Week 3

t Confidence intervals

In the previous lesson, we discussed creating a confidence interval using the CLT.

- They took the form $\text{Estimate} \pm \text{ZQ} \times SE_{\text{Est}}$.

In this lecture, we are going to discuss some methods for small samples.

- Notably, we are going to talk about Student's or Gosset's t distribution and t confidence intervals.
- These intervals are going to be of the form $\text{Estimate} \pm \text{TQ} \times SE_{\text{Est}}$, where TQ is a quantile from the t distribution.

The t distribution has heavier tails than the normal distribution, so these intervals are going to be a little bit wider.

Gosset's t distribution

Invented by William Gosset (under the pseudonym "Student") in 1908.

Has thicker tails than the normal.

Is indexed by degrees of freedom (df).

- Gets more like a standard normal as df gets larger.

It assumes that the underlying data are i.i.d. Gaussian with the result that $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ follows a t distribution with $n - 1$ degrees of freedom.

- Note that, if we had σ in the place of S, the statistic would follow a standard distribution.
- As n increases, the distinction between the t and standard normal distribution decreases.

The confidence interval is given by $\bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}}$, where t_{n-1} is the relevant quantile.

Notes about the t interval:

- Technically assumes that the data are i.i.d. normal, though it is robust to this assumption;
- Works well whenever the distribution of the data is roughly symmetric and mound shaped;
- Paired observations are often analyzed using the t interval by taking differences;
- For large degrees of freedom, t quantiles become the same as standard normal quantiles; therefore this interval converges to the same interval as the CLT yielded;
- For skewed distributions, the spirit of the t interval assumptions is violated:
  - Also, for skewed distributions, it does not make a lot of sense to center the interval at the mean;
  - In this case, consider taking logs or using a different summary like the median;
- For highly discrete data, like binary, other intervals are available.

Example:

```r
g1 <- sleep$extra[1:10]; g2 <- sleep$extra[11:20]
difference <- g2 - g1
mn <- mean(difference); s <- sd(difference); n <- 10

mn + c(-1,1) * qt(.975, n-1) * s / sqrt(n)
t.test(difference)
t.test(g2, g1, paired = TRUE)
t.test(extra ~ I(relevel(group, 2)), paired = TRUE, data = sleep)
```

Independent group t confidence intervals

Suppose that we want to compare the mean blood pressure between two groups in a randomized trial: those who received the treatment to those who received a placebo.

We cannot use the paired t test because the groups are independent and may have different sample sizes:

- We now present methods for comparing independent groups.

Confidence interval:

Therefore a $(1-\alpha)\times 100\%$ confidence interval for $\mu_y - \mu_x$ is

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2,\ 1-\alpha/2}\ S_p\left(\frac{1}{n_x}+\frac{1}{n_y}\right)^{\frac{1}{2}}$$

where $t_{n_x+n_y-2,\ 1-\alpha/2}$ is the $\left[(1-\alpha/2)\times 100\right]^{\text{th}}$ quantile of the t distribution for (n_x+n_y-2) degrees of freedom; and $S_p\left(\frac{1}{n_x}+\frac{1}{n_y}\right)^{\frac{1}{2}}$ is the standard error of the difference between the groups, with n_x and n_y being the number of observations in each group.

The pooled variance estimator is

$$S_p^2 = \frac{(n_x-1)S_x^2 + (n_y-1)S_y^2}{n_x+n_y-2}$$

where S_x^2 and S_y^2 are variances in each group.

- If there is an equal number of observations in each group, S_p^2 would be the simple average of the variances of each group.
- Remember this is assuming a constant variance across the two groups.
    - If there is some doubt, assume a different variance per group, which we will cover later.

Example:

Based on Rosner, Fundamentals of Biostatistics.

- Comparing SBP for 8 oral contraceptive users versus 21 controls:
    - $\bar{X}_{\text{OC}} = 132.86$ mmHg with $s_{\text{OC}} = 15.34$ mmHg.
    - $\bar{X}_{\text{C}} = 127.44$ mmHg with $s_{\text{C}} = 18.23$ mmHg.
- Pooled variance estimate:

```r
sp <- sqrt((7 * 15.34^2 + 20 * 18.23^2) / (8 + 21 - 2))

# interval:
132.86 - 127.44 + c(-1,1) * qt(.975, 27) * sp * (1/8 + 1/21)^.5
```

Unequal variances:

The confidence interval is given by:

$$\bar{Y} - \bar{X} \pm t_{\text{df}} \times \left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} \right)^{\frac{1}{2}}$$

- The relevant statistic does not follow a t distribution, but it can be approximated as one, as long as the degrees of freedom are calculated as:

$$\text{df} = \frac{\left(\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y} \right)^2}{\frac{1}{n_x - 1} \left(\frac{S_x^2}{n_x} \right)^2 + \frac{1}{n_y - 1} \left(\frac{S_y^2}{n_y} \right)^2}$$

- In R, `t.test(..., var.equal = FALSE)`.

Comparing other kinds of data:

For binomial data, there are lots of ways to compare two groups:

- Relative risk, risk difference, odds ratio.
- χ^2 tests, normal approximations, exact tests.

For count data, there are also χ^2 tests and exact tests.

Hypothesis Testing

A null hypothesis is specified that represents the *status quo*, usually labeled H_0.

- The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research r alternative hypothesis H_1.
- The alternative hypotheses are typically of the form $<, >, \neq$.

Note that there are four possible outcomes of our statistical decision process.

Truth	Decide	Result
H_0	H_0	Correctly accept null
H_0	H_1	Type I error
H_1	H_1	Correctly reject null
H_1	H_0	Type II error

A reasonable strategy would reject the null hypothesis if the estimator, say $\bar{X}$, were larger than some constant C.

- Typically, C is chosen so that the probability of a type I error (α), is 0.05 (or some other relevant constant).
- α: type I error rate = probability of rejecting the null hypothesis when, in fact, it is correct.

Example:

A respiratory disturbance index (RDI) of more than 30 events per hour, say, is considered evidence of severe sleep disordered breathing (SDB).

Suppose that in a sample of 100 overweight subjects with other risk factors for sleep disordered breathing at a sleep clinic, the mean RDI was 32 events per hour, with a standard deviation of 10 events per hour.

We might want to test the hypothesis that:

$$H_0 : \mu = 30$$
$$H_1 : \mu > 30$$

where μ is the population mean RDI.

- Standard error of the mean: $\frac{s}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$;
- Then, under H_0, $\bar{X} \sim \mathcal{N}(30, 1)$.

We want to choose C so that $\mathrm{P}(\bar{X} > C; H_0)$ is 5%.

- The 95^{th} percentile of a normal distribution is 1.645 standard deviations from the mean: $C = 30 + 1 \times 1.645 = 31.645$.
  - Then, the probability that a $\mathcal{N}(30, 1)$ is larger than C is 5% (one-tailed test).
  - So, the rule "Reject H_0 when $\bar{X} \geq 31.645$" has the property that the probability of rejection is 5% when H_0 is true (for the given μ_0, σ and n).

In general we do not convert C back to the original scale.

- We would just reject because the z-score (which is how many std. errors the sample mean is above the hypothesized mean) $Z := \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{32 - 30}{10/\sqrt{100}} = 2$ is greater than 1.645.
- Or, whenever $Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} > Z_{1-\alpha}$.

Reconsidering:

Consider our example again. Now, suppose that $n = 16$ (rather than 100).

- The statistic $\frac{\bar{X} - 30}{s/\sqrt{16}}$ follows a t distribution with 15 df under H_0 (rather than a standard normal distribution, due to the reduced sample size).
- Under H_0, the probability that it is larger than the 95^{th} percentile of the t distribution is 5%.
- The 95^{th} of the t distribution with 15 df is 1.7531 (obtained via `qt(.95, 15)`).
- So that our test statistic is now $\frac{32 - 30}{10/\sqrt{16}} = 0.8$.
- Since $t_{\text{test}} = 0.8 < t_{0.95,\ 15=1.7531}$, we now fail to reject.

Discussion:

Consider a court of law: the null hypothesis is that the defendant is innocent.

- We require a standard on the available evidence to reject the null hypothesis (convict).
- If we set a low standard, the we would increase the percentage of innocent people convicted (type I errors); however, we would also increase the percentage of guilty people convicted (correctly rejecting the null).
- If we set a high standard, then we increase the percentage of innocent people let free (correctly accepting the null), while we would also increase the percentage of guilty people let free (type II errors).

Two sided (or tailed) tests

Suppose that we would reject the null hypothesis if in fact the mean was too large or too small.

- That is, we want to estimate the alternative $H_1 : \mu \neq 30$.
- We will reject the test statistic is either too large or too small.
- Then we want the probability of rejecting under the null to be α, split equally as $\alpha/2$ in the upper tail and $\alpha/2$ in the lower tail.
    - For large samples, the z distribution (std. normal) ca be used.
    - For small samples, use the t distribution.
- Thus, we reject if our statistic is larger than $t_{1-\alpha/2,\ \mathrm{df}}$ or smaller than $t_{\alpha/2,\ \mathrm{df}}$.
    - Since both the t and z distributions are symmetric around the mean, this is the same as saying: reject if the absolute value of the test statistic is larger than the $(\alpha/2)^{\text{th}}$, that is, $|t_{\text{test}}| > t_{1-\alpha/2,\ \mathrm{df}}$.

Connections with confidence intervals:

Consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

- Take the set of all possible values for which you fail to reject H_0, this set is a $(1-\alpha) \times 100\%$ confidence interval for μ.
- The same works in reverse: if a $(1-\alpha) \times 100\%$ confidence interval contains μ_0, then we fail to reject H_0.

P-values

- Most common measure of statistical significance.
- Their ubiquity, along with concern over their interpretation and use, makes them controversial among statisticians.

Idea: how unusual is the result we got if the null hypothesis is true?

Approach:

1. Define the hypothetical distribution of a data summary (statistic) when "nothing is going on" (null hypothesis).
2. Calculate the summary/statistic with the data we have (test statistic).
3. Compare what we calculated to our hypothetical distribution and see if the value is "extreme" (p-value).
    - **If the p-value is small, what you are saying is the probability of observing a test statistic as extreme as we saw is low if the null hypothesis is true**.

Definition: The p-value is the probability under the null hypothesis of obtaining evidence as extreme or more extreme than that obtained.

If the p-value is small, then *either* H_0 is true and we have observed a rare event *or* H_0 is false.

- Suppose that you get a t statistic of 2.5 for 15 degrees of freedom testing $H_0 : \mu = \mu_0$ *versus* $H_1 : \mu > \mu_0$.
    - What is the probability of getting a t statistic as large as 2.5.
    - `pt(2.5, 15, lower.tail = FALSE)` produces 0.01225.
    - The probability of seeing evidence as or more extreme than actually obtained under the null hypothesis is 1%.
        - Either the null hypothesis is true and we have seen an exceptionally large t statistic;
        - Or the null hypothesis is false.
    - If the p-value is less than α, reject the null hypothesis.

The attained significance level:

Our test statistic was 2 for $H_0 : \mu = 30$ *vs.* $H_1 : \mu > 30$.

- Notice that we rejected the one sided test when $\alpha = 0.05$, would we reject if $\alpha = 0.01$?
- At any rate, another way to think about the p-value is the smallest value of α for which you would still reject the null hypothesis.
- Because of that they call it the attained significance level.
    - What this means is that the p-value is an extremely convenient test statistic to communicate to people, because when you give it to them, they can test it against whatever alpha level they would like.

Example (binomial):

Suppose a friend has 8 children, 7 of which are girls and none are twins.

If each gender has an independent 50% probability for each birth, what is the probability of getting 7 or more girls out of 8 births?

$$\begin{aligned} H_0 &: p = 0.5 \\ H_1 &: p > 0.5 \end{aligned}$$

According to the binomial distribution:

$$\begin{aligned} \mathrm{P}(X \geq 7) &= \mathrm{P}\big(n = 8, k = 7\big) + \mathrm{P}\big(n = 8, k = 7\big) \\ &= \binom{8}{7} \cdot 0.5^7 \cdot (1 - 0.5)^1 + \binom{8}{8} \cdot 0.5^8 \cdot (1 - 0.5)^0 \\ &= 0.03516 \end{aligned}$$

- In R, it can be obtained by running `pbinom(6, size = 8, prob = 0.5, lower.tail = FALSE)`.
- If we were testing that hypothesis, we would reject at a 5% and a 4% significance level, but we would fail to reject at a type I error rate of 3%.

> Simple trick: If you wanted to test whether p is equal to or different from 0.5 (two-tailed test), then you just calculate the two one-sided p-values.
>
> In this case, the probability of being 7 or larger would be one one-sided p-value, and the probability of being 7 or smaller would be the other one-sided p-value.
>
> You take those two one-sided p-values, you take the smaller one, and you double it. And that is the procedure for getting a two-sided p-value in these exact binomial calculations.

Example (Poisson):

Suppose that a hospital has an infection rate of 10 infections per 100 person/days at risk (rate of 0.1) during the last monitoring period.

- Assume that an infection rate of 0.05 is an important benchmark.
    - If the rate goes above that they would implement some quality control procedures, let us say.
- We are going to assume that the count of infections is Poisson.

$$\begin{aligned} H_0 &: \lambda = 0.05 \ (\lambda_{100} = 5) \\ H_1 &: \lambda > 0.05 \end{aligned}$$

- `ppois(9, 5, lower.tail = FALSE)` produces 0.03183.

- "Remember this little quirk of R: if you want the upper tail and you are doing a discreet distribution, you actually have to drop the number down by one".
- This is the probability of obtaining 10 or more infections if, in fact, the true rate of infections we should have seen on a 100 person/days at risk is 5.
  - It turns out that that is a relatively low probability.
  - It is unlikely for us to have seen as many as 10 infections for a 100 person/days at risk.

`knitr`

File > New File > R Markdown.

Embedding R codes:

````
```{r echo=TRUE}
summary(cars)
```
````

Embedding plots:

````
```{r fig.width=7, fig.height=6}
plot(cars)
```
````

Save as `<file_name>.Rmd`.

You can click `Knit HTML` to view the result (and create a HTML document).

Run `browseURL("<file_name>.html")`.