  3. The area outside ± 2.29 on the Student's curve with 13 degrees of freedom is 0.04
- Our p-value from the z-test was 0.02, which overstates the evidence against the null hypothesis

z-tests vs. t-tests

- For reasonably large sample sizes (> 50), the z- and t-tests are essentially the same
- However, it is difficult to justify z-tests and z-confidence intervals, as their p-values and coverage probabilities are not correct
- So, in practice, no one uses z-tests for one-sample, continuous data
- On the other hand, t-tests are probably the most common type of statistical test on the planet

Procedure for finding confidence intervals

- The procedure for calculating an $x\%$ confidence interval for the mean is similar to calculating an approximate interval for percentages:
  1. Calculate the standard error: $\mathrm{SE} = \mathrm{SD}/\sqrt{n}$
  2. Determine the values of the t-distribution with $n-1$ degrees of freedom that contain the middle $x\%$ of the data; denote these values $\pm t_{x\%}$
  3. Calculate the confidence interval:

$$(\hat{\mu} - t_{x\%}\mathrm{SE}, \hat{\mu} + t_{x\%}\mathrm{SE})$$

FVC example: Patients taking drug

- For patients taking the drug in the cystic fibrosis crossover experiment, the mean reduction in FVC was 160, with standard deviation 197
- Let's calculate a 95% confidence interval for the average reduction in lung function that individuals with cystic fibrosis in the population would be likely to experience over a 25-week period, if they took this drug:
  (1) The standard error is $197/\sqrt{14} = 53$
  (2) The values ± 2.16 contain the middle 95% of Student's curve with 13 degrees of freedom
  (3) Thus, my confidence interval is:

$$\begin{aligned}(160 - 2.16 \times 53, 160 + 2.16 \times 53)\\ = (46, 274)\end{aligned}$$

FVC example: Patients taking placebo

- For patients taking the placebo, the mean reduction in FVC was 296, with standard deviation 297
  (1) The standard error is $297/\sqrt{14} = 79$
  (2) The values ± 2.16 still contain the middle 95% of Student's curve with 13 degrees of freedom
  (3) Thus, my confidence interval is:

$$\begin{aligned}(296 - 2.16 \times 79, 296 + 2.16 \times 79) \\ = (125, 467)\end{aligned}$$

Comparing drug and placebo

- Note that our two confidence intervals, (46, 274) and (125, 467), overlap quite a bit
- On the surface, this would seem to indicate a lack of evidence that the drug is effective
- However, recall that paired designs are powerful ways to reduce noise; constructing separate confidence intervals does not take advantage of this design
- To assess whether drug is more effective than placebo, we should instead construct a single confidence interval for the **difference** in FVC reduction for each patient

FVC example: Difference between two treatments

- The mean difference in reduction in FVC (placebo − drug) was 136, with standard deviation 223
  1. The standard error is $223/\sqrt{14} = 60$
  2. Once again, the values ± 2.16 contain the middle 95% of Student's curve with 13 degrees of freedom
  3. Thus, the confidence interval is:

$$\begin{aligned}(136 - 2.16 \times 60, 136 + 2.16 \times 60)& \\ = (7, 267)&\end{aligned}$$

- This gives us a range of likely values by which taking the drug would slow the decline of lung function in cystic fibrosis patients
- Note that all of the values are positive, indicating benefit from taking the drug, which agrees with the hypothesis test

Remark: CIs for categorical data (percentages)

- When calculating approximate confidence intervals for percentages, you still use the normal curve, not Student's curve
- The reason is that you're not estimating a standard deviation separately from the mean; you just estimate π, so there is still just one source of uncertainty
- Of course, if you use the exact method, you don't need to worry about this

Concerns about t-tests

- The t-test fixes an important problem with the z-test (correcting for the uncertainty in the sample standard deviation), but still relies on the normal approximation
- It is important to keep in mind that the t-test treats the sampling distribution of the mean as if it were a normal curve
- Thus, the t-test relies on the same central limit theorem arguments as the z-test
- If the sample size is small and the data is skewed (or odd-shaped in some other way), the sampling distribution may not look particularly normal, and the t-test will be questionable

Binomial vs. t-tests

- For this reason, people sometimes split continuous data into categories so that they can use the exact results from the previous lectures
- This is exactly what we did when we recorded the number of patients who did better on drug than on placebo
- Recall that when we used the binomial test, we calculated a p-value of 0.06 (as opposed to the t-test p-value of 0.04)
- These are two different p-values, calculated using the same data
- Neither one is wrong, they are just two different ways of performing the hypothesis test, and in fact are testing slightly different hypotheses

Advantages and disadvantages

- Each approach has advantages and disadvantages
- The advantage of the binomial test is that it makes fewer assumptions – is $n = 14$ large enough to rely on the central limit theorem? If not, our p-value from the t-test may be unreliable
- The advantage of the paired t-test is that it is generally more powerful than the binomial test – the binomial test throws away the magnitude of the difference, while the t-test uses all the information
- There are other approaches as well, which strike a balance between these advantages and disadvantages – we will cover them later

Summary

- z-tests fail for continuous data because they ignore uncertainty about SD – this is especially problematic for small sample sizes
- t-tests fix this problem (although they still rely on the CLT):
  - Know how to calculate the one-sample t-test (also known as the paired t-test)
  - Know how to construct confidence intervals for one-sample continuous data using Student's distribution
- Constructing a CI for the difference between two groups is not the same as constructing two CIs, one for each group, then seeing if they overlap