

Stata Workshop 5:

Merging Data and Inference about Means and Proportions

Tao Wang, SSQL @ Swarthmore, swarthmore.edu/ssql

1. Workshop Objectives

Welcome! In this workshop, we will cover merging two datasets and making inferences about means and proportions.

By the end of this session, you will be able to:

- **Merge** two datasets using unique identifiers.
 - Calculate and interpret a **confidence interval** for a population mean or proportion.
 - Understand the logic of **hypothesis testing**: the null H_0 , alternative H_a , and p -value.
 - Perform **one-sample** t -tests, z -tests, and proportion tests.
 - Perform **two-sample** tests for paired means, independent means, and independent proportions.
 - Distinguish between a command (e.g. `ci`) and its "immediate" calculator version (e.g., `cii`).
-

2. New Commands to be Mastered

Merging datasets: `isid, merge`

Confidence Intervals: `ci means`, `cii means`, `ci prop`, `cii prop`

Hypothesis tests: `ttest`, `ztest`, `prtest`

Hypothesis tests calculators: `ttesti`, `ztesti`, `prtesti`

Repeat command by groups: `bys var:`

3. Workshop Exercises

Exercise 1:

How would you calculate a 95% CI for hourly wages if the population standard deviation was known to be 12.5?

Hint: Use the `invnormal()` function instead of `invttail()` for the critical z-value. (e.g., `invnormal(0.975)`).

```
sum hourwage2
di "A 95% confidence interval for the mean of hourly wages:"
di "Lower bound: " r(mean) - invnormal(0.975) * r(sd) / sqrt(r(N))
di "Upper bound: " r(mean) + invnormal(0.975) * r(sd) / sqrt(r(N))
```

Exercise 2:

Create a 90% confidence interval for the mean of the 'uhrsworkt' variable in the dataset. Use multiple ways to obtain your answers.

Does the estimated interval make sense? If not, how can you resolve the issue?

```
ci mean uhrsworkt if uhrsworkt < 997, level(90)
```

Be careful with how the variable is coded. Note that there are values of 997 and 999.

Exercise 3:

A YouGov Poll conducted between March 16 and 19 of 2024 shows that 36% of the 1,682 U.S. adults in the sample approve of the Supreme Court. Create a 90% confidence interval for the Supreme Court approval rating at the time based on the sample data. Try to use multiple methods and compare your results.

```
di "A 90% confidence interval for the approval rate:"
di .36-invnormal(1-0.1/2)*sqrt(.36*(1-.36)/1682)
di .36+invnormal(1-0.1/2)*sqrt(.36*(1-.36)/1682)
di 1682*.36
cii prop 1682 606, level(90)
cii prop 1682 606, wald level(90)
```

Exercise 4:

Based on the current sample data, can you conclude that on average U.S. workers work for less than 40 hours a week?

Hint: Use variable 'uhrsworkt' or 'ahrsworkt'. Look at the one-sided p-value.

```
ttest uhrsworkt == 40 if uhrsworkt < 997
ttest ahrsworkt == 40 if ahrsworkt < 999
```

Exercise 5:

A YouGov Poll conducted between March 16 and 19 of 2024 shows that 36% of the 1,682 U.S. adults in the sample approve of the Supreme Court. Can you conclude that less than half of U.S. adults approve of the Supreme Court's job performance at the time?

Hint: Test $H_0: p = 0.5$ vs. $H_a: p < 0.5$. Use `prtesti`.

```
prtesti 1682 .36 .5
```