Refresher Course in Calculus, Probability, and Statistics

Day 4: Inferential Statistics

Instructor: Dr. Guillaume Wald

10 September 2026

Introduction

- **Inferential statistics** is the science of using data to learn about the world.

→ Statistical tools help answer questions about unknown characteristics of distributions in populations of interest:

  - What is the mean earning among recent economics graduates?
  - Do earnings differ for men and women? How much?

- **Statistical inference:** learn about a population distribution from a random sample.

→ Three types of statistical methods:

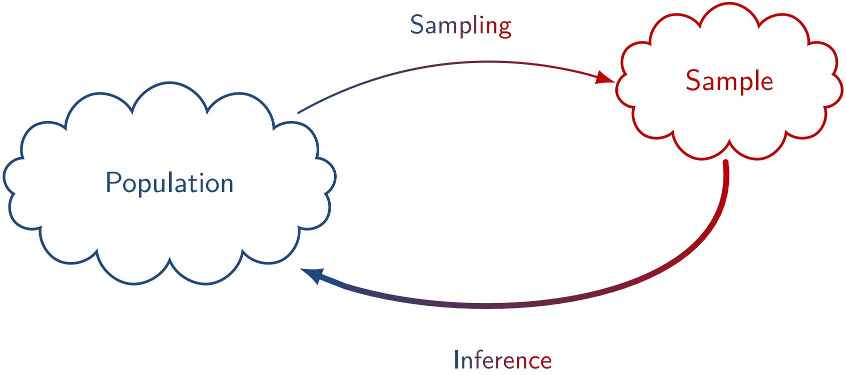
  - Estimation
  - Confidence intervals
  - Hypothesis testing

→ Measures of association:

  - OLS: Ordinary Least Squares

Statistical Inference

- **Definition:** use of *inductive methods* to reason about the distribution of a population on the basis of knowledge obtained in a sample from that population.



Random Sampling

- Samples of a population: **random sampling**.
- **Simple random sampling:**
  - each member of the population is equally likely to be included in the sample
  - the value of the random variable for the ith randomly drawn object is denoted Y_i
  - $Y_1, Y_2, \ldots, Y_n$: a sample of size n from the population is a *data set*
- If the sampling is random, then:
  - two observations are drawn randomly, so the value of Y_i contains no information about the value of Y_j: $Y_1, \ldots, Y_n$ are **independently** distributed
  - Y_i and Y_j are drawn from the same distribution: $Y_1, \ldots, Y_n$ are **identically** distributed
  - Y_i for $i = 1, \ldots, n$ are **independently and identically distributed** (i.i.d.).

Random Sampling

Population

Variable	Obs	Mean
age	2,246	39.15316
wage	2,246	7.766949

Sample 1

Var.	Obs	Mean
age	10	39.4
wage	10	6.946883

Sample 2

Var.	Obs	Mean
age	10	39.5
wage	10	8.278549

- Many different samples

$\rightarrow$ many different values of the mean.

$\Rightarrow$ A whole **probability distribution** of the sample mean emerges.

Distribution of the Sample Average

- Sample average: $\overline{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$.
- For a random sample, $\overline{Y}$ is a **random variable** and has a probability distribution, called the **sampling distribution**.

Distribution of the Sample Average

- Sample average: $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
- For a random sample, $\overline{Y}$ is a **random variable** and has a probability distribution, called the **sampling distribution**.
- Suppose $Y_1, \ldots, Y_n$ are i.i.d. and the population has mean μ_Y and variance σ_Y^2:
  - Expected value:

$$E(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \mu_Y.$$

Distribution of the Sample Average

- Sample average: $\overline{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$.
- For a random sample, $\overline{Y}$ is a **random variable** and has a probability distribution, called the **sampling distribution**.
- Suppose $Y_1, \ldots, Y_n$ are i.i.d. and the population has mean μ_Y and variance σ_Y^2:
  - Expected value:
  $$E(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \mu_Y.$$
  - Variance:
  $$\operatorname{var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}(Y_i) + \underbrace{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\operatorname{cov}(Y_i, Y_j)}_{=\,0\ \ (\text{iid}\ \Rightarrow\ \text{indep.})} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2 = \frac{1}{n^2}\, n\,\sigma_Y^2 = \frac{\sigma_Y^2}{n}.$$

$\Rightarrow$ If $Y_i \sim N(\mu_Y, \sigma_Y^2) \ \Rightarrow\ \overline{Y} \sim N\left(\mu_Y, \dfrac{\sigma_Y^2}{n}\right)$.

Large-Sample Approximations to Sampling Distributions

- **Asymptotic distribution:** large-sample approximation to the sampling distribution; the approximation becomes exact as $n \to \infty$.

Large-Sample Approximations to Sampling Distributions

- **Asymptotic distribution:** large-sample approximation to the sampling distribution; the approximation becomes exact as $n \to \infty$.
- **Law of large numbers:**
    - as n becomes large, $\overline{Y}$ is near μ_Y with very high probability
    - convergence in probability (consistency): $\overline{Y} \xrightarrow{p} \mu_Y$.

Large-Sample Approximations to Sampling Distributions

- **Asymptotic distribution:** large-sample approximation to the sampling distribution; the approximation becomes exact as $n \to \infty$.
- **Law of large numbers:**
  - as n becomes large, $\overline{Y}$ is near μ_Y with very high probability
  - convergence in probability (consistency): $\overline{Y} \xrightarrow{p} \mu_Y$.
- **Central Limit Theorem (CLT):**
  - under general conditions, when n is large the distribution of $\overline{Y}$ is well approximated by a normal distribution
  - this holds *independently* of the distribution of Y
  - if Y_i $(i = 1, \ldots, n)$ are i.i.d. with $E(Y_i) = \mu_Y$ and $\text{var}(Y_i) = \sigma_Y^2$, then as $n \to \infty$

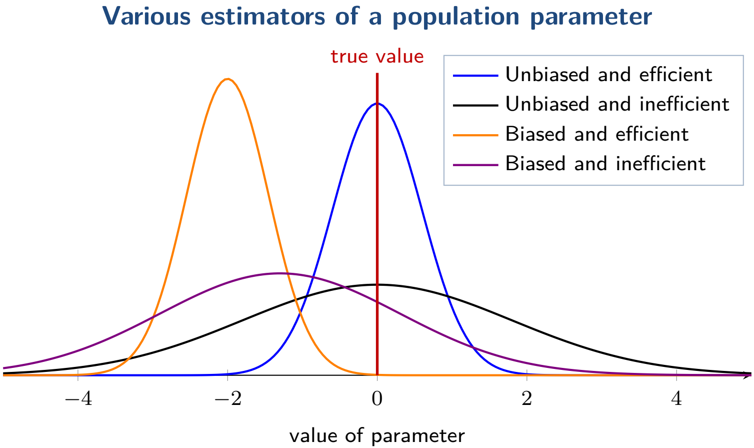
$$\overline{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right) \qquad \text{or} \qquad \frac{\overline{Y} - \mu_Y}{\sigma_Y / \sqrt{n}} \sim N(0, 1).$$

Estimators and their Properties

- **Estimator:** a function of the sample data used to infer the value of an unknown population parameter.
  - true (unknown) parameter: μ
  - estimator: $\hat{\mu}$ (read "mu hat").
- **Estimate:** the numerical value of the estimator actually computed from a specific sample.
- Desirable characteristics of an estimator:
  - **Unbiasedness:** $E(\hat{\mu}) = \mu$ (bias $= E(\hat{\mu}) - \mu$)
  - **Consistency:** $\hat{\mu} \xrightarrow{p} \mu$ $(\mathrm{var}(\hat{\mu}) \to 0$ as $n \to \infty)$
  - **Efficiency:** between two unbiased estimators $\hat{\mu}$ and $\tilde{\mu}$, $\hat{\mu}$ is more efficient if $\mathrm{var}(\hat{\mu}) < \mathrm{var}(\tilde{\mu})$.

Estimators and their Properties

(continued)



Estimation of the Population Mean

- Suppose you want to know the mean value of Y in a population

$\rightarrow$ You could estimate μ_Y exactly only with *all* values of Y

- You have at hand a sample of n i.i.d. observations $Y_1, \ldots, Y_n$ with mean μ_Y and variance σ_Y^2.
- Possible estimators:
    - Y_1
    - $\overline{Y} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} Y_i$;
    - $\widetilde{Y} = \dfrac{1}{n} \left(\frac{1}{2} Y_1 + \frac{3}{2} Y_2 + \cdots + \frac{1}{2} Y_{n-1} + \frac{3}{2} Y_n \right)$ [assume n is even].

How do these estimators fare judged by the three criteria? First, **unbiasedness**:

- $E(Y_1) = \mu_Y$ $\Rightarrow$ unbiased
- $E(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_Y = \mu_Y$ $\Rightarrow$ unbiased
- $$E(\widetilde{Y}) = \frac{1}{n}\left(\tfrac{1}{2}E(Y_1) + \tfrac{3}{2}E(Y_2) + \cdots + \tfrac{1}{2}E(Y_{n-1}) + \tfrac{3}{2}E(Y_n)\right)$$
  $$= \frac{1}{n}\left(\tfrac{1}{2}\cdot\tfrac{n}{2}\,\mu_Y + \tfrac{3}{2}\cdot\tfrac{n}{2}\,\mu_Y\right) = \mu_Y$$ $\Rightarrow$ unbiased

Estimation of the Population Mean

How do these estimators fare? — Efficiency

Next, **efficiency**. With the Y_i i.i.d. ($\mathrm{var}(Y_i) = \sigma_Y^2$, covariances vanish):

- $\mathrm{var}(Y_1) = \sigma_Y^2$
- $\mathrm{var}(\overline{Y}) = \dfrac{1}{n^2}\displaystyle\sum_{i=1}^{n} \mathrm{var}(Y_i) = \dfrac{\sigma_Y^2}{n}$
- $\mathrm{var}(\widetilde{Y}) = \dfrac{1}{n^2}\displaystyle\sum_{i=1}^{n} c_i^2 \ \mathrm{var}(Y_i), \qquad c_i = \tfrac{1}{2}$ (odd i), $\tfrac{3}{2}$ (even i)

$$= \frac{\sigma_Y^2}{n^2}\left(\underbrace{\tfrac{n}{2}\left(\tfrac{1}{2}\right)^2}_{\text{odd}} + \underbrace{\tfrac{n}{2}\left(\tfrac{3}{2}\right)^2}_{\text{even}}\right) = \frac{\sigma_Y^2}{n^2}\left(\tfrac{n}{2}\cdot\tfrac{1}{4} + \tfrac{n}{2}\cdot\tfrac{9}{4}\right)$$

$$= \frac{\sigma_Y^2}{n^2}\cdot\frac{5n}{4} = \frac{5}{4}\,\frac{\sigma_Y^2}{n}$$

Estimation of the Population Mean

How do these estimators fare? — Efficiency

Next, **efficiency**. With the Y_i i.i.d. ($\mathrm{var}(Y_i) = \sigma_Y^2$, covariances vanish):

- $\mathrm{var}(Y_1) = \sigma_Y^2$
- $\mathrm{var}(\overline{Y}) = \dfrac{1}{n^2} \displaystyle\sum_{i=1}^{n} \mathrm{var}(Y_i) = \dfrac{\sigma_Y^2}{n}$
- $\mathrm{var}(\widetilde{Y}) = \dfrac{5}{4} \dfrac{\sigma_Y^2}{n}$

Ranking: $\quad \mathrm{var}(\overline{Y}) = \dfrac{\sigma_Y^2}{n} \; < \; \mathrm{var}(\widetilde{Y}) = \dfrac{5}{4} \dfrac{\sigma_Y^2}{n} \; < \; \mathrm{var}(Y_1) = \sigma_Y^2$

$\Rightarrow$ $\overline{Y}$ is the **most efficient** of all unbiased estimators of μ_Y, called the **Best Linear Unbiased Estimator** (BLUE):

$\rightarrow$ for any $\hat{\mu} = \sum_{i=1}^{n} a_i Y_i$ with non-random constants a_i satisfying $\sum_i a_i = 1$, one has

$\mathrm{var}(\overline{Y}) \leq \mathrm{var}(\hat{\mu})$.

Estimation of the Population Variance

- Population variance: $\text{var}(Y) = \sigma_Y^2 = E\left[(Y - \mu_Y)^2\right]$.
  - If we knew μ_Y, we could estimate σ_Y^2 by:

$$\widetilde{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)^2;$$

  - But μ_Y is unknown: it must be replaced by $\overline{Y}$.

- **Unbiased** estimator of σ_Y^2:

$$\widehat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2;$$

→ Division by $n-1$ instead of n: *degrees-of-freedom correction* as the mean uses up one degree of freedom.

- Estimator of the population **standard deviation** σ_Y: $\widehat{\sigma}_Y = \sqrt{\widehat{\sigma}_Y^2}$.

How Exact are Estimates of the Population Mean?

Population

$\mu_{\text{age}} = 39.2, \qquad \mu_{\text{wage}} = 7.77$

Sample 1

Variable	Obs	Mean
age	10	39.4
wage	10	6.946883

Sample 2

Variable	Obs	Mean
age	10	39.5
wage	10	8.278549

- Because of random sampling error, it is impossible to learn the population mean μ *exactly* from a sample.
- Different samples give different estimates that scatter around μ.

Standard Error (of the Estimated Mean)

- From the CLT: as $n \to \infty$, $\overline{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$.
- Estimated variance of the estimator $\hat{\mu} = \overline{Y}$:

$$\widehat{\mathrm{var}}(\overline{Y}) = \widehat{\sigma}^2_{\overline{Y}} = \frac{\widehat{\sigma}^2_Y}{n} = \frac{1}{n} \cdot \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Standard Error (of the Estimated Mean)

- From the CLT: as $n \to \infty$, $\overline{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$.
- Estimated variance of the estimator $\hat{\mu} = \overline{Y}$:

$$\widehat{\text{var}}(\overline{Y}) = \widehat{\sigma}^2_{\overline{Y}} = \frac{\widehat{\sigma}^2_Y}{n} = \frac{1}{n} \cdot \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

- **Standard error** of $\overline{Y}$:

$$\text{se}(\overline{Y}) = \widehat{\sigma}_{\overline{Y}} = \frac{\widehat{\sigma}_Y}{\sqrt{n}}.$$

  - estimator of the **standard deviation of the sampling distribution** of $\overline{Y}$;

  $\rightarrow$ referred to as the **standard error** of the mean / the estimate.

	Mean	Std. Err.
age	39.4	.8969083
wage	6.946883	.9061781

Confidence Intervals for the Population Mean

- **Confidence interval:** a *random* interval, computed from the sample, that covers the fixed unknown μ with a prescribed probability — the **confidence level** $1 - \alpha$.

→ a 95% interval is constructed so that, over repeated random samples, it contains the true μ in 95% of them.

Confidence Intervals for the Population Mean

- **Confidence interval:** a *random* interval, computed from the sample, that covers the fixed unknown μ with a prescribed probability — the **confidence level** $1-\alpha$.

$\rightarrow$ a 95% interval is constructed so that, over repeated random samples, it contains the true μ in 95% of them.

- Using Student's t-distribution with $n-1$ degrees of freedom:

$$P\big(\overline{Y} - t_{1-\frac{\alpha}{2},n-1} \cdot \mathrm{se}(\overline{Y}) \ \leq \ \mu \ \leq \ \overline{Y} + t_{1-\frac{\alpha}{2},n-1} \cdot \mathrm{se}(\overline{Y})\big) = 1-\alpha,$$

where $t_{1-\frac{\alpha}{2},n-1}$ is the $(1-\frac{\alpha}{2})$-quantile of t_{n-1}

$\Rightarrow$ the Student t critical value places mass $\frac{\alpha}{2}$ in each tail.

	Mean	Std. Err.	[95% Conf.	Interval]
age	39.4	.8969083	37.37105	41.42895
wage	6.946883	.9061781	4.896966	8.996800

Confidence Intervals for the Population Mean

(continued)

- Assuming a large sample size, the distribution is approximately normal, so we may replace the t-multiplier by the standard-normal critical value:

$$P\Big(\overline{Y} - z_{\text{critical}} \cdot \text{se}(\overline{Y}) \ \leq \ \mu \ \leq \ \overline{Y} + z_{\text{critical}} \cdot \text{se}(\overline{Y})\Big) = 1 - \alpha.$$

- Common two-sided critical values:
  - 90% CI: $z_{\text{critical}} = 1.645$;
  - 95% CI: $z_{\text{critical}} = 1.96$;
  - 99% CI: $z_{\text{critical}} = 2.58$.

Hypothesis Testing

- A hypothesis is a proposition about a population — a yes/no question.
- Hypotheses we might want to test:
  - Do university graduates earn CHF 6,000 per month on average?
  - $\rightarrow$ hypothesis about the population mean of a *single* population;
  - Are mean earnings the same for men and women?
  - $\rightarrow$ hypothesis about differences in means between *two* populations.

- Components of a hypothesis test:
  - Null hypothesis, H_0
  - Alternative hypothesis, H_1
  - Test statistic
  - Rejection region
  - Conclusion

Null and Alternative Hypotheses

- The **null hypothesis** H_0 specifies a value c for a parameter:

$$H_0 : \ \mu = c.$$

  - H_0 is what we believe until the sample provides evidence against it; then we reject H_0.

- **Alternative hypotheses:**

$$\begin{aligned} H_1 &: \ \mu \neq c \qquad \text{(two-sided alternative)} \\ H_1 &: \ \mu > c \qquad \text{(one-sided alternative)} \\ H_1 &: \ \mu < c \qquad \text{(one-sided alternative)} \end{aligned}$$

Test Statistic

- The sample information about H_0 is embodied in a **test statistic**.
- Based on the value of the test statistic, we decide whether it is reasonable to reject H_0 or not.
- Consider $H_0 : \mu = c$. If the sample comes from a population $N(\mu, \sigma)$:

$$t = \frac{\overline{Y} - c}{\widehat{\sigma}/\sqrt{n}} \sim t_{n-1}.$$

- When do we use the t distribution and when the normal distribution?
  - in practice, the t *distribution* is used for *small* samples
  - when the **sample size is large**, the *normal distribution* is an accurate enough approximation.

Rejection Region

- **Rejection region:** the range of values of the t-statistic that leads to rejecting H_0.

$\rightarrow$ If the t-statistic falls in a region of low probability, H_0 is probably not true.

- If H_0 is false (H_1 true), the test statistic is unusually "large" or unusually "small" given the sampling distribution.

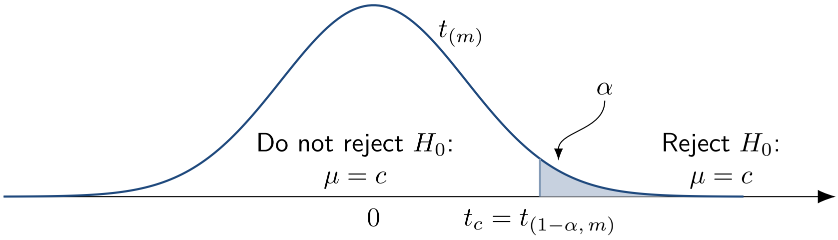
  Depending on the chosen probability α, we may reject H_0 incorrectly (or fail to reject it):
  - α is the **level of significance**
  - common values: $\alpha = 0.1,\ 0.05,\ 0.01.$

One-Tailed Test $(>)$

- $H_0 : \mu \leq c$ and $H_1 : \mu > c$.
- Critical value $t_{\text{critical}} = t_{(1-\alpha,\, n-1)}$ is the $100(1-\alpha)$-percentile of a t-distribution with $n-1$ degrees of freedom:

$$P(t \leq t_{\text{critical}}) = 1 - \alpha$$

- If the test statistic $\geq t_{\text{critical}}$ $\Rightarrow$ reject H_0.



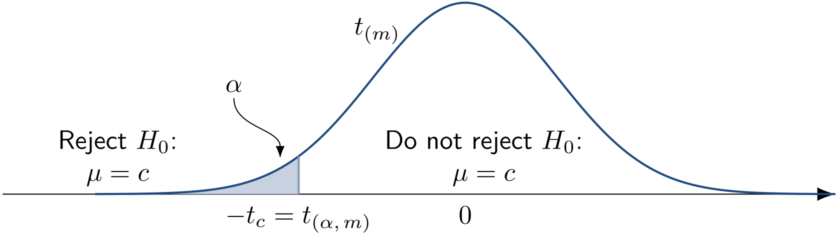
Hill, Griffiths, & Lim (2008)

One-Tailed Test $(<)$

- $H_0 : \mu \geq c$ and $H_1 : \mu < c$.
- Critical value $-t_{\text{critical}} = t_{(\alpha, n-1)}$ is the 100α-percentile of a t-distribution with $n-1$ degrees of freedom:

$$P(t \leq -t_{\text{critical}}) = \alpha$$

- If the test statistic $\leq -t_{\text{critical}} \Rightarrow$ reject H_0.



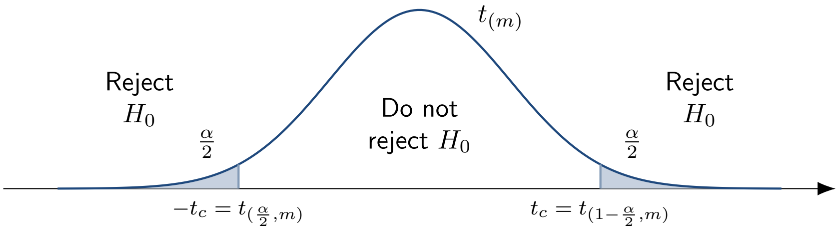
Hill, Griffiths, & Lim (2008)

Two-Tailed Test ($\neq$)

- $H_0 : \mu = c$ and $H_1 : \mu \neq c$.
- Critical value $t_{\text{critical}} = t_{(1-\frac{\alpha}{2}, n-1)}$ is the $100(1-\frac{\alpha}{2})$-percentile of a t-distribution with $n-1$ degrees of freedom:

$$P(t \geq t_{\text{critical}}) = P(t \leq -t_{\text{critical}}) = \frac{\alpha}{2}$$

- If the test statistic $\geq |t_{\text{critical}}|$ or $\leq -|t_{\text{critical}}|$ $\Rightarrow$ reject H_0.

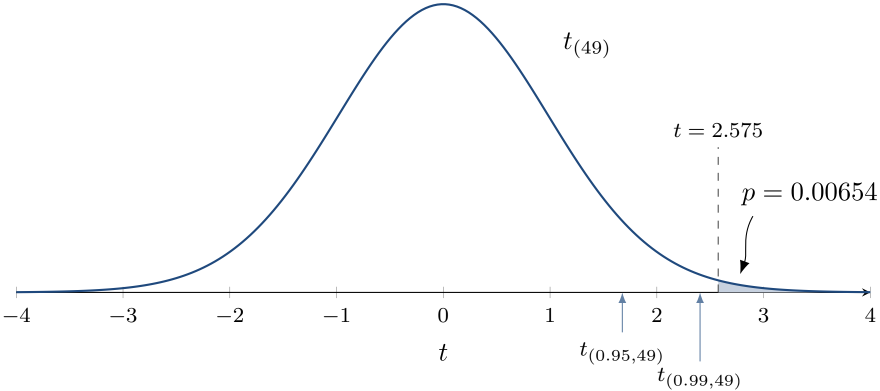


Hill, Griffiths, & Lim (2008)

The p-Value

- The p-value lets us determine the outcome of a test by comparing it to the level of significance α, without computing the critical value.
- **p-value rule:**
  - $p < \alpha \ \Rightarrow$ reject H_0;
  - $p \geq \alpha \ \Rightarrow$ do not reject H_0.
- Computation of the p-value (with t the observed statistic):
  - if $H_1 : \mu < c \ \Rightarrow \ p =$ probability to the *left* of t;
  - if $H_1 : \mu > c \ \Rightarrow \ p =$ probability to the *right* of t;
  - if $H_1 : \mu \neq c \ \Rightarrow \ p =$ sum of the probabilities beyond $\pm|t|$ (both tails).
- When n is large (using the standard normal Φ):
  - if $H_1 : \mu < c \ \Rightarrow \ p = \Phi(t)$;
  - if $H_1 : \mu > c \ \Rightarrow \ p = 1 - \Phi(t)$;
  - if $H_1 : \mu \neq c \ \Rightarrow \ p = 2\big[1 - \Phi(|t|)\big]$.

p-Value for a One-Tailed Test $(>)$



- $t = 2.575 > t_{(0.99,49)} = 2.405$ (equivalently $p = 0.00654 < 0.01 = \alpha$):

$\Rightarrow$ **Reject** H_0 at the 1% level.

Hill, Griffiths, & Lim (2008)

Conclusion of the Hypothesis Test

- When the test is completed, the conclusion is one of:
  - reject H_0;
  - fail to reject H_0.
- H_0 is never said to be *accepted*: absence of evidence is not evidence of absence!
- Do not forget to interpret the test results in a meaningful way for the economic / financial problem you are working on.
- Possible mistakes:
  - **Type I error** (α)
    - False positive (of an effect / a difference)
    - H_0 rejected when it is true
  - **Type II error** (β)
    - False negative
    - H_0 not rejected when it is false

- Is the mean income of US women significantly different from 10?

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
wage	2,246	7.766949	.1214451	5.755523	7.528793	8.005105

mean = mean(wage) $\quad t = -18.39$

H_0: mean $= 10$ $\quad$ degrees of freedom $= 2{,}245$

H_a: mean < 10	H_a: mean $\neq 10$	H_a: mean > 10
$\Pr(T < t) = 0.0000$	$\Pr(\lvert T\rvert > \lvert t\rvert) = 0.0000$	$\Pr(T > t) = 1.0000$

Testing the Equality of Two Population Means

- Let two populations be distributed as $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$.
- To test the difference $\mu_1 - \mu_2$, take a random sample from each:
  - sample of size n_i from population i;
  - sample mean $\overline{Y}_i$ and sample variance $\widehat{\sigma}_i^2$.
- Null hypothesis: $H_0: \ \mu_1 - \mu_2 = c$.

Testing the Equality of Two Population Means

- Let two populations be distributed as $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$.
- To test the difference $\mu_1 - \mu_2$, take a random sample from each:
  - sample of size n_i from population i;
  - sample mean $\overline{Y}_i$ and sample variance $\widehat{\sigma}_i^2$.
- Null hypothesis: $H_0: \; \mu_1 - \mu_2 = c$.
- **Case (a): population variances are equal.** Use both samples to estimate $\sigma_p^2 = \sigma_1^2 = \sigma_2^2$:

$$\widehat{\sigma}_p^2 = \frac{(n_1 - 1)\widehat{\sigma}_1^2 + (n_2 - 1)\widehat{\sigma}_2^2}{n_1 + n_2 - 2}.$$

  If H_0 is true:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2 - c}{\sqrt{\widehat{\sigma}_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \;\sim\; t_{n_1+n_2-2}.$$

- **Case (b): population variances are unequal.** If H_0 is true:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2 - c}{\sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}}}.$$

  - the exact distribution of this statistic is neither normal nor a standard t;
  - it is approximated by a t-distribution with degrees of freedom

$$df = \frac{\left(\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}\right)^2}{\frac{\left(\widehat{\sigma}_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(\widehat{\sigma}_2^2/n_2\right)^2}{n_2 - 1}};$$

  - when both n_1 and n_2 are large, the t-statistic has a standard normal distribution.

Testing the Equality of Two Population Means

Example — Hypotheses and test statistic

Question: is mean log wage the same for the two groups (`white`$=0$ vs. `white`$=1$)?
Summary statistics:

Variable	Obs	Mean	Std. Dev.	Min	Max
$\rightarrow$ white $=0$					
ln_wage	609	1.755669	.5722596	.140951	3.707372
$\rightarrow$ white $=1$					
ln_wage	1,637	1.910674	.5699247	.0049396	3.693819

- **Step 1 — hypotheses** (difference $c=0$; two-sided):
  $H_0: \ \mu_0 - \mu_1 = 0$ vs. $H_1: \ \mu_0 - \mu_1 \neq 0$.
- **Step 2 — pooled SD & standard error** (variances assumed equal):
  $\widehat{\sigma}_p^2 = \frac{(n_0-1)\widehat{\sigma_0^2}+(n_1-1)\widehat{\sigma_1^2}}{n_0+n_1-2} \approx 0.3256$, so $\widehat{\sigma}_p \approx 0.5706$ and $\mathrm{se}(\text{diff}) = \widehat{\sigma}_p\sqrt{\frac{1}{n_0}+\frac{1}{n_1}} = 0.0271$.
- **Step 3 — test statistic:** $t = \frac{\overline{Y}_0-\overline{Y}_1-c}{\mathrm{se}(\text{diff})} = -\frac{0.1550}{0.0271} = -5.7237, \quad df = n_0+n_1-2 = 2{,}244.$

Testing the Equality of Two Population Means

Example — Decision and conclusion

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Int.]
0	609	1.755669	.0231891	.5722596	1.710128	1.801210
1	1,637	1.910674	.0140862	.5699247	1.883045	1.938303
combined	2,246	1.868645	.0121240	.5745800	1.844870	1.892420
diff		−.1550052	.0270814		−.2081123	−.1018977

$t = -5.7237$ $df = 2{,}244$ $\texttt{Pr}(T < t) = 0.0000$ $\texttt{Pr}(|T| > |t|) = 0.0000$ $\texttt{Pr}(T > t) = 1.0000$

- **Step 4 — rejection region / p-value** (two-sided, $\alpha = 0.05$): $|t| = 5.72 \gg t_{0.975,\,2244} \approx 1.96$, so we reject; equivalently $p = \Pr(|T| > |t|) = 0.0000 < \alpha$ (also below 0.01). The 95% CI for the difference, $[-0.2081, -0.1019]$, does not contain 0.
- **Step 5 — conclusion:** reject H_0. Mean log wage differs significantly between the two groups; the estimated gap is $\overline{Y}_0 - \overline{Y}_1 \approx -0.155$ log points (roughly 15% in wage terms). NB: this is a descriptive difference in group means, not a ceteris-paribus effect.