

Introduction to Probability and Applied Statistics

Week 6: Inference

Melissa Sands

Department of Government, Cornell University · Fall 2026

Roadmap

Estimators and their properties

Interval Estimation

Statistical Hypothesis Testing

Example: One-Sample t -test

p -Values, Rejection Regions, and CIs

Substantive versus Statistical Significance

Testing with Two Random Variables

Two Sample t -Test for Equality of Means

Paired t -Test

Summary: Hypothesis Testing

Mean squared error

MEAN SQUARED ERROR

The **mean squared error** of $\hat{\theta}$ is

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] .$$

MSE = Bias² + Variance

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

The definition.

$$= \mathbb{V}[\hat{\theta} - \theta] + (\mathbb{E}[\hat{\theta} - \theta])^2$$

$\mathbb{E}[W^2] = \mathbb{V}[W] + (\mathbb{E}[W])^2$, with $W = \hat{\theta} - \theta$.

$$= \mathbb{V}[\hat{\theta}] + (\mathbb{E}[\hat{\theta}] - \theta)^2$$

θ is a constant.

$$= \mathbb{V}[\hat{\theta}] + \text{Bias}(\hat{\theta})^2$$

The definition of bias.

Properties of an estimator

Property	Definition	For $\bar{X}$
Unbiased	$\mathbb{E}[\hat{\theta}] = \theta$	$\mathbb{E}[\bar{X}] = \mu$
Mean squared error	$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{V} + \text{Bias}^2$	σ^2/n
Efficient	smaller $\mathbb{V}[\hat{\theta}]$ among unbiased	least variance among unbiased weighted averages
Consistent	$\hat{\theta}_n \xrightarrow{p} \theta$	law of large numbers
Asymptotically Normal	$\frac{\hat{\theta}_n - \theta}{\text{SD}[\hat{\theta}_n]} \xrightarrow{d} \mathcal{N}(0, 1)$	central limit theorem

Roadmap

Estimators and their properties

Interval Estimation

Statistical Hypothesis Testing

Example: One-Sample t -test

p -Values, Rejection Regions, and CIs

Substantive versus Statistical Significance

Testing with Two Random Variables

Two Sample t -Test for Equality of Means

Paired t -Test

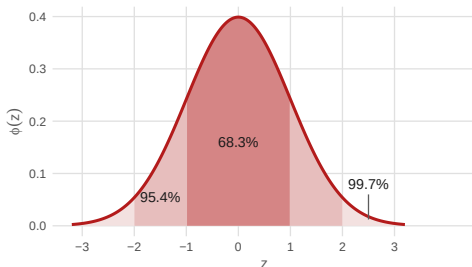
Summary: Hypothesis Testing

From the central limit theorem to the standard normal

CENTRAL LIMIT THEOREM

If $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F$ with mean μ and finite variance σ^2 , then as $n \rightarrow \infty$

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$



From the central limit theorem to the standard normal

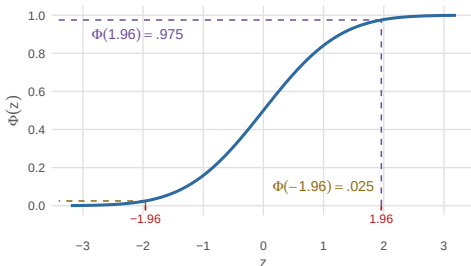
CENTRAL LIMIT THEOREM

If $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F$ with mean μ and finite variance σ^2 , then as $n \rightarrow \infty$

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

So for large n , at every z ,

$$\mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) \approx \Phi(z).$$

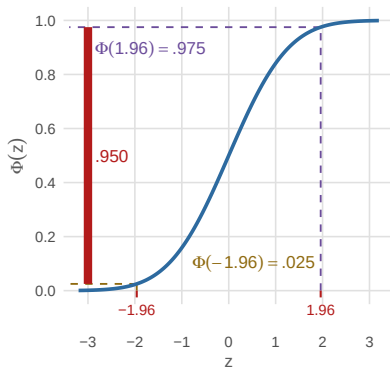
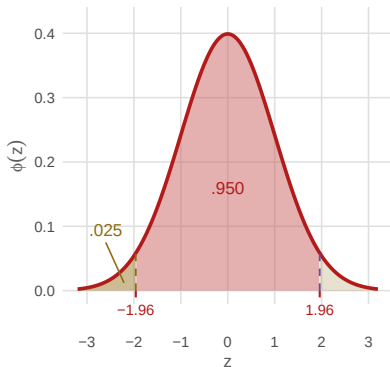


From the central limit theorem to the standard normal

CENTRAL LIMIT THEOREM

If $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F$ with mean μ and finite variance σ^2 , then as $n \rightarrow \infty$

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$



What is Interval Estimation?

A point estimator $\hat{\theta}$ estimates a population parameter θ , such as μ , with a single number, but we might want:

- More information about the uncertainty in our estimate.
- A bound for an estimate, instead of a single number.

Interval estimation accomplishes both of these goals. For a parameter θ that is a single number, such as μ or p , an **interval estimator** takes the following form:

$$[\hat{\theta}_{\text{lower}}, \hat{\theta}_{\text{upper}}]$$

where the lower and upper bounds are random quantities that vary from sample to sample.

An **interval estimate** is a realized value from an interval estimator. We often call these **confidence intervals**.

Confidence Interval: Example with Known σ^2

Suppose the population is $X \sim \mathcal{N}(\mu, \sigma^2 = 1)$ and we draw an iid random sample $X_1, \dots, X_n$ where n may be large or small.

Given the assumption of a normal population with $\sigma^2 = 1$, we know that the sampling distribution of the sample average is:

$$\bar{X} \sim \mathcal{N}(\mu, 1/n)$$

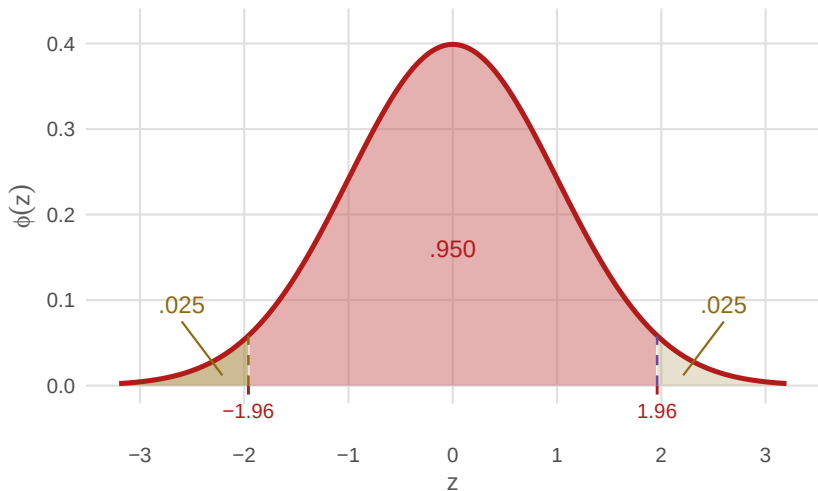
And the standardized sample average follows a standard normal distribution:

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

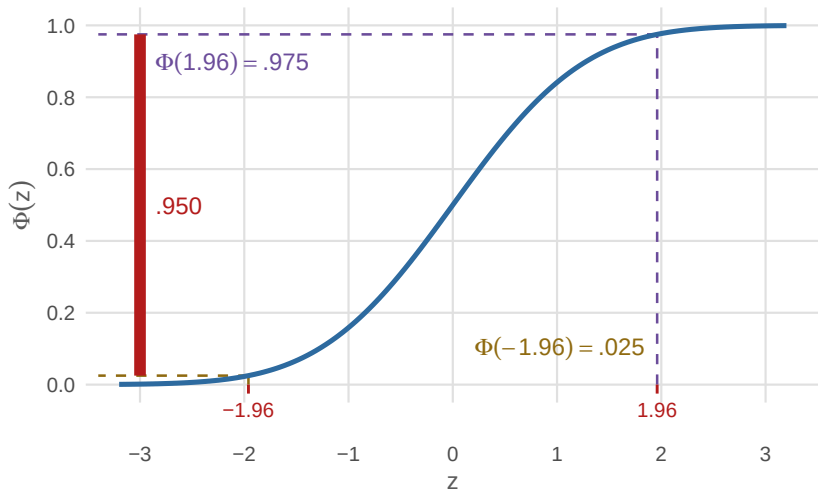
So we know that:

$$\mathbb{P}\left(-1.96 < \frac{\bar{X} - \mu}{1/\sqrt{n}} < 1.96\right) = .95 \quad \text{Why?}$$

Remember the Standard Normal Distribution



Remember the Standard Normal Distribution



Constructing a Confidence Interval ($\sigma^2 = \text{known}$)

Suppose the population is $X \sim \mathcal{N}(\mu, \sigma^2 = 1)$.

So we know that:

$$\mathbb{P}\left(-1.96 < \frac{\bar{X} - \mu}{1/\sqrt{n}} < 1.96\right) = .95$$

Re-arranging yields:

$$\mathbb{P}\left(\bar{X} - 1.96/\sqrt{n} < \mu < \bar{X} + 1.96/\sqrt{n}\right) = .95$$

So the probability that the confidence interval estimator

$$[\bar{X} - 1.96/\sqrt{n}, \bar{X} + 1.96/\sqrt{n}]$$

contains the true population mean μ is .95 or 95%.

We use $\bar{X} \pm 1.96/\sqrt{n}$.

Interpreting a Confidence Interval

Suppose the population is $X \sim \mathcal{N}(\mu, \sigma^2 = 1)$.

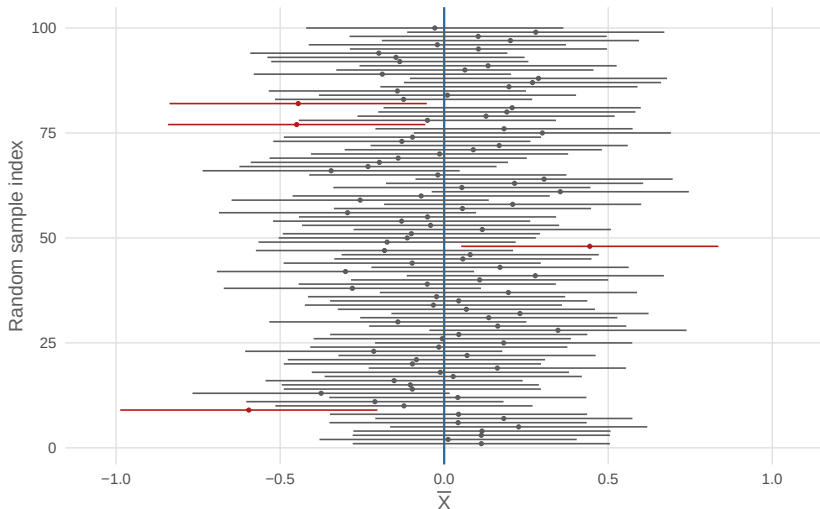
Imagine we implement the interval estimator $\bar{X} \pm 1.96/\sqrt{n}$ for a particular sample and obtain [2.5, 4].

Does this mean that there is a .95 probability that the true parameter value μ lies between these two numbers? **No!**

Confidence intervals are easy to construct but harder to interpret:

- Each confidence interval estimate from a particular sample either contains μ or not (we never know).
- All we can say is that if we were to repeat the interval estimator over repeated samples, 95% of the time the constructed confidence intervals will cover μ .
- We therefore refer to .95 as the **coverage probability**.

Confidence Intervals (100 Samples, $\mathcal{N}(\mu = 0, \sigma^2 = 1)$)



Constructing a Confidence Interval ($\sigma^2 = \text{known}$)

The same principle easily extends to normal populations with any known variance $\mathcal{N}(\mu, \sigma^2)$.

In general, if the population is $\mathcal{N}(\mu, \sigma^2)$ and we draw an iid random sample $X_1, \dots, X_n$ where n may be large or small, a .95 confidence interval for the population mean μ is constructed by:

$$\mathbb{P}\left(\bar{X} - 1.96 \sigma / \sqrt{n} < \mu < \bar{X} + 1.96 \sigma / \sqrt{n}\right) = .95$$

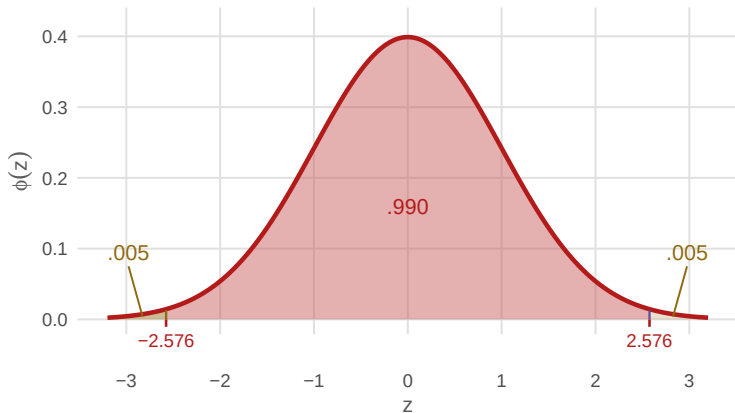
We use:

$$\bar{X} \pm 1.96 \sigma / \sqrt{n}$$

$$\bar{X} \pm 1.96 \underbrace{\text{SD}[\bar{X}]}_{\text{standard error}}$$

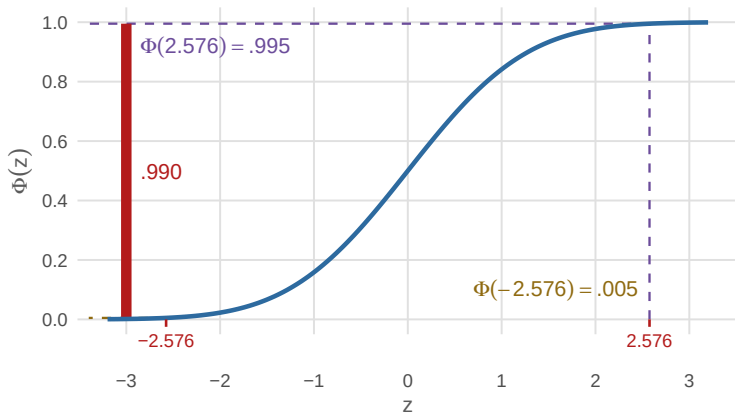
Constructing a Confidence Interval ($\sigma^2 = \text{known}$)

What if we want to change the coverage probability to .99?



Constructing a Confidence Interval ($\sigma^2 = \text{known}$)

What if we want to change the coverage probability to .99?



Constructing a Confidence Interval ($\sigma^2 = \text{known}$)

If the population is $\mathcal{N}(\mu, \sigma^2)$ and we draw an iid random sample $X_1, \dots, X_n$ where n may be large or small, a .99 confidence interval for the population mean μ is constructed by:

$$\mathbb{P}\left(\bar{X} - 2.576 \sigma / \sqrt{n} < \mu < \bar{X} + 2.576 \sigma / \sqrt{n}\right) = .99$$

We use:

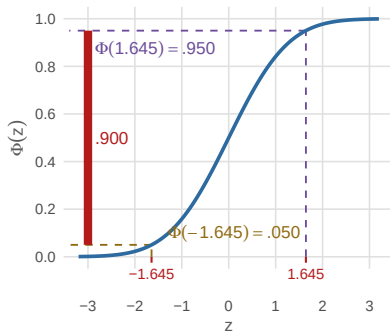
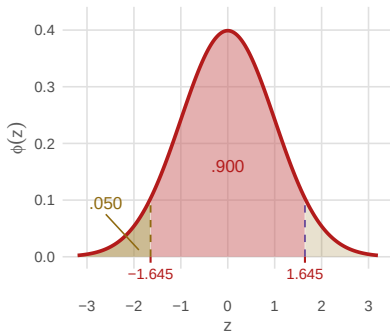
$$\bar{X} \pm 2.576 \sigma / \sqrt{n}$$

$$\bar{X} \pm 2.576 \underbrace{\text{SD}[\bar{X}]}_{\text{standard error}}$$

Constructing a Confidence Interval ($\sigma^2 = \text{known}$)

For coverage $1 - \alpha$ we use $\bar{X} \pm z_{\alpha/2} \sigma / \sqrt{n}$.

α	0.01	0.05	0.10
$z_{\alpha/2}$	2.576	1.960	1.645



Constructing a Confidence Interval ($\sigma^2 = \text{known}$)

DEFINITION

Coverage probability is the probability that an interval estimator contains the true value of the parameter:

$$\mathbb{P}(\hat{\theta}_{\text{lower}} \leq \theta \leq \hat{\theta}_{\text{upper}}) = 1 - \alpha$$

$$\mathbb{P}\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$z_{\alpha/2}$ is the $100(1-\alpha/2)$ th percentile of the standard normal distribution.

Example: Feeling Thermometer Scores (ANES)

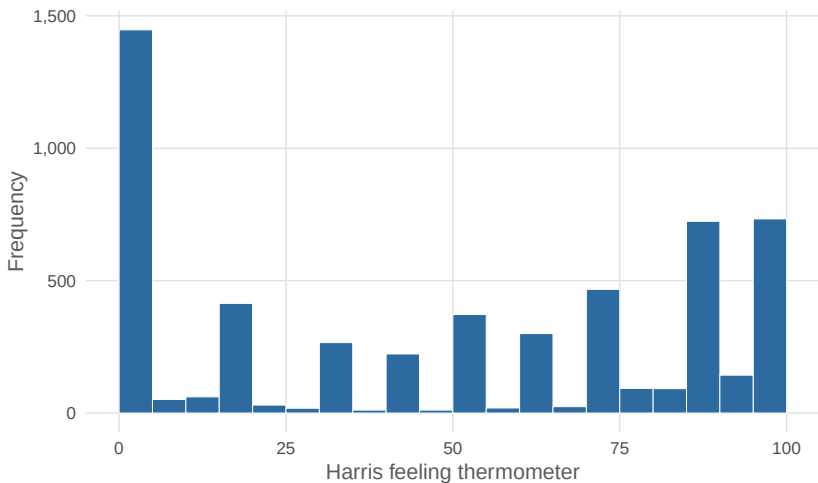
```
=====
V241156  PRE: FEELING THERMOMETER: KAMALA HARRIS,
        DEMOCRATIC PRESIDENTIAL CANDIDATE
=====
```

```
Question:      (How would you rate:) Kamala Harris
Value labels:  -9. Refused
               -8. Don't know
```

```
=====
V241157  PRE: FEELING THERMOMETER: DONALD TRUMP,
        REPUBLICAN PRESIDENTIAL CANDIDATE
=====
```

```
Question:      (How would you rate:) Donald Trump
Value labels:  -9. Refused
               -8. Don't know
               -1. Inapplicable
```

Feeling Thermometer Scores for Kamala Harris



Feeling Therm Example: $X \sim \mathcal{N}(\mu = ?, \sigma^2 = \text{known})$

Suppose we assume that Harris feeling thermometer scores in the population are normally distributed, and we know (somehow) that $\sigma = 37.5$.

The ANES sample has $n = 5,499$ and is randomly drawn, and we find that the sample mean is $\bar{X} = 47.2$.

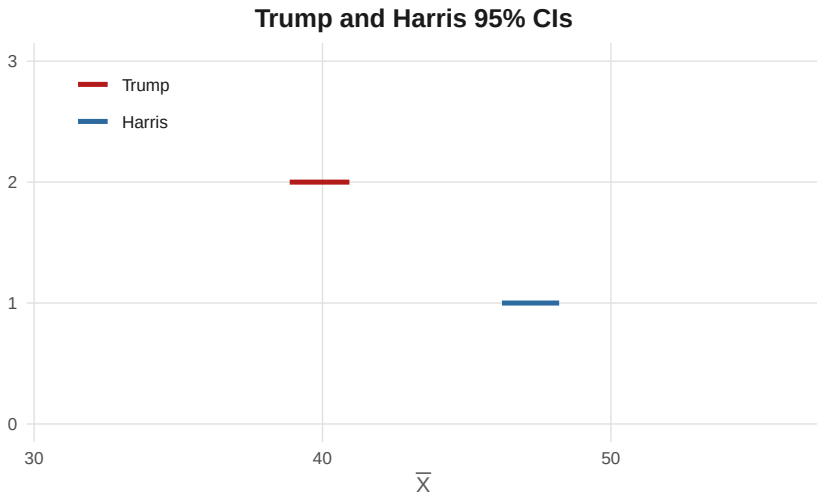
Q: What is the .95 confidence interval for Harris's average feeling thermometer score?

The .95 CI is given by: $\mathbb{P}(\bar{X} - 1.96 \sigma / \sqrt{n} < \mu < \bar{X} + 1.96 \sigma / \sqrt{n}) = .95$

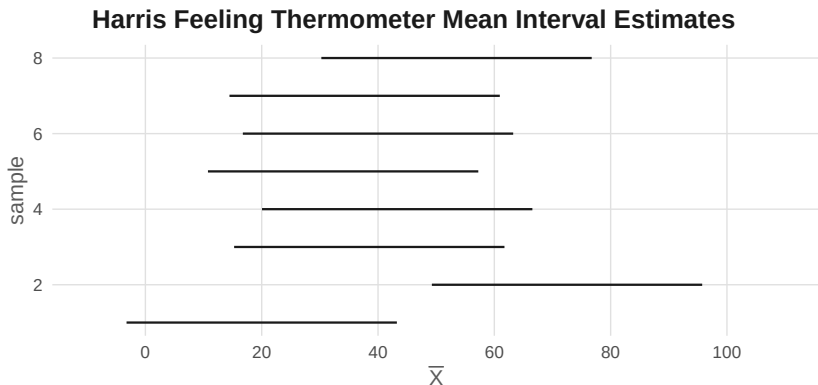
$$\mathbb{P}\left(47.2 - 1.96 \cdot 37.5 / \sqrt{5,499} < \mu < 47.2 + 1.96 \cdot 37.5 / \sqrt{5,499}\right) = .95$$

$$\mathbb{P}(46.2 < \mu < 48.2) = .95$$

Example: Harris Feeling Therm Interval Estimates



Sampling Distribution for Feeling Therm Interval Estimator



Feeling Therm Example: $X \sim \mathcal{N}(\mu = ?, \sigma^2 = ?)$

Suppose we model Harris feeling thermometer scores as normally distributed with σ unknown. Recall that if $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n) \quad \text{and so} \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

Question: Why can't our previous interval be used?

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

We need an estimator of σ^2 .

Estimator for the Population Variance σ^2

Assuming an iid sample, here are two possible estimators of the population variance:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \qquad s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Which one do we prefer?

After some tedious algebra,* you can show that $\mathbb{E}[\hat{\sigma}^2] = \left(\frac{n-1}{n}\right) \sigma^2$ (biased) and that $\mathbb{E}[s^2] = \left(\frac{1}{n-1}\right) (n-1) \sigma^2 = \sigma^2$ (unbiased).

What happens in large samples? $\hat{\sigma}^2 \approx s^2$ and both will converge in probability to σ^2 (both estimators are consistent).

*See `handout_estimation.pdf` for the derivation.

Estimating σ and the SE

Using the unbiased estimator of the population variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

the sample standard deviation is $s = \sqrt{s^2}$ and we can plug in s for σ .

So our estimator for the standard error is now

$$\widehat{SE}[\bar{X}] = \frac{s}{\sqrt{n}} \quad \text{instead of} \quad SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}$$

and to construct confidence intervals we can now use $\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim ?$

The problem is that now a constant σ is replaced with a random variable s , so the standardized sample mean will not exactly follow a standard normal distribution.

Sampling Distribution with Estimated s

Instead the standardized mean follows a so-called **Student t -distribution** with $n - 1$ **degrees of freedom**:

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

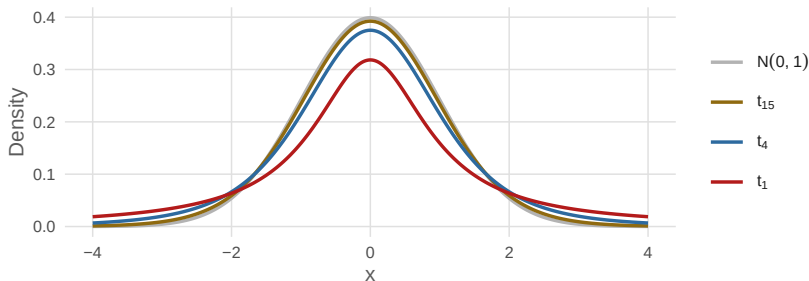
The reason for this is that the sample mean follows a normal distribution and the sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ satisfies $(n - 1)s^2/\sigma^2 \sim \chi_{n-1}^2$, a chi-square distribution, independent of $\bar{X}$.

The t -distribution arises as a standard normal divided by the square root of an independent chi-square over its degrees of freedom.

Use this t -value to construct CIs when σ^2 is unknown. The CIs constructed from the t -distribution will retain the correct coverage.

The t Distribution

1. The mean of a t_ν random variable is 0 as long as $\nu > 1$, and its variance is $\nu/(\nu - 2)$ as long as $\nu > 2$.
2. Has fatter tails than the normal for small ν , but as $\nu \rightarrow \infty$ we have $t_\nu \xrightarrow{d} \mathcal{N}(0, 1)$.
3. In R `dt(x, df)`, `pt(q, df)`, `qt(p, df)`, `rt(n, df)` give the PDF, CDF, quantiles, and random draws.



$(1 - \alpha)$ t -Intervals

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

$$\mathbb{P}\left(-t_{n-1,\alpha/2} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) = 1 - \alpha$$

$$\mathbb{P}\left(\bar{X} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Notice: $t_{n-1,\alpha/2}$ refers to the $100(1 - \alpha/2)$ th percentile of the t_{n-1} distribution. We usually construct the $(1 - \alpha)$ confidence interval with the following formula:

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \quad \text{or simply} \quad \bar{X} \pm t_{n-1,\alpha/2} \widehat{SE}[\bar{X}]$$

For a 95% confidence interval, $t_{n-1,\alpha/2}$ is often close to 2.

Feeling Therm Example: $X \sim ?(\mu = ?, \sigma^2 = ?)$

Suppose we do not assume a distribution for Trump's feeling thermometer score. Recall that if $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} ?(\mu = ?, \sigma^2 = ?)$, then it follows from the central limit theorem that the sampling distribution of the sample mean is approximately

$$\bar{X} \approx \mathcal{N}(\mu, \sigma^2/n)$$

and as $s_n \xrightarrow{P} \sigma$ the standardized mean follows

$$\frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Therefore, our confidence intervals constructed from the standard normal (or t -distribution) will have approximately correct coverage in large samples. This process is called **normal approximation**.

Margin of Error

A poll's **margin of error** is the half-width of its 95% confidence interval:

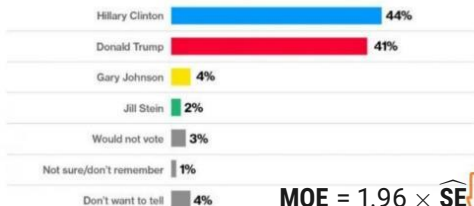
$$\text{MOE} = 1.96 \times \widehat{\text{SE}}, \quad \bar{X} \pm \text{MOE}$$

$\hat{p} = \bar{X}$ for a 0/1 variable.

Margin of Error

2016 Presidential Race

If the election were held today and the candidates were Clinton for the Democrats, Trump for the Republicans, Johnson for the Libertarians, and Stein for the Green Party, for whom would you vote?



Source: Bloomberg Politics national poll conducted by Selzer & Co. Nov. 4-6, 2016, among 799 likely voters. Margin of error +/- 3.5 percentage points. Respondents who already voted were asked whom they chose. Totals may not add to 100% due to rounding.

$$\begin{aligned} CI_{\text{Clinton}} &= (44 - 3.5, 44 + 3.5) = (40.5, 47.5) \\ CI_{\text{Trump}} &= (41 - 3.5, 41 + 3.5) = (37.5, 44.5) \end{aligned} \left. \vphantom{\begin{aligned} CI_{\text{Clinton}} \\ CI_{\text{Trump}} \end{aligned}} \right\} \text{do the intervals overlap?}$$

What Can We Learn from Overlap?

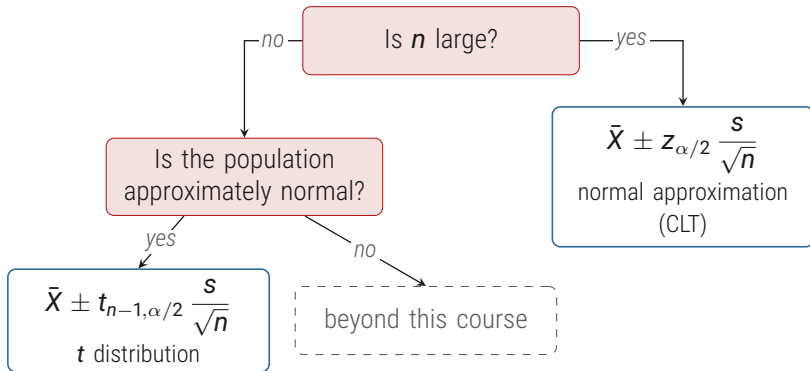
- Two 95% intervals that **don't overlap**: the difference is statistically significant at $\alpha = 0.05$. If anything the test is conservative.
- Two intervals that **overlap**: the amount of overlap matters. If one interval contains the other's point estimate, the difference is not statistically significant.

Rule of thumb: in a single poll, a lead between two candidates is statistically significant at $\alpha = 0.05$ only if it is larger than about twice the reported MOE.

Here: lead = $44 - 41 = 3$ points; $2 \times \text{MOE} = 7$ points. The lead is not statistically significant.

Why twice: the variance of the lead is the sum of the two shares' variances plus twice their (negative) covariance. When the two candidates split nearly all the vote, the SE of the lead is about twice the SE of one share.

Summary: Interval Estimation



Roadmap

Estimators and their properties

Interval Estimation

Statistical Hypothesis Testing

Example: One-Sample t -test

p -Values, Rejection Regions, and CIs

Substantive versus Statistical Significance

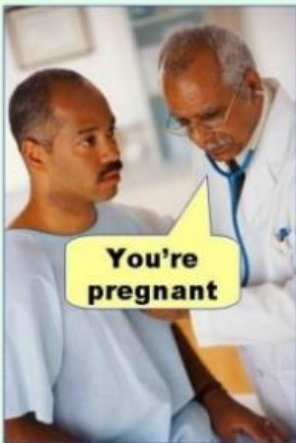
Testing with Two Random Variables

Two Sample t -Test for Equality of Means

Paired t -Test

Summary: Hypothesis Testing

Type I and Type II Errors



Type I error (false positive)



Type II error (false negative)

Statistical Hypothesis Testing: The Trial Analogy

Often we are interested in testing hypotheses.

Suppose we must decide whether to convict or acquit a defendant based on evidence presented at a trial. There are four possible outcomes.

Decisions and Outcomes

		State of the World	
		Guilty	Innocent
Decision	Convict	Correct	Type I Error
	Acquit	Type II Error	Correct

Our goal is to limit the probability of error.

Statistical Hypothesis Testing: The Trial Analogy

Often we are interested in testing hypotheses.

Suppose we can somehow model the probabilities for the various outcomes conditional on the true state of the world.

Probabilities given the true state of the world

		State of the World	
		Guilty	Innocent
Decision	Convict	$1 - \beta$	α
	Acquit	β	$1 - \alpha$

We would like α and β to be small, but it may be difficult to achieve both goals.

Statistical Hypothesis Testing: Step by Step

1. We start by specifying a particular **null hypothesis**. For example: H_0 : The defendant is innocent.

H_0 is presumed to be true unless the data/evidence strongly suggest otherwise and we may be willing to reject the null hypothesis in favor of an **alternative hypothesis**.

Probabilities given the true state of the world

		State of the World	
		Alternative Hypothesis	Null Hypothesis
Decision	Reject	$1 - \beta$	α
	Fail to Reject	β	$1 - \alpha$

2. Next we pick a small **significance level** (also called the level of the test) $\alpha = \mathbb{P}(\text{Reject } H_0 \mid H_0)$ (e.g. 5%), and then minimize β given this constraint.

Statistical Hypothesis Testing: Step by Step

1. We start by specifying a particular **null hypothesis**. For example: H_0 : The defendant is innocent.

H_0 is presumed to be true unless the data/evidence strongly suggest otherwise and we may be willing to reject the null hypothesis in favor of an **alternative hypothesis**.

Probabilities given the true state of the world

		State of the World	
		Alternative Hypothesis	Null Hypothesis
Decision	Reject	$1 - \beta$	α
	Fail to Reject	β	$1 - \alpha$

If $\alpha = 0.05$ we are willing to falsely reject H_0 5% of the time in order to detect deviations from H_0 .

Statistical Hypothesis Testing: Step by Step

3. We choose a **test statistic** t which is some function of the sample and the null hypothesis (and may provide evidence against the null hypothesis). Once we compute the test, we obtain a realized value of this test statistic, t_{obs} .
4. We derive the **null distribution**, which is the sampling distribution for the test statistic t assuming that the null hypothesis is true. The null distribution allows us to partition the possible values of t into those for which the null hypothesis is rejected, the so-called **rejection** or **critical region**, and those for which it is not.
5. Using the null distribution, we look up the **critical value** t_{crit} that corresponds to our chosen significance level α . This critical value determines the rejection region.
6. We finally plug in the numbers and obtain the realized value of t , which we call t_{obs} . We then check if t_{obs} falls in the rejection region or not, and either reject H_0 or fail to reject H_0 .

Roadmap

Estimators and their properties

Interval Estimation

Statistical Hypothesis Testing

Example: One-Sample t -test

p -Values, Rejection Regions, and CIs

Substantive versus Statistical Significance

Testing with Two Random Variables

Two Sample t -Test for Equality of Means

Paired t -Test

Summary: Hypothesis Testing

Harris Feeling Therm Example: Test about Population Mean

Let μ_H be the expected value of Kamala Harris's feeling thermometer score in the population.

Let's assume we want to test whether μ_H is equal to some "null" value of the parameter μ_0 , for example $\mu_H = 47$.

We need to start by specifying a null hypothesis and our alternative hypothesis. Here are two possible hypothesis tests:

1. $H_0 : \mu_H = 47$ and $H_1 : \mu_H \neq 47$ (two-sided alternative)
2. $H_0 : \mu_H = 47$ and $H_1 : \mu_H > 47$ (one-sided alternative)

Also we fix the significance level at $\alpha = 0.05$.

Test Statistics

We choose the standardized version of the sample mean as our test statistic. This test statistic is called the **one-sample t -statistic** or **t -score**:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

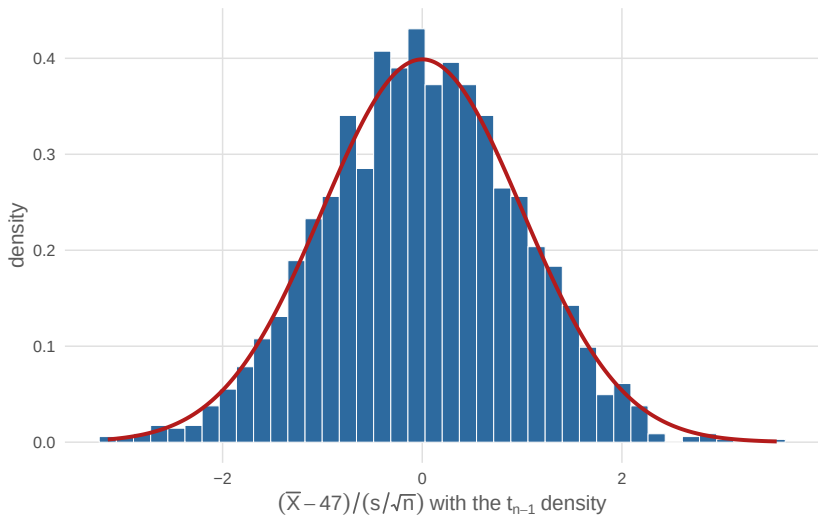
where μ_0 is the “null” value of the parameter μ_H . In our case $\mu_0 = 47$.

Every t -statistic has this form:

$$t = \frac{\text{estimate} - \text{null value}}{\widehat{\text{SE}}}$$

Being a function of the sample and H_0 , this t -statistic has a sampling distribution under the null called the **null distribution**, which we can use to determine the probability of observing a realized value of $t = t_{\text{obs}}$.

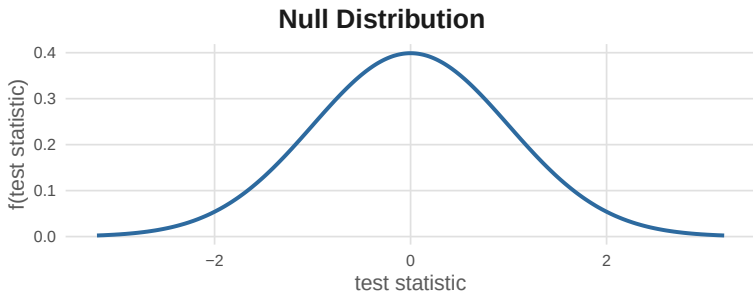
Where Does the Null Distribution Come From?



The Null Distribution for the t -Statistic

Assuming a null value of $\mu_H = \mu_0 = 47$ and a large enough n , the **null distribution** of the t -statistic is approximately

$$\frac{\bar{X} - 47}{s/\sqrt{n}} \approx t_{n-1}$$



The Null Distribution and Critical Values

Assuming a null value of $\mu_H = \mu_0 = 47$ and a large enough n , the null distribution of the t -statistic is approximately

$$\frac{\bar{X} - 47}{s/\sqrt{n}} \approx t_{n-1}$$

Obtain the critical value t_{crit} :

- For a two-sided test t_{crit} is the $100(1 - \alpha/2)$ th percentile of the t_{n-1} distribution, i.e. find t_{crit} such that $\mathbb{P}(|t| > t_{\text{crit}} \mid H_0) = \alpha$.

This t_{crit} is the same $t_{n-1, \alpha/2}$ we used for the confidence interval.

- For a one-sided test t_{crit} is the $100(1 - \alpha)$ th percentile and we want $\mathbb{P}(t > t_{\text{crit}} \mid H_0) = \alpha$.

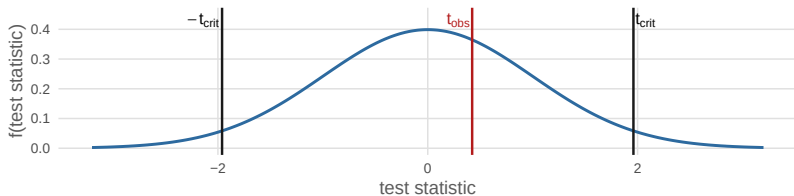
With $n = 5,499$ and $\alpha = 0.05$ the critical value is $t_{\text{crit}} = 1.96$ for a two-sided test and $t_{\text{crit}} = 1.65$ for a one-sided test.

Plugging in the sample mean and standard deviation gives $t_{\text{obs}} = 0.42$, which we compare to t_{crit} to decide whether to reject H_0 .

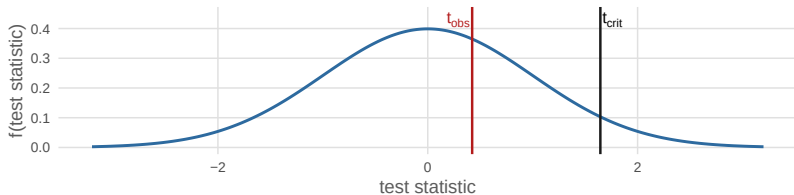
Null Distribution for Harris Feeling Therm and Rejection Regions

For our two-sided test: $|t_{\text{obs}}| > t_{\text{crit}}$. For our one-sided test: $t_{\text{obs}} > t_{\text{crit}}$.

Two Sided Rejection Region ($\alpha = 5\%$)



One Sided Rejection Region ($\alpha = 5\%$)



One-Sample t -Test in R

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} \approx t_{n-1} \text{ with } \mu_0 = 47$$

```
t.test(x = harris_ft, mu = 47)
```

One Sample t-test

```
data: harris_ft
```

```
t = 0.42338, df = 5498, p-value = 0.672
```

```
alternative hypothesis: true mean is not equal to 47
```

```
95 percent confidence interval:
```

```
46.22296 48.20512
```

```
sample estimates:
```

```
mean of x
```

```
47.21404
```

One-Sample t -Test in R

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} \approx t_{n-1} \text{ with } \mu_0 = 46$$

```
t.test(x = harris_ft, mu = 46)
```

One Sample t-test

```
data:  harris_ft
t = 2.4014, df = 5498, p-value = 0.01636
alternative hypothesis: true mean is not equal to 46
95 percent confidence interval:
 46.22296 48.20512
sample estimates:
mean of x
 47.21404
```

Roadmap

Estimators and their properties

Interval Estimation

Statistical Hypothesis Testing

Example: One-Sample t -test

p -Values, Rejection Regions, and CIs

Substantive versus Statistical Significance

Testing with Two Random Variables

Two Sample t -Test for Equality of Means

Paired t -Test

Summary: Hypothesis Testing

p-Value

The **p-value** is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

The **p-value** summarizes how much evidence the data provide against the null hypothesis: it is the largest significance level α at which we could carry out the test and still fail to reject H_0 .

“Extreme” is defined by the alternative hypothesis. For a one-sided test with $H_1 : \mu_H > 47$ we have:

$$p\text{-value} = \mathbb{P}(t \geq t_{\text{obs}} \mid \mu_H = 47) = 1 - F_{t_{n-1}}(0.42) = 0.336$$

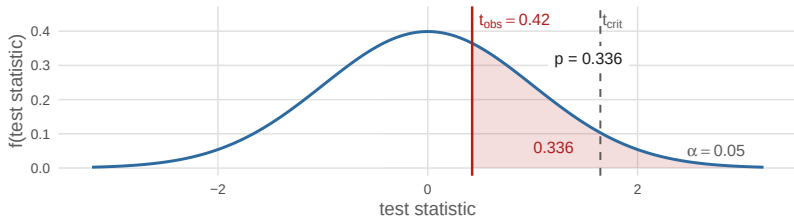
What does this mean? **Weak evidence against H_0 !** If H_0 were true we would expect a value as large as t_{obs} about 34% of the time.

For a two-sided test with $H_1 : \mu_H \neq 47$ we have:

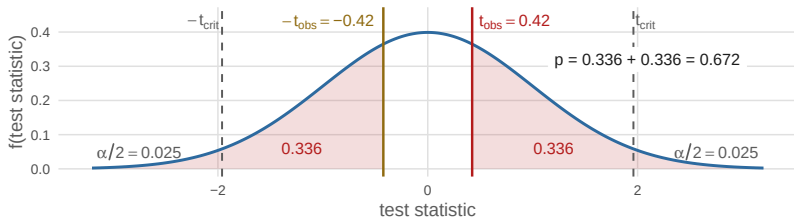
$$\begin{aligned} p\text{-value} &= \mathbb{P}(t \geq t_{\text{obs}} \text{ or } t \leq -t_{\text{obs}} \mid \mu_H = 47) \\ &= 2 \cdot \mathbb{P}(t \geq t_{\text{obs}} \mid \mu_H = 47) = 2 \cdot 0.336 = 0.672 \end{aligned}$$

One and Two Sided p -values

One Sided p -value



Two Sided p -value



α Rejection Regions and $1 - \alpha$ CIs

CIs can be used to test against two-sided alternatives. If we construct a $1 - \alpha$ CI for μ , then we can reject $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at the α level for all hypothesized values of μ_0 that are outside of the constructed CI.

Example: .95 CI for μ and rejection regions for two-sided test against $H_0 : \mu = 47$.

```
t.test(x = harris_ft, mu = 47)
```

One Sample t-test

```
data: harris_ft
```

```
t = 0.42338, df = 5498, p-value = 0.672
```

```
alternative hypothesis: true mean is not equal to 47
```

```
95 percent confidence interval:
```

```
46.22296 48.20512
```

```
sample estimates:
```

```
mean of x
```

```
47.21404
```

α Rejection Regions and $1 - \alpha$ CIs

CIs can be used to test against two-sided alternatives. If we construct a $1 - \alpha$ CI for μ , then we can reject $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at the α level for all hypothesized values of μ_0 that are outside of the constructed CI.

Example: .95 CI for μ and rejection regions for two-sided test against $H_0 : \mu = 46$.

```
t.test(x = harris_ft, mu = 46)
```

One Sample t-test

```
data: harris_ft
```

```
t = 2.4014, df = 5498, p-value = 0.01636
```

```
alternative hypothesis: true mean is not equal to 46
```

```
95 percent confidence interval:
```

```
46.22296 48.20512
```

```
sample estimates:
```

```
mean of x
```

```
47.21404
```

Roadmap

Estimators and their properties

Interval Estimation

Statistical Hypothesis Testing

Example: One-Sample t -test

p -Values, Rejection Regions, and CIs

Substantive versus Statistical Significance

Testing with Two Random Variables

Two Sample t -Test for Equality of Means

Paired t -Test

Summary: Hypothesis Testing

Substantive versus Statistical Significance

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} \approx t_{n-1}$$

You reject the null hypothesis, why is that?

1. Maybe $\bar{X} - \mu_0$ is large (big difference between sample mean and mean assumed by H_0)
2. Maybe n is large (you have a lot of data so you have a lot of precision)
3. Maybe the sample standard deviation, s , is small

We need to be careful to distinguish:

- **substantive significance** (e.g. a big effect)
- **statistical significance** (i.e. we reject the null)

In large samples even tiny effects will be statistically significant, but the results may not be very important substantively. Always discuss both!

Roadmap

Estimators and their properties

Interval Estimation

Statistical Hypothesis Testing

Example: One-Sample t -test

p -Values, Rejection Regions, and CIs

Substantive versus Statistical Significance

Testing with Two Random Variables

Two Sample t -Test for Equality of Means

Paired t -Test

Summary: Hypothesis Testing

Roadmap

Estimators and their properties

Interval Estimation

Statistical Hypothesis Testing

Example: One-Sample t -test

p -Values, Rejection Regions, and CIs

Substantive versus Statistical Significance

Testing with Two Random Variables

Two Sample t -Test for Equality of Means

Paired t -Test

Summary: Hypothesis Testing

Testing Equality of Means

We can use the same framework to conduct other tests. For example, we can test equality of means for the distributions of two variables:

$$Y_A \sim ?(\mu_{Y_A}, \sigma_{Y_A}^2) \quad \text{and} \quad Y_B \sim ?(\mu_{Y_B}, \sigma_{Y_B}^2)$$

The variables may be outcomes from two different experimental treatments or two different populations.

Suppose that we have two iid samples from the distributions of Y_A and Y_B respectively with sizes n_A and n_B .

$$H_0 : \mu_{Y_A} = \mu_{Y_B}, \quad H_1 : \mu_{Y_A} \neq \mu_{Y_B}$$

In terms of the difference, the general hypotheses are

$$H_0 : \mu_{Y_A} - \mu_{Y_B} = \delta_0, \quad H_1 : \mu_{Y_A} - \mu_{Y_B} \neq \delta_0$$

To test whether the means are equal, set $\delta_0 = 0$.

Two-Sample t -Test (unequal variances)

If the two samples are independent and n_A and n_B are large, under the null hypothesis,

$$\bar{Y}_A - \bar{Y}_B \approx \mathcal{N}\left(\delta_0, \frac{\sigma_{Y_A}^2}{n_A} + \frac{\sigma_{Y_B}^2}{n_B}\right),$$

or

$$t = \frac{\bar{Y}_A - \bar{Y}_B - \delta_0}{\sqrt{\frac{s_{Y_A}^2}{n_A} + \frac{s_{Y_B}^2}{n_B}}} \approx \mathcal{N}(0, 1)$$

approximately when n_A and n_B are large. We reject H_0 at significance level $\alpha = 0.05$ if $|t_{\text{obs}}| > 1.96$.

Can simplify further if we assume that the two distributions have the same variance.

Example: Testing Equality of Means

Was Kamala Harris as popular as Donald Trump?

```
mean(trump_ft)
```

```
[1] 39.89815
```

```
mean(harris_ft)
```

```
[1] 47.21404
```

Two Sample Test:

```
t.test(x = harris_ft, y = trump_ft)
```

Welch Two Sample t-test

data: harris_ft and trump_ft

t = 10.008, df = 10985, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

5.883011 8.748770

sample estimates:

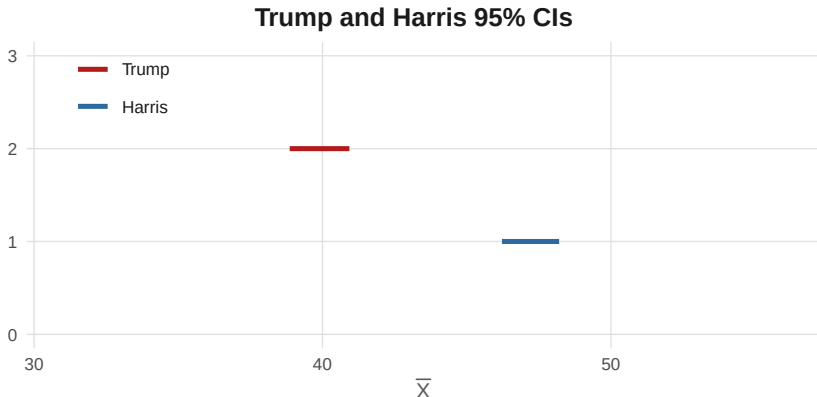
mean of x mean of y

47.21404 39.89815

For illustration purposes only. In reality, the Harris and Trump ratings come from the same respondents; the two samples are not independent.

Example: Testing Equality of Means

Was Kamala Harris as popular as Donald Trump?



For illustration purposes only. In reality, the Harris and Trump ratings come from the same respondents; the two samples are not independent.

Roadmap

Estimators and their properties

Interval Estimation

Statistical Hypothesis Testing

Example: One-Sample t -test

p -Values, Rejection Regions, and CIs

Substantive versus Statistical Significance

Testing with Two Random Variables

Two Sample t -Test for Equality of Means

Paired t -Test

Summary: Hypothesis Testing

Paired t -Test

The two-sample t -test assumes that the two samples of Y_A and Y_B are independent. Often, however, there is only one sample of subjects that have been tested twice (repeated measures or paired observations).

We can increase the power/precision of our test (i.e. obtain a higher probability of rejecting a false null hypothesis) by using a **paired t -test** that considers the distribution of within-subject differences in outcomes

$$D_i = Y_{A,i} - Y_{B,i}$$

Assume $D_1, \dots, D_n \stackrel{\text{i.i.d.}}{\sim} (\mu_D, \sigma_D^2)$; then we may test $H_0 : \mu_D = 0$, the average within-subject outcome difference is zero.

When n is large, under the null we have:

$$t = \frac{\bar{D} - \mu_0}{\sqrt{s_D^2/n}} \approx \mathcal{N}(0, 1)$$

Note: s_D^2 is the sample variance of the within-subject differences.

Paired t -Test

Did ratings of Harris change between the pre- and post-election interviews?

```
t.test(x = harris_pre, y = harris_post, paired = TRUE)
```

Paired t-test

```
data: harris_pre and harris_post  
t = 2.1215, df = 4868, p-value = 0.03393  
alternative hypothesis: true mean difference is not equal to 0  
95 percent confidence interval:  
 0.04065666 1.03061054  
sample estimates:  
mean difference  
 0.5356336
```

Paired t -Test

Did ratings of Harris change between the pre- and post-election interviews?

```
t.test(x = harris_pre,  
      y = harris_post,  
      paired = TRUE)
```

```
t = 2.1215  
df = 4868  
p-value = 0.03393  
95 percent confidence interval:  
 0.04065666 1.03061054  
mean difference  
 0.5356336
```

```
t.test(x = harris_pre,  
      y = harris_post)
```

```
t = 0.71123  
df = 9730.1  
p-value = 0.477  
95 percent confidence interval:  
 -0.9406263 2.0118935  
mean of x mean of y  
 47.88704 47.35141
```

Roadmap

Estimators and their properties

Interval Estimation

Statistical Hypothesis Testing

Example: One-Sample t -test

p -Values, Rejection Regions, and CIs

Substantive versus Statistical Significance

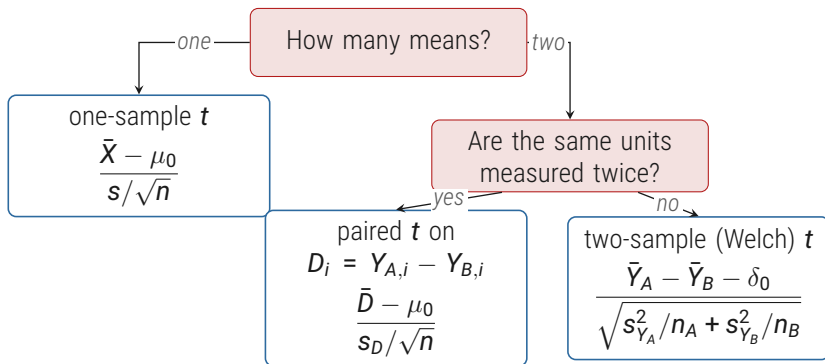
Testing with Two Random Variables

Two Sample t -Test for Equality of Means

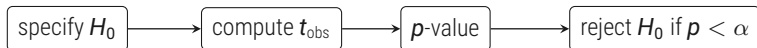
Paired t -Test

Summary: Hypothesis Testing

Summary



For a two-sided test:



or

