The t-distribution

Patrick Breheny

October 3

Introduction

- So far we've (thoroughly!) discussed how to carry out hypothesis tests and construct confidence intervals for categorical outcomes: success versus failure, life versus death
- This week we'll 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
- It should come as no surprise, then, that there are also big differences in how these data must be analyzed

Notation

- We'll use the following notation:
  - The true population mean is denoted μ
  - The observed sample mean is denoted either $\bar{x}$ or $\hat{\mu}$
  - For hypothesis testing, H_0 is shorthand for the null hypothesis, as in $H_0 : \mu = \mu_0$
- Unlike the case for binary outcomes, we also need some notation for the standard deviation:
  - The true population variance is denoted σ^2 (i.e. σ is the SD)
  - The observed sample variance is denoted $\hat{\sigma}^2$ or s^2:

$$\hat{\sigma}^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1},$$

    with $\hat{\sigma}$ and s the square root of the above quantity

Using the central limit theorem

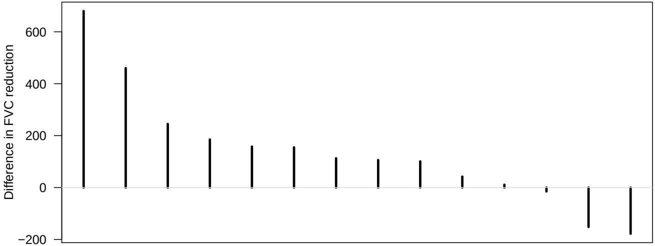
- We've already used the central limit theorem to construct confidence intervals and perform hypothesis tests for categorical data
- The same logic can be applied to continuous data as well, with one wrinkle
- For categorical data, the parameter we were interested in (π) also determined the standard 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

FVC example

- Let's revisit the cystic fibrosis crossover study that we've discussed a few times now, but 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:



- Let's carry out a z-test for this data, plugging in $\hat{\sigma}$ for σ

FVC example (cont'd)

- In the study, the mean difference in reduction in FVC (placebo − drug) was 137, with standard deviation 223
- Performing the z-test of $H_0 : \mu = 0$:
  #1 $\text{SE} = 223/\sqrt{14} = 60$
  #2

$$z = \frac{137 - 0}{60} = 2.28$$

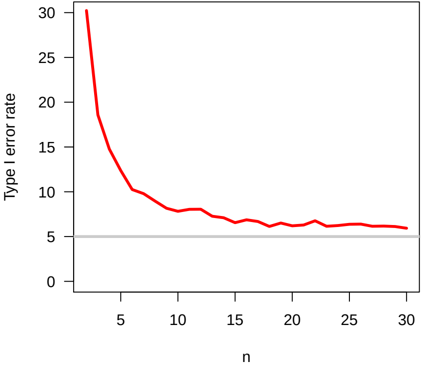
  #3 The area outside ± 2.28 is $2\Phi(-2.28) = 2(0.011) = 0.022$
- This is fairly substantial evidence that the drug helps prevent deterioration in lung function

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?" in using the central limit theorem
- 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 carrying out this test and recording the type I error rate

Simulation results

Using $p < 0.05$ as a rejection rule:



What would a simulation involving confidence intervals 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

Distribution of the sample variance

- Before we get into the business of fixing the z-test, we need to discuss a more basic issue: what does the sampling distribution of the variance look like?
- We have this beautiful central limit theorem describing what the sampling distribution of the mean looks like for *any* underlying distribution
- Unfortunately, there is no corresponding theorem for the sample variance

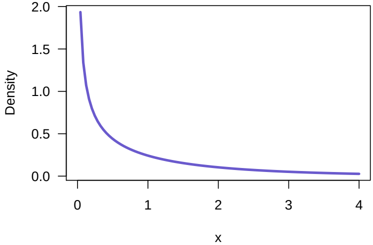
Special case: The normal distribution

- We may, however, consider the important special case of the normal distribution
- If the underlying distribution is normal, we can derive many useful results concerning the sample variance
- Keep in mind, however, that unlike the results we established in the central limit theorem lecture, these results only apply to random variables that follow a normal distribution

The χ^2 distribution

- An important distribution derived from the normal distribution is the χ^2-distribution
- Suppose $Z \sim \mathrm{N}(0, 1)$; then Z^2 is said to follow a χ^2_1 distribution, with pdf:

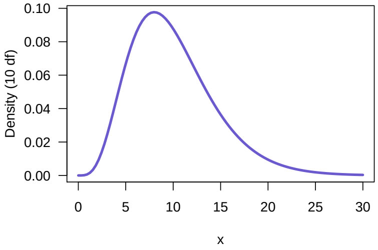
$$f(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}$$



The χ^2 distribution: Degrees of freedom

- An important generalization is to consider sums of squared observations from the normal distribution
- Suppose $Z_1, Z_2, \ldots, Z_p \sim \mathrm{N}(0,1)$ and are mutually independent; then $\sum_{i=1}^p Z_i^2$ is said to follow a chi-squared distribution with p degrees of freedom, denoted χ_p^2:

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{p/2-1} e^{-x/2}$$



Distribution of the sample variance (normal case)

- From the previous slide, it immediately follows that if $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ are mutually independent, then

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n$$

- In other words, letting $\tilde{S}^2 = \sum (x_i - \mu)^2/n$, we have $n\tilde{S}^2/\sigma^2 \sim \chi^2_n$
- It can also be shown (not so immediately) that if $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ are mutually independent, then

$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$$

Independence of mean and variance

- By working out the joint distribution of $\bar{X}$ and $X_2 - \bar{X}, X_3 - \bar{X}, \ldots, X_n - \bar{X}$, we also arrive at the useful conclusion that the sampling distributions of $\bar{X}$ and S^2 are independent
- In other words, for normally distributed variables, the mean and variance have no relationship whatsoever
- This is obviously not true for other distributions – for example, we saw that the binomial distribution has $\mathrm{Var}(X) = n\mathrm{E}(X)(1 - \mathrm{E}(X))$

Distribution of the sample mean (normal case)

- Finally, it is worth mentioning that when a random variable follows a normal distribution, the distribution of its sample mean is exactly normal (i.e., the central limit theorem is an exact result, not an approximation)
- More formally, suppose $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ are mutually independent; then

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim \mathrm{N}(0, 1)$$

Revisiting our earlier test statistic

- When we carried out our z-test from earlier, we looked at the quantity

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

and acted as if it followed a normal distribution

- But of course, it really doesn't: the numerator is normal, but then we're dividing it by another random variable

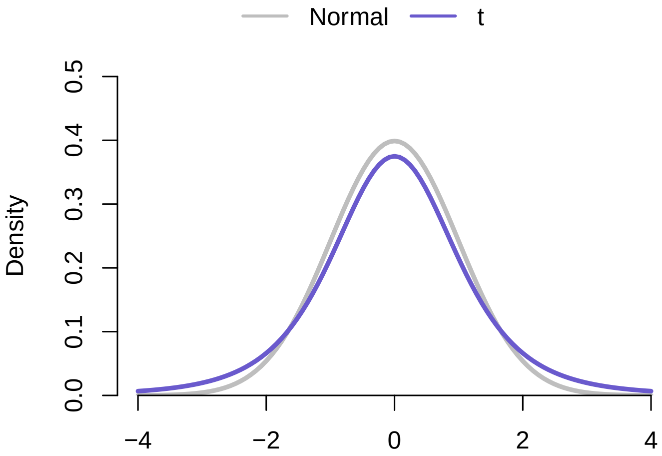
The t-distribution

- The problem of "What is the resulting distribution when you divide one random variable by another?" was studied by a statistician named W. S. Gosset, who showed the following
- Suppose that $Z \sim \mathrm{N}(0,1)$, $X^2 \sim \chi^2_n$, and that Z and X^2 are independent; then

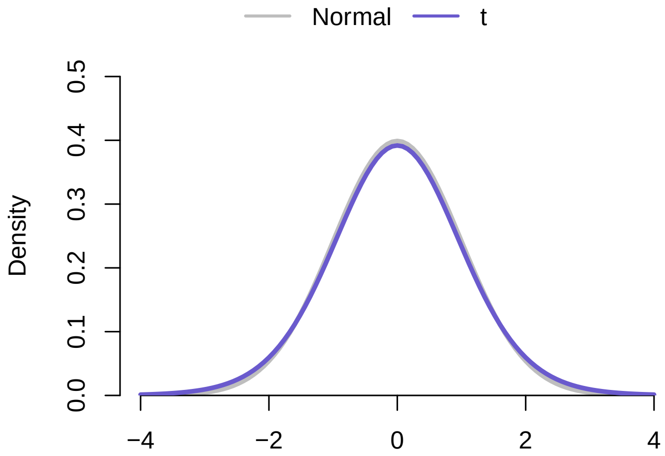
$$\frac{Z}{\sqrt{X^2/n}} \sim t_n,$$

the t*-distribution with* n *degrees of freedom*

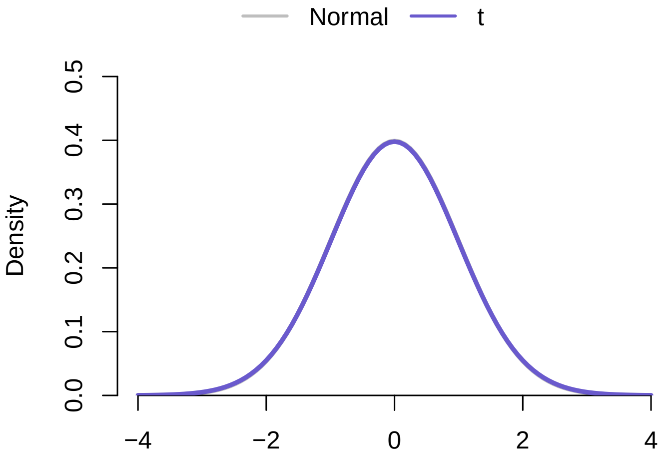
t-distribution vs. normal distribution, $df = 4$



t-distribution vs. normal distribution, $df = 14$



t-distribution vs. normal distribution, $df = 99$



t-distribution vs. normal distribution

- There are many similarities between the normal distribution and t-distribution:
  - Both are symmetric around 0
  - Both have positive support over the entire real line
  - As the degrees of freedom go up, the t-distribution converges to the normal distribution
- However, there is one very important difference:
  - The tails of the t-distribution are thicker than those of the normal distribution
  - This difference can be quite pronounced when df is small

The t-distribution and the sample mean

- Returning to our test statistic for one-sample inference concerning the mean of a continuous random variable, we have the following result:
- Suppose $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ are mutually independent; then

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

- In other words, our test statistic from earlier *does* have a known, well-defined distribution – it's just not $\mathrm{N}(0, 1)$
- Thus, we can still derive hypothesis tests and confidence intervals, we'll just have to use the t-distribution instead of the normal distribution; this will be the subject of the next lecture

Summary

- z-tests fail for continuous data because they ignore uncertainty about SD – this is especially problematic for small sample sizes
- $Z_1, Z_2, \ldots, Z_n \sim \mathrm{N}(0,1) \implies \sum Z_i^2 \sim \chi_n^2$
- $Z \sim \mathrm{N}(0,1), X^2 \sim \chi_n^2$, and $Z \amalg X^2 \implies Z/\sqrt{X^2/n} \sim t_n$
- For $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$,
  - $\sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathrm{N}(0,1)$
  - $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$
  - $\bar{X}$ and S^2 are independent
  - Thus, $\sqrt{n}(\bar{X} - \mu)/S \sim t_{n-1}$