# STAT 390 A
# Statistical Methods in Engineering and Science
# Week 09 Lectures - Part 1 & 2 – Spring 2023
# Introduction to Hypothesis Testing

Alexander Giessing

Department of Statistics

University of Washington

May 20, 2023

# Outline

# Hypothesis Testing

- Recall the ways of learning about an unknown parameter from data:
  - point estimates of parameters (MLE, MME).
  - interval estimates of parameters (CIs for the mean and other quantities).
- Often numerical values are not of primary interest, but one would like to use the data to answer *'yes or no'* questions:
  - Is the Omicron variant more contagious than the Delta variant?
  - Does Pfizer's Covid-19 Antiviral Candidate PAXLOVID reduce hospitalization and death?
  - Are mRNA vaccines more effective than conventional vaccines?
  - Does a new marketing policy increase market share?
  - Is the average nicotine content $< 1.5mg$ per pack of cigarettes?
  - Are data consistent with a theory/ a statistical model (i.e. does the introduction of a minimum wage increase inflation or are waiting times geometrically distributed)?
- **Statistical Hypothesis Testing** can be used to address such questions.

# Example: Tesla Factory in Fremont, CA

*Suppose that under the standard factory setup it is known that the time to assemble a Model 3 has distribution* $N(90, 1)$. *Elon Musk changes the production setup to reduce the amount of time it takes to assemble a Model 3. He samples* 35 *completion times under the new setup and finds that the average time is* 87. *Does the new production setup indeed reduce completion time of a Model 3?*

- Let $X_1, \ldots X_{35}$ be a random sample of completion times with mean $\mu$ and variance 1. Denote by $\bar{X}_{35}$ and $\bar{x}_{35}$ the sample mean and its realization.
- Compute how likely it is to observe a sample mean of at most 87 if the true population mean $\mu$ equals 90, i.e.

$$P(\bar{X}_{35} \leq 87 \mid \mu = 90).$$

- If above probability is very small, then it is extremely unlikely to observe $\bar{x}_{35} = 87$ if the true population mean $\mu$ equals 90.
- Reversing this reasoning: If above probability is small and yet we observe $\bar{x}_{35} = 87$, then most likely the true population mean is not equal to 90; in particular, $\mu < 90$.

# Example: Tesla Factory in Fremont, CA (Cont.)

- Compute

$$
\begin{aligned}
P\left(\bar{X}_{35} \leq 87 \mid \mu = 90\right) &= P\left(\frac{\bar{X}_{35} - \mu}{1/\sqrt{35}} \leq \frac{87 - \mu}{1/\sqrt{35}} \mid \mu = 90\right) \\
&= P\left(\frac{\bar{X}_{35} - 90}{1/\sqrt{35}} \leq \frac{87 - 90}{1/\sqrt{35}} \mid \mu = 90\right) \\
&\overset{(a)}{=} \Phi(-\sqrt{35} \times 3) = 8.62 \times 10^{-71} \approx 0,
\end{aligned}
$$

  where (a) holds since under the hypothesis $\mu = 90$ the distribution of the standardized sample mean is $\sqrt{35}(\bar{X}_{35} - 90) \sim N(0, 1)$.

  $\Longrightarrow$ If $\mu = 90$, it is very unlikely to observe $\bar{x} = 87$.

- Since $P\left(\bar{X}_{35} \leq 87 \mid \mu = \mu_0\right) = \Phi\big(-\sqrt{35} \times (\mu_0 - 87)\big)$, the probability of observing $\bar{x}_n = 87$ is even smaller for any $\mu_0 > 90$.

  $\Longrightarrow$ If $\mu \geq 90$, it is very unlikely to observe $\bar{x} = 87$. Hence, there is strong evidence in the data that $\mu < 90$.

# Hypothesis Testing: Concepts and Definitions

## Statistical Hypotheses

Statistical hypotheses are statements about (features of) the probability distribution of a population whose plausibility one would like to assess after collecting data.

- Null hypotheses describe 'prior beliefs' which one hopes to refute using the collected data.
- Alternative hypotheses are the statements that one hopes to establish using the collected data.

We denote null and alternative hypotheses by $H_0$ and $H_1$, respectively.

- In the Tesla factory example the feature of the distribution, in which we are interested, is the mean $\mu$.
- Null and alternative hypothesis in the Tesla factory example are

$$H_0 : \mu = 90 \qquad vs. \qquad H_1 : \mu < 90.$$

# Hypothesis Testing: Concepts and Definitions

## Null and Alternative Model

The null model is the probability distribution of the population under $H_0$ (i.e assuming that the null hypothesis is true). The alternative model is the probability distribution of the population under $H_1$.

- Reconsider the Tesla factory example:
  - Null model: completion times follow $N(90, 1)$.
  - Alternative model: completion times follow $N(\mu, 1)$ for some $\mu < 90$.

## Test Statistic

Suppose that the data set is modeled as the realization of a random sample $X_1, \ldots, X_n$. A test statistic is any sample statistic $T = h(X_1, \ldots, X_n)$, whose numerical value is used to decide whether to reject $H_0$ or not.

- In the Tesla factory example the test statistic is the standardized sample mean, i.e.

$$T = \frac{\bar{X}_{35} - \mathrm{E}[\bar{X}_{35}]}{\sqrt{\mathrm{Var}(\bar{X}_{35})}} \overset{(a)}{=} \frac{\bar{X}_{35} - 90}{1/\sqrt{35}},$$

  where (a) holds only (!) under $H_0$ (i.e. when $\mathrm{E}[\bar{X}_{35}] = \mu = 90$).

# Hypothesis Testing: Concepts and Definitions

*P*-Values (Heuristic Definition)

The $p$-value is the probability of collecting evidence against $H_0$ that is at least as strong as the one observed.

- If the $p$-value is sufficiently small we reject $H_0$ in favor of $H_1$.
- The smaller the $p$-value, the stronger the evidence against $H_0$.
- Reconsider the Tesla factory example:
  - The observed evidence against $H_0$ is $\bar{x}_{35} = 87$.
  - The probability of collecting evidence against $H_0$ that is at least as strong as the observed value of 87 is $P(\bar{X}_{35} \leq 87 \mid \mu = 90)$.
  - The p-value is approx. 0, hence we reject $H_0$ in favor of $H_1$.

# Hypothesis Testing: Concepts and Definitions

*P*-Values (Definition based on Test Statistics)

Let the data set $x_1, \ldots, x_n$ be a realization of the random sample $X_1, \ldots, X_n$. Let $T = h(X_1, \ldots, X_n)$ be a test statistic and $t = h(x_1, \ldots, x_n)$ its realization. The $p$-value is the probability of the test statistic $T$ taking on values at least as extreme as the observed value $t$ based on the data set.

- The phrase "at least as extreme as the observed value" needs to be interpreted in accordance with the hypotheses being tested.
- Reconsider the Tesla factory example:
  - The test statistic and its realization are $T = \sqrt{35}(\bar{X}_{35} - 90)$ and $t = -\sqrt{35} \times 3 \approx -17.75$.
  - We want to infer whether Elon's changes to the production process reduced the completion time of a Model 3 to below 90 min.
  - Therefore, the smaller the value of $T$, the stronger the evidence provided against $H_0$. For this reason, the $p$-value is the left tail probability

$$P(T \leq -17.75 \mid \mu = 90) = 8.62 \times 10^{-71}.$$

# Hypothesis Testing: Concepts and Definitions

- Recall the Tesla factory example:
  - Based on a sample of size 35 we cannot prove or disprove $H_0$. While it is extremely unlikely to observe $\bar{x}_{35} = 87$ under $H_0$, it is not impossible.
  - We can make two possible decisions: Either we reject $H_0$ in favor of $H_1$ or we do not reject $H_0$.
  - This leads to four possible situations, summarized below:

| | | **True state of nature** | |
|---|---|---|---|
| | | $H_0$ is true | $H_1$ is true |
| **Our decision on the basis of the data** | Reject $H_0$ | *Type I error* | Correct decision |
| | Not reject $H_0$ | Correct decision | *Type II error* |

# Hypothesis Testing: Concepts and Definitions

## Type I and II Errors

A type I error occurs if we incorrectly reject $H_0$, i.e. we reject $H_0$ even though $H_0$ is correct. A type II error occurs if we incorrectly fail to reject $H_0$, i.e. we do not reject $H_0$ even though $H_1$ is correct.

- What error could we have made in the Tesla factory example?
  - We rejected $H_0$. Hence, we may have committed a type I error.
- In the medical and life sciences ...
  - ... the type I error is also known as "false discovery" or "false positive";
  - ... the type II error is also known as "missed discovery" or "false negative".

  Explanation: Suppose that you want to establish that gene $A$ causes a hereditary disease $D$. You therefore want to test: $H_0$ : *"$A$ does not cause $D$"* versus $H_1$ : *"$A$ causes $D$"*. If you incorrectly rejected $H_0$, you incorrectly decided to believe that $A$ causes $D$, i.e. you discovered a causal connection between $A$ and $D$ even though there is none; hence "false discovery".

# Outline

# Neyman-Pearson Framework of Hypothesis Testing

- Recall: We reject $H_0$ in favor of $H_1$ if the $p$-value is sufficiently small.
  - What does "sufficiently small" mean?
  - Less than 10%, 1%, 0.01%, ...?
- Whatever our decision is, there is always a chance that we err and commit a type I or type II error.
  - What are the probabilities of type I and type II errors?
- In some applications a wrong decisions may have disastrous consequences:
  - How can we construct tests with a prescribed type I error probability?
- The **Neyman-Pearson framework** allows us to answer these questions.

# Example: Testing for Speeding

*To test whether a driver is speeding on a freeway with a speed limit of 65 mph, a device takes three measurements $x_1, x_2, x_3$ of the speed of a passing vehicle, modeled as the realization of a random sample $X_1, X_2, X_3$. Denote the sample average and its realization by $\bar{X}_3$ and $\bar{x}_3$. For what values $\bar{x}_3$ of $\bar{X}_3$ should we fine the driver, if we allow that 5% of the drivers are fined unjustly?*

- To make this a statistical problem, assume that each measurement $X_i$ can be thought of as

$$X_i = \mu + \epsilon_i, \quad i = 1, 2, 3,$$

  where $\mu$ is the true speed of the vehicle and $\epsilon_i$ is a measurement error.

- Suppose that the device is calibrated so that $\epsilon_i \sim N(0, 4)$. This implies that $X_i \sim N(\mu, 4)$. (Why?)

- We can now formulate our testing problem as

$$H_0 : \mu = 65 \qquad vs. \qquad H_1 : \mu > 65,$$

  with test statistic and observed value

$$T = \frac{X_1 + X_2 + X_3}{3} = \bar{X}_3 \qquad \text{and} \qquad t = \frac{x_1 + x_2 + x_3}{3} = \bar{x}_3.$$

# Example: Testing for Speeding (Cont.)

- Idea: If the observed value $t$ is much larger than 65 we interpret this as strong evidence against the null hypothesis.
  - Thus, we will reject $H_0$ in favor of $H_1$ for large values $t$ of $T$.
  - Unjustly fining a driver corresponds to falsely rejecting $H_0$, i.e. committing a type I error.

- Since we want to unjustly fine only 5% of the drivers, we have to answer the following question:

  *How large does $T$ have to be to incorrectly reject $H_0$ with only 5% probability?*

- Mathematically speaking, we want to find the critical value $c$ such that

$$P(T \geq c \mid \mu = 65) = 0.05$$

  and we will reject $H_0$ whenever $t > c$, where $t$ is a realization of $T$.

# Example: Testing for Speeding (Cont.)

- Solving for the critical value $c$:
  - Obviously (why?), we have

$$0.05 = P(T \geq c \mid \mu = 65) = P\left(\frac{T - 65}{2/\sqrt{3}} \geq \frac{c - 65}{2/\sqrt{3}}\right) = 1 - \Phi\left(\frac{c - 65}{2/\sqrt{3}}\right). \quad (1)$$

  - We also know (using R) that

$$0.05 = 1 - \Phi(z_{0.05}) \qquad \Longleftrightarrow \qquad z_{0.05} = 1.645. \quad (2)$$

  - Combining eq. (1) and (2) we obtain

$$1.645 = \frac{c - 65}{2/\sqrt{3}} \qquad \Longleftrightarrow \qquad c = 65 + 1.645 \times \frac{2}{\sqrt{3}} = 66.9.$$

- Thus, if we reject $H_0$ for values $t$ of $T$ larger than 66.9, we will commit a type I error in 5% of the cases only/ with probability 5% only.
- **Note:** Either we made the correct decision or we didn't. However, above procedure guarantees that the chance of a type I error is 5%. We cannot (yet) say anything about the chance of a type II error.

# Significance Level

**Significance Level**

The significance level is the largest acceptable probability of committing a type I error and is denoted by $\alpha$, $0 < \alpha < 1$.

- The significance level is the level below which the $p$-value is sufficiently small to reject $H_0$.
- We speak of "performing a test at (significance) level $\alpha$" or "'rejecting $H_0$ in favor of $H_1$ at (significance) level $\alpha$" whenever we construct a test that controls the type I error as in above speeding example.
- When performing a test at level $\alpha$, we will incorrectly reject $H_0$ in favor of $H_1$ $100\alpha\%$ of the time.

# Critical Region and Critical Values

## Critical Region and Critical Values

Suppose that we test $H_0$ against $H_1$ at significance level $\alpha$ by means of a test statistic $T$. The set $K \subset \mathbb{R}$ that corresponds to all values of $T$ for which we reject $H_0$ in favor of $H_1$ is called the critical region. Value(s) on the boundary of the critical region are called critical value(s).

- The precise shape of the critical region $K$ depends on the significance level $\alpha$ and the test statistic $T$. Above definition implies that

$$P(T \in K \mid H_0 \text{ is true}) \leq \alpha.$$

- Reconsider the speeding example:
  - We decided to reject $H_0$ in favor of $H_1$ at a 5% significance level. We showed that this is equivalent to rejecting $H_1$ whenever $t \geq 66.9$.
  - Rejection region $K = [66.9, \infty)$.
  - Critical value $c_{0.05} = 66.9$.

# Example: $p$-value, sig. level $\alpha$, and rejection region $K$

- Reconsider the speeding example: We decided to reject $H_0$ in favor of $H_1$ at a 5% significance level. We showed that this is equivalent to rejecting $H_0$ whenever $t \geq 66.9$.

- Suppose that we observe $t = 67$.
  - We reject $H_0$ at level 5% because $67 > 66.9 = c_{0.05}$.
  - $p$-value corresponding to $t = 67$ is $P(T \geq 67 \mid \mu = 65) = 0.042$.

$$P(T \geq 67 \mid \mu = 65) = 1 - \Phi\left(\frac{67 - 65}{2/\sqrt{3}}\right) = 1 - \Phi(1.73) = 0.042 < 5\%.$$

- **In general:** If we test $H_0$ against $H_1$ at level $\alpha$, then

$$t \in K \quad \Longleftrightarrow \quad p\text{-value corresponding to } t \text{ is less than or equal to } \alpha.$$

*Deciding whether to reject $H_0$ at a given significance level $\alpha$ can be done by either comparing the observed test statistic $t$ with the critical value $c_\alpha$ or the $p$-value corresponding to $t$ with $\alpha$.*

# Power and Power Function of a Test

## Power of a Test

The power of a test is the probability of correctly rejection the null hypothesis, i.e.

$$\text{power } = 1 - P(\text{type II error}).$$

## Power Function of a Test

Consider testing a feature $\theta \in \mathbb{R}$ of a probability distribution:

$$H_0 : \theta \in \Theta_0 \qquad vs. \qquad H_1 : \theta \in \mathbb{R} \setminus \Theta_0 =: \Theta_1.$$

If we test $H_0$ against $H_1$ by means of a test statistic $T$ and a critical region $K$, then the power function of the test is defined as

$$f(x) = P(T \in K \mid \theta = x) = \begin{cases} P(\text{type I error}) & \text{if } x \in \Theta_0, \\ 1 - P(\text{type II error}) & \text{if } x \in \Theta_1. \end{cases}$$

# Example: Power of the Speeding Test

*Let's return to the speeding example. We decided to reject $H_0$ in favor of $H_1$ at a 5% significance level. We showed that this is equivalent to rejecting $H_0$ whenever $t \geq 66.9$. What is the probability of committing a type II error?*

- Suppose that the true speed of a car is $\mu = 75$ mph. A type II error occurs when $T < 66.9$. Since $T = \bar{X}_3 \sim N(75, 4/3)$ (Why?),

$$P(T < 66.9 \mid \mu = 75) = P\left(\frac{T - 75}{2/\sqrt{3}} < \frac{66.9 - 75}{2/\sqrt{3}}\right) = \Phi(-7.01) \approx 0.$$

  $\implies$ power at $\mu = 75$ mph $\approx 100\%$, i.e. we catch (almost) everyone driving 75 mph.

- Suppose that the true speed of a car is $\mu = 68$ mph. The probability of a type II error is now

$$P(T < 66.9 \mid \mu = 68) = P\left(\frac{T - 75}{2/\sqrt{3}} < \frac{66.9 - 68}{2/\sqrt{3}}\right) = \Phi(-0.95) \approx 17.11\%.$$

  $\implies$ power at $\mu = 68$ mph $= 100 - 17.11\% = 82.89\%$.

- **Take-away:** We can control $\alpha$, but we cannot control the power of a test.

# Example: Power of the Speeding Test (Cont.)

*(Stylized sketch of null and alternative models in the testing problem.)*

# Outline

# Configuration of Tests for the Population Mean

## Tests for the Population Mean

Given a 'prior belief' $\mu_0$ about the population mean we have the following three testing problems named after the alternative hypothesis that one hopes to establish:

(a) One-sided upper tail test:

$$H_0 : \mu \leq \mu_0 \qquad vs. \qquad H_1 : \mu > \mu_0.$$

(b) One-sided lower tail test:

$$H_0 : \mu \geq \mu_0 \qquad vs. \qquad H_1 : \mu < \mu_0.$$

(c) Two-sided (tail) test:

$$H_0 : \mu = \mu_0 \qquad vs. \qquad H_1 : \mu \neq \mu_0.$$

- Note that the testing problem

$$H_0 : \mu = \mu_0 \qquad vs. \qquad H_1 : \mu > \mu_0$$

is the same as (a) because both try to establish the same $H_1$, etc. ...

# Large Sample Test for the Population Mean

### Large Sample Test for the Mean (Known Variance)

Let $x_1, \ldots, x_n$ be a realization of a random sample with mean $\mu$ and variance $\sigma^2$. Let $\alpha \in (0, 1)$ and $z_\alpha$ be the critical value of the standard normal distribution and define

$$t = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}.$$

(a) One-sided upper tail test at level $\alpha$:

$$\textit{Reject } H_0 \textit{ in favor of } H_1 \textit{ whenever } t > z_\alpha.$$

(b) One-sided lower tail test at level $\alpha$:

$$\textit{Reject } H_0 \textit{ in favor of } H_1 \textit{ whenever } t < -z_\alpha.$$

(c) Two-sided (tail) test at level $\alpha$:

$$\textit{Reject } H_0 \textit{ in favor of } H_1 \textit{ whenever } |t| > z_{\alpha/2}.$$

# Large Sample Test for the Population Mean (Cont.)

*Derivation the 'rejection rule' of the two-sided (tail) test at level $\alpha$. (The other cases are similar. Convince yourself!)*

- We want to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ at a $\alpha$ sig. level when the variance $\sigma^2$ is known and the sample is large.
  - Idea: If the observed sample average $\bar{x}_n$ is either much larger or much smaller than $\mu_0$, we interpret this as (strong) evidence against $H_0$.
  - Since we want to control the type I error at level $\alpha$, we have to answer the following question:

    *How large (or small) does $\bar{X}_n$ have to be to incorrectly reject $H_0$ with $\alpha$ probability?*

- Mathematically speaking, we want to find the critical values $c_l, c_u$ such that

$$P(\bar{X}_n > c_u \text{ or } \bar{X}_n < c_l \mid H_0 \text{ is true}) = \alpha.$$

  and we will reject $H_0$ whenever $\bar{x}_n > c_u$ or $\bar{x}_n < c_l$.

# Large Sample Test for the Population Mean (Cont.)

- We compute

$$P(\bar{X}_n > c_u \text{ or } \bar{X}_n < c_l \mid H_0 \text{ is true}) = \alpha$$

$$\iff \quad P(c_l \leq \bar{X}_n \leq c_u \mid H_0 \text{ is true}) = 1 - \alpha$$

$$\iff \quad P\left(\frac{c_l - \mu_0}{\sigma/\sqrt{n}} \leq \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \leq \frac{c_u - \mu_0}{\sigma/\sqrt{n}} \mid H_0 \text{ is true}\right) = 1 - \alpha. \qquad (3)$$

- **If $H_0$ is true**, the $X_i$'s have mean $\mu_0$ and **the CLT implies** that the test statistic

$$T = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mathrm{E}[\bar{X}_n])}{\sqrt{\mathrm{Var}(\bar{X}_n)}}$$

has asymptotic distribution $N(0,1)$. Hence,

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \mid H_0 \text{ is true}\right) \approx 1 - \alpha, \qquad (4)$$

where $z_{\alpha/2}$ is the $\alpha/2$-critical value of the standard normal distribution.

# Large Sample Test for the Population Mean (Cont.)

- Combining eq. (3) and (4) we obtain

$$\frac{c_l - \mu_0}{\sigma/\sqrt{n}} = -z_{\alpha/2} \qquad \text{and} \qquad \frac{c_u - \mu_0}{\sigma/\sqrt{n}} = z_{\alpha/2}$$

$$\iff \quad c_l = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \text{and} \qquad c_u = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

- Thus, we will reject $H_0$ whenever

$$\bar{x}_n > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \text{or} \qquad \bar{x}_n < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

- This is equivalent to the more succinct formulation:

$$\text{Reject } H_0 \text{ whenever } \left|\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}.$$

# Large Sample Test for the Population Mean (Cont.)

### Large Sample Test for the Mean (Unknown Variance)

Let $x_1, \ldots, x_n$ be a realization of a random sample with mean $\mu$ and unknown variance $\sigma^2$. Let $\alpha \in (0,1)$ and $z_\alpha$ be the critical value of the standard normal distribution and define

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}, \qquad \text{where} \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

(a) One-sided upper tail test at level $\alpha$:

$$\textit{Reject } H_0 \textit{ in favor of } H_1 \textit{ whenever } t > z_\alpha.$$

(b) One-sided lower tail test at level $\alpha$:

$$\textit{Reject } H_0 \textit{ in favor of } H_1 \textit{ whenever } t < -z_\alpha.$$

(c) Two-sided (tail) test at level $\alpha$:

$$\textit{Reject } H_0 \textit{ in favor of } H_1 \textit{ whenever } |t| > z_{\alpha/2}.$$

# Small Sample Test for the Population Mean

### Small Sample Test for the Mean (Known Variance)

Let $x_1, \ldots, x_n$ be a realization of a random sample from $N(\mu, \sigma^2)$ with known variance $\sigma^2$. Let $\alpha \in (0, 1)$ and $z_\alpha$ be the critical value of the standard normal distribution and define

$$t = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}.$$

(a) One-sided upper tail test at level $\alpha$:

$$\textit{Reject } H_0 \textit{ in favor of } H_1 \textit{ whenever } t > z_\alpha.$$

(b) One-sided lower tail test at level $\alpha$:

$$\textit{Reject } H_0 \textit{ in favor of } H_1 \textit{ whenever } t < -z_\alpha.$$

(c) Two-sided (tail) test at level $\alpha$:

$$\textit{Reject } H_0 \textit{ in favor of } H_1 \textit{ whenever } |t| > z_{\alpha/2}.$$

# Small Sample Test for the Population Mean (Cont.)

## Small Sample Test for the Mean (Unknown Variance)

Let $x_1, \ldots, x_n$ be a realization of a random sample from $N(\mu, \sigma^2)$ with unknown variance $\sigma^2$. Let $\alpha \in (0,1)$ and $t_{n-1,\alpha}$ be the critical value of the $t$-distribution with $n-1$ degrees of freedom and define

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}, \qquad \text{where} \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

(a) One-sided upper tail test at level $\alpha$:

*Reject* $H_0$ *in favor of* $H_1$ *whenever* $t > t_{n-1,\alpha}$.

(b) One-sided lower tail test at level $\alpha$:

*Reject* $H_0$ *in favor of* $H_1$ *whenever* $t < -t_{n-1,\alpha}$.

(c) Two-sided (tail) test at level $\alpha$:

*Reject* $H_0$ *in favor of* $H_1$ *whenever* $|t| > t_{n-1,\alpha/2}$.

# Small Sample Test for the Population Mean (Cont.)

*Derivation the 'rejection rule' of the two-sided (tail) test at level $\alpha$. (The other cases are similar. Convince yourself!)*

- We want to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ at a $\alpha$ sig. level when the variance $\sigma^2$ is unknown, the data are normally distributed, and the sample is small.
  - The idea and basic arguments are identical to the case of a known variance and a large sample, except that we will use the critical values from the $t_{n-1}$-distribution.
- **If $H_0$ is true**, the $X_i$'s follow $N(\mu_0, \sigma^2)$ and hence the test statistic is **$\mathbf{t(n-1)}$-distributed**, i.e.

$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t(n-1), \quad \text{where} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Therefore,

$$P\left(-t_{n-1,\alpha/2} \leq \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \leq t_{n-1,\alpha/2} \mid H_0 \text{ is true}\right) = 1 - \alpha, \tag{5}$$

where $t_{n-1,\alpha/2}$ is the $\alpha/2$-critical value of the $t(n-1)$-distribution.

# Small Sample Test for the Population Mean (Cont.)

- Combining eq. (3) and (5), we conclude that

$$\frac{c_l - \mu_0}{s_n/\sqrt{n}} = -t_{n-1,\alpha/2} \qquad \text{and} \qquad \frac{c_u - \mu_0}{s_n/\sqrt{n}} = t_{n-1,\alpha/2}$$

$$\iff \quad c_l = \mu_0 - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}} \qquad \text{and} \qquad c_u = \mu_0 + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}.$$

- Thus, we will reject $H_0$ whenever

$$\bar{x}_n > \mu_0 + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}} \qquad \text{or} \qquad \bar{x}_n < \mu_0 - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}.$$

- This is equivalent to the more succinct formulation:

$$\text{Reject } H_0 \text{ whenever } \left|\frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}\right| > t_{n-1,\alpha/2}.$$

# Comments on Tests for the Population Mean

- Instead of comparing the realized test statistic $t$ against the critical value of the standard normal/ the $t$-distribution with $n-1$ degrees of freedom, one can equivalently compare the $p$-value corresponding to $t$ with the significance level $\alpha$:

  *Reject $H_0$ in favor of $H_1$ whenever $p$-value corresponding to $t \leq \alpha$.*

- The small sample tests are "exact" in the sense that the type I error is exactly controlled

$$P\left(\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} > z_\alpha \mid \mu = \mu_0\right) = 1 - \Phi(z_\alpha) = \alpha,$$

  and

$$P\left(\frac{\bar{X}_n - \mu_0}{s_n/\sqrt{n}} > t_{n-1,\alpha} \mid \mu = \mu_0\right) = \alpha,$$

  etc. ...

- The large sample tests are "asymptotically exact" because the type I error is only controlled approximately thanks to the CLT.

# Outline

# Type II Error, Sample Size, and Variance

*If the data $X_1, \ldots, X_n$ are normally distributed with known variance $\sigma^2$ we can compute explicitly the probability of committing a type II error.*

| Alternative Hypothesis | Type II Error Probability for given $\alpha$ |
|---|---|
| $H_a : \mu > \mu_0$ | $\beta(\mu) = \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right)$ |
| $H_a : \mu < \mu_0$ | $\beta(\mu) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_\alpha\right)$ |
| $H_a : \mu \neq \mu_0$ | $\beta(\mu) = \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{\alpha/2}\right)$ |

- For fixed sample size $n$, decreasing Type I error increases Type II error.
- For fixed sample size $n$ and fixed $\alpha$, the type II decreases as the variance $\sigma^2$ decreases.
- For fixed variance $\sigma^2$ and fixed $\alpha$, the type II decreases as the sample size increases.

# Type II Error, Sample Size, and Variance (Cont.)

*(Derivation.)*

# Type II Error, Sample Size, and Variance (Cont.)

- For a fixed $\alpha$ (significance level) type II error decreases as the sample size increases.

- Consider the upper tail test at level $\alpha$. Then, the probability of type II error is
$$\beta = \beta(\mu) = \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right).$$

- For a fixed $\beta$ (amount of tolerated type II error probability) we can determine the necessary sample size.
$$\beta = \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right) \quad \Longrightarrow \quad -z_\beta = \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha$$
$$\Longrightarrow \quad n = \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu}\right)^2$$

- Think about how this looks like for other tests...

# Outline

# Example: Flex-time in Tesla Factory Fremont, CA

*The Tesla Factory in Fremont, CA introduces a "flex-time" plan to verify its effectiveness in reduction of "absenteeism", a random sample of 100 employees is drawn; it shows an average number of days off-work of 5.4 days with an SD of 3 days. Before introducing the plan, the average number of days off-work was 6.3 days.*

1.1 Is there any statistical difference before and after introducing the plan? Answer the question at $\alpha = 5\%$.

1.2 What is the power of the test at $\mu = 6$?

1.3 Construct the 95% confidence interval.

2.1 Does the plan reduce "absenteeism"? Formulate the hypothesis and answer the question at $\alpha = 5\%$.

2.2 At significance level $\alpha = 5\%$, what is the probability of a type II error $\beta(\mu)$ at $\mu = 6$?

2.3 At significance level $\alpha = 5\%$, what should be the sample size $n$ so that $\beta(6) = 10\%$?

2.4 Construct the 95% upper confidence bound for the population mean.

# Example: Flex-time in Tesla Factory Fremont (Cont.)

1.1 Is there any statistical difference before and after introducing the plan? Answer the question at $\alpha = 5\%$.

$$H_0 : \mu = 6.3 \qquad vs. \qquad H_1 : \mu \neq 6.3.$$

- The observed test statistic is $t = \frac{5.4-6.3}{3/\sqrt{100}} = -3$, the critical value $z_{0.025} = 1.96$.
- We reject $H_0$ because
  - $|t| > z_{0.025}$, and also
  - $p$-value $= P(|T| > 3 \mid \mu_0 = 6.3) = 2\Phi(-3) = 2 \times 0.0013 = 0.26\% < 5\%$.

  **Note:** You only need to check one of these two conditions.

# Example: Flex-time in Tesla Factory Fremont (Cont.)

1.2 What is the power of the test at $\mu = 6$?

$$f(6) = P\left(\left|\frac{\bar{X}_n - 6.3}{3/\sqrt{100}}\right| \geq 1.96 \mid \mu = 6\right)$$

$$= P(\bar{X}_n \geq 6.888 \mid \mu = 6) + P(\bar{X}_n \leq 5.712 \mid \mu = 6)$$

$$= 1 - \Phi\left(\frac{6.888 - 6}{0.3}\right) + \Phi\left(\frac{5.712 - 6}{0.3}\right)$$

$$= 17\%.$$

- **Interpretation:** If $\mu$ is the actually 6 days, the power of the test (at level 5%) is 17% and probability of a type II error is 83%.

# Example: Flex-time in Tesla Factory Fremont (Cont.)

1.3 Construct the 95% confidence interval.

$$5.4 \pm 1.96 \times 0.3 = 5.4 \pm 0.588 \approx [4.8; 6.0]$$

$\Longrightarrow$ The mean $\mu_0 = 6.3$ is not included in this 95%-CI.

- **In general:** Consider the two-sided testing problem for the mean

$$H_0 : \mu = \mu_0 \qquad vs. \qquad H_1 : \mu \neq \mu_0.$$

If we test $H_0$ versus $H_1$ at level $\alpha$, then

two sided $100(1-\alpha)\%$ CI does not cover $\mu_0$ $\iff$ reject $H_0$ at level $\alpha$.

# Example: Flex-time in Tesla Factory Fremont (Cont.)

2.1 Does the plan reduce "absenteeism"? Formulate the hypothesis and answer the question at $\alpha = 5\%$.

$$H_0 : \mu = 6.3 \qquad vs. \qquad H_1 : \mu \leq 6.3.$$

- The observed test statistic is (as before!) $t = \frac{5.4-6.3}{3/\sqrt{100}} = -3$, the critical value is now (!) $z_{0.05} = 1.645$.
- We reject $H_0$ because
  - $t < -z_{0.05}$, and also
  - $p$-value $= P(T < -3 \mid \mu_0 = 6.3) = \Phi(-3) = 0.13\% < 5\%$.

  **Again:** You only need to check one of these two conditions.

# Example: Flex-time in Tesla Factory Fremont (Cont.)

2.2 At significance level $\alpha = 5\%$, what is the probability of a type II error at $\mu = 6$?

Recall that $z_{0.05} = 1.645$. Therefore, the power of the test at $\mu = 6$ is

$$f(6) = P\left(\frac{\bar{X}_n - 6.3}{3/\sqrt{100}} \leq -1.645 \mid \mu = 6\right)$$

$$= P(\bar{X}_n \leq 5.807 \mid \mu = 6)$$

$$= \Phi\left(\frac{5.807 - 6}{0.3}\right) = 26\%,$$

and the probability of a type II error is

$$\beta(6) = 1 - 26\% = 74\%.$$

# Example: Flex-time in Tesla Factory Fremont (Cont.)

2.3 At significance level $\alpha = 5\%$, what should be the sample size $n$ so that $\beta(6) = 10\%$?

$$\beta(\mu) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{0.05}\right) = 10\%$$

$$\iff \quad \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{0.05}\right) = 0.9$$

$$\iff \quad \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{0.05} = -z_{0.9} = z_{0.1}$$

$$\iff \quad n = \left(\frac{\sigma(z_{0.1} + z_{0.05})}{\mu_0 - \mu}\right)^2$$

Thus, the required sample size is approximately

$$n \approx \left(\frac{3(1.282 + 1.645)}{6.3 - 6}\right)^2 = 856.7.$$

# Example: Flex-time in Tesla Factory Fremont (Cont.)

2.4 Construct the upper 95% CI for the mean.

$$(-\infty, 5.4 + 1.645 \times 0.3] = (-\infty, 5.89]$$

$\Longrightarrow$ The mean $\mu_0 = 6.3$ is not included in this upper 95% CI.

- **In general:** Consider the one-sided lower tail testing problem for the mean

$$H_0 : \mu \geq \mu_0 \qquad vs. \qquad H_1 : \mu < \mu_0.$$

If we test $H_0$ versus $H_1$ at level $\alpha$, then

upper $100(1-\alpha)\%$ CI does not cover $\mu_0$ $\iff$ reject $H_0$ at level $\alpha$.