ST 516: Foundations of Data Analytics

t-based procedures for the mean

Recall: estimating a population mean

We have a random sample, $X_1, \ldots, X_n$, of size n from the population. We calculate the sample mean, $\overline{X}$ and wish to perform inferences about the population mean, μ.

Last module we constructed confidence intervals for μ of the form,

$$\overline{X} \pm z_{\alpha/2} \times SE_{\overline{X}}$$

where $SE_{\overline{X}} = \frac{s}{\sqrt{n}}$. We performed hypothesis tests for $\mu = \mu_0$ by comparing

$$\frac{\overline{X} - \mu_0}{SE_{\overline{X}}}$$

to a standard Normal distribution.

Motivation

To construct our tests and confidence intervals, we relied on the result that the sampling distribution of $\overline{X}$ was approximately Normal with mean μ and standard deviation $\frac{s}{\sqrt{n}}$.

The standard deviation should really be $\frac{\sigma}{\sqrt{n}}$ but we don't know σ. Instead we used s as an estimate for σ.

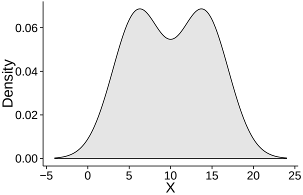
Using an estimate for σ introduces extra uncertainty. This lecture, we are going to investigate how ignoring this extra uncertainty can be problematic in small samples and introduce an approach to remedy it.

σ: the population standard deviation
s: the sample standard deviation

An example

Let's consider an example where our sample size is relatively small, $n = 15$.

Consider the following population distribution (with mean, $\mu = 10$):



Let's look at the sampling distribution for the statistic $t = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$.

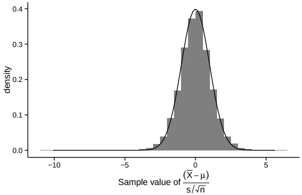
Our intervals and tests from last week, relied on the sampling distribution for this quantity being a Normal with zero mean and standard deviation 1.

An example

We repeat many times by simulation:

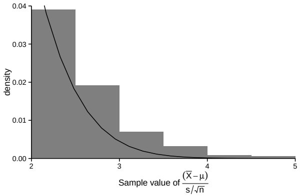
- take a sample of size 15 from the population
- use the sample to calculate t.

This histogram of the resulting statistics is shown below. I've overlaid a Normal curve on our result.



Doesn't look too bad, but let's take a closer look at the extremes

An example



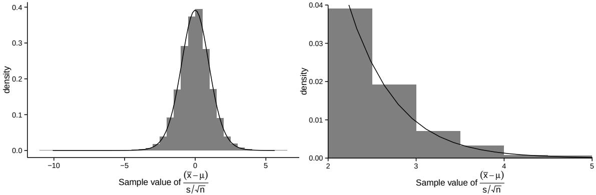
The Normal curve is under estimating how likely it is to see t values that are large (and similarly extreme and negative).

The additional uncertainty due to the estimation of σ results in a sampling distribution with more probability of extreme values than a Normal curve would suggest.

If you consider 95% confidence intervals created the way we did last week ($\overline{X} \pm 1.96 \times SE_{\overline{X}}$), about 92.7% cover the true parameter, not the 95% we would expect.

Enter the t

Turns out there is a better distribution to describe the sampling distribution of $\dfrac{\overline{X}-\mu}{s/\sqrt{n}}$, it's called Student's t-distribution, or the t-distribution for short.



When we base our confidence intervals on our statistic having this distribution they cover 95.2% of the time

A few facts about the t-distribution

The t-distribution is centered around zero.

It has a parameter called the degrees of freedom, which controls the shape of the distribution.

As degrees of freedom increases, the shape of the t-distribution gets closer and closer to the shape of the Normal distribution.

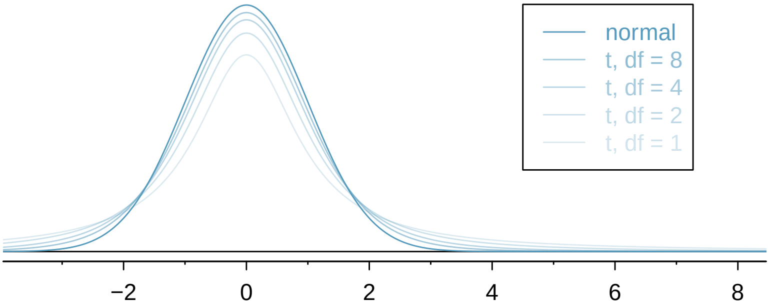


Figure from OpenIntro

In practice

To use the t-based methods for inference on a population mean, the calculations are mostly the same, but the reference distribution changes.

For **hypothesis tests** we calculate the statistic the same way,

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

but to find a p-value (or critical value) we now compare our statistic to the **t-distribution with** $n-1$ **degrees of freedom**.

For a t-based **confidence interval**, the form is the same

$$\overline{X} \pm \text{critical value} \times \frac{s}{\sqrt{n}}$$

but the critical value is now the corresponding quantile from a **t-distribution with** $n-1$ **degrees of freedom**.

Returning to our example

To construct a 95% t based confidence intervals, we need the 0.975 quantile from a t-distribution with 14 d.f.

```r
qt(0.975, df = 14)
```

```
## [1] 2.144787
```

If, in our simulation, we calculate our confidence interval with

$$\overline{X} \pm 2.14 \times \frac{s}{\sqrt{n}}$$

the coverage of our intervals is 95.2%. (Much closer to our claim of 95%)

Approximate versus exact

For small samples using the t-distribution as a model for the sampling distribution of the t-statistic is better than using the Normal distribution because it takes into account the added uncertainty due to estimating σ.

However, the t-distribution still might not be quite the right distribution.

If the sampling distribution of the t-statistic is exactly t-distributed, we say our procedures are **exact**.

This means our 95% confidence intervals will have exactly 95% coverage, and our level 0.05 t-test will reject the null exactly 5% of the time when the null is true.

When procedures aren't exact, but there is some reason to believe they are close to exact, we might label them as **approximate**.

CLT to the rescue

If the population distribution is Normal, then the t-statistic is exactly t-distributed, regardless of the sample size.

However, it's pretty rare (impossible?) to know your population distribution is exactly Normal.

We do know for large samples $\overline{X}$ is approximately Normal, regardless of the population distribution (CLT) and we know that the t-distribution is approximately Normal for reasonably large sample sizes.

So, our t-based procedures should be approximately correct for large samples.

It turns out that *large* here is surprisingly small for populations that aren't too non-Normal. So the t-based methods work very well for small samples even when the population distribution isn't Normal.

We'll demonstrate this in the next module.