Lecture 10: Z-test, T-test, and test the difference in the mean

Lecturer: Jie Fu

EEL 3850

Null hypothesis testing. H_0: ?

Binary : H_0, H_1

Z-test: Example

z-score in z test: $P\left(\frac{\hat{\mu}_X - 1000}{\sqrt{50}} \le \frac{990-1000}{\sqrt{50}}\right)$

z-score.

- Example: A battery manufacturer claims that the ~~average~~ lifespan of their batteries is 1000 hours on average and standard deviation of 50 hours. A consumer protection agency takes a random sample of 50 batteries and finds that the sample mean lifespan is $\boxed{990}$ hours.
- We want to test whether the manufacturer's claim is true at a **5% significance level**.

Null H_0: claim true. X lifespan $\mu_X = 1000$, $\sigma_X = 50$

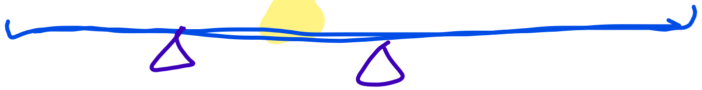
P-value observation: $P(\text{observation} \mid H_0) < 5\%$

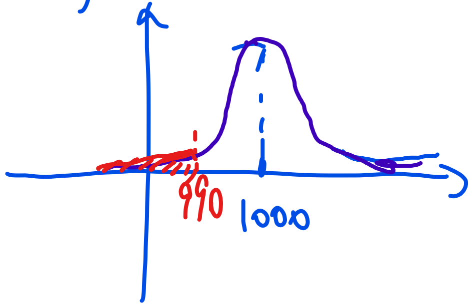
a sample of 50. a sample of $\hat{\mu}_X \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$

$n=50$

$\times$ $P(\hat{\mu}_X = 990 \mid H_0)$ $\hat{\mu}_X \sim N\left(1000, \frac{50^2}{50}\right)$

? $\Rightarrow P(\hat{\mu}_X \le 990 \mid H_0)$





$N(1000, 50)$

Z-test

sample value from observation

Z-test: whether the sample mean $\boxed{\bar{X}}$ differs significantly from a known population mean μ_0.

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

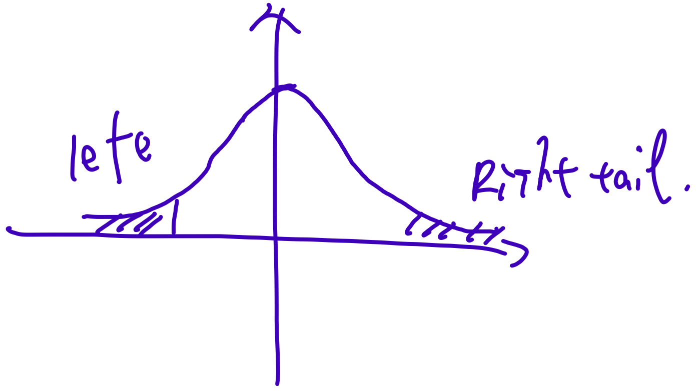
$G_0 \sim N(0, 1)$

$$z = \frac{990 - 1000}{50/\sqrt{50}} = \frac{-10}{\sqrt{50}}$$

$$P(\text{observation} \mid H_0) = G_0.\text{cdf}(z) \overset{?}{<} 0.05$$

One-sided or two-sided

- Use a one-sided test when you want to test whether the population parameter is greater than or less than a specific value.
  - **Left-tailed test**: $P(Z < z\text{-score} \mid H_0)$
  - **Right-tailed test:** $P(Z > z\text{-score} \mid H_0)$



- A **one-sided test** is used when your research hypothesis is **directional**—that is, you expect the parameter to be either greater than or less than a specific value, but not both.
  - if prior research or theory strongly suggests that a new drug increases response time compared to a known standard, you would use a one-sided test to check if the mean response time is significantly higher.
  - If our study strongly suggests that the manufacture may be lie about the average lifetime about their product, then you would use a one-sided test to check if the lifetime is significantly lower than what they claims.

(significantly higher/lower) → with (rare event) probability smaller than a given threshold.

Two-sided test

$\hat{\mu}_x \sim N(\mu_x, \frac{\sigma_x^2}{n})$

- Use when you are testing for any deviation from a specified value without assuming a direction.
  - For instance, if you are testing whether a new process leads to a different average performance (without a clear expectation of it being higher or lower), a two-sided test is used.

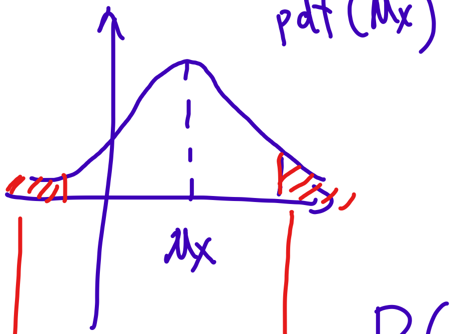
Example: Evaluating the effect of a curriculum re-design on student performance.

- 1. The previous average score was 75.
- 2. We collect a sample of 100 students who followed the new curriculum, and their average score is 78.
- 3. The population standard deviation is known to be 10.

a sample

$\hat{\mu}_x \sim N(75, \frac{10^2}{100})$

Is the new mean significantly different from the previous mean? With significance level 0.05.



$$P(\underbrace{|\hat{\mu}_x - \mu_x|}_{\text{difference}} > \text{observed difference} \mid H_0) \overset{?}{>} 0.05 \text{ : reject } H_0$$

$$< \text{ : not reject}$$

$$P(|\hat{\mu}_x-\mu_x| > 78-75) = P(\hat{\mu}_x-\mu_x>3) + P(\hat{\mu}_x-\mu_x<-3) \overset{?}{<} 0.05$$

$$2\cdot P(\hat{\mu}_x-\mu_x>3) \overset{?}{<} 0.05$$

$$P(\hat{\mu}_x-\mu_x>3) < 0.025$$

T-test: Unknown variance.

$\mu_x = 1000$? σ_x

- A battery manufacturer claims that the **average lifespan** of their batteries is **1000 hours**. A consumer protection agency takes a **random sample of ~~10~~ 50 batteries** and the lifetime is reported as follows:

array([960., 1074., 947., 974., 971., 1075., 1008., 1048., 1026., 993., 1016., 1035., 974., 953., 986., 937., 1002., 928., 989., 1010., 1037., 974., 964., 1005., 985., 1002., 1033., 1085., 925., 1013., 881., 986., 951., 1003., 973., 994., 1023., 991., 945., 881., 919., 995., 1055., 996., 965., 1001., 1097., 1015., 953., 1042.])

sample mean $\hat{\mu}_x \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$

σ_x^2 unknown $\rightarrow$ estimate $\sigma_x^2 \rightarrow \hat{\sigma}_x^2$

$\hat{\mu}_x \sim T_v\left(\mu_x, \frac{\hat{\sigma}_x^2}{n}\right)$ dof $:= n-1 = 49$

? $P(\hat{\mu}_x <$ sample_mean_from_data $| H_0)$

$P(\hat{\mu}_x < 991 | H_0)$

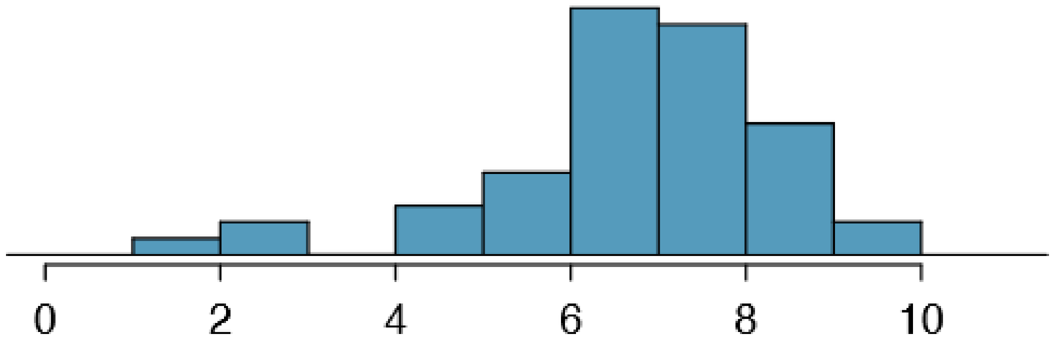
$P\left(\frac{\hat{\mu}_x - 1000}{\hat{\sigma}_x/\sqrt{n}} < \frac{991 - 1000}{\hat{\sigma}_x/\sqrt{n}}\right) = T_0.\text{cdf}\left(\frac{991-1000}{\hat{\sigma}_x/\sqrt{n}}\right) < 0.05$

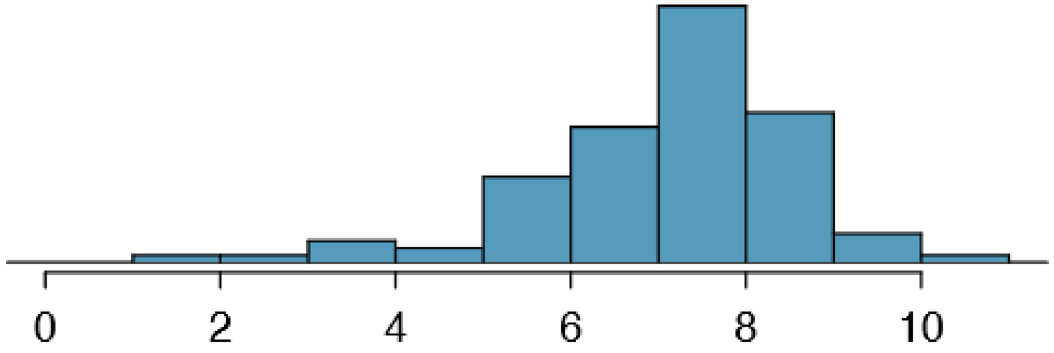
$\rightarrow$ t-value.

reject.

Test of difference in the mean

- A data set called baby smoke represents a random sample of 150 cases of mothers and their newborns in North Carolina over a year.
- We are particularly interested in two variables: weight and smoke.
- The weight variable represents the weights of the newborns and the smoke variable describes which mothers smoked during pregnancy.
- We would like to know if there is convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?





Hypothesis testing: Difference in the mean

- The null hypothesis represents the case of no difference between the groups.

- H_0: There is no difference in average birth weight for newborns from mothers who did and did not smoke.

X : birth weight from smoke mom. μ_X, σ_X

Y : birth weight " non-smoker mom. μ_Y, σ_Y

$\Rightarrow \mu_X = \mu_Y$.

- H_1: There is some difference in average newborn weights from mothers who did and did not smoke.

$\mu_X \neq \mu_Y$

Testing a difference of two means

- Testing the difference of two means:

$n_X = n$ $\quad$ $n_Y = m$

Given two data samples $\mathbf{X} = [X_0, X_1, \cdots, X_{n_X-1}]$ and $\mathbf{Y} = [Y_0, Y_1, \cdots, Y_{n_Y-1}]$, where all data samples are assumed to be independent

$\hat{\mu}_X \sim N(\mu_X, \frac{\sigma_X^2}{n})$

$\hat{\mu}_Y \sim N(\mu_Y, \frac{\sigma_Y^2}{m})$

- Test statistics: difference between sample mean estimators.

$$N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}) \sim \quad T = \hat{\mu}_X - \hat{\mu}_Y.$$

- Given the observed difference t, what is the probability to observe a difference this big assuming the null hypothesis is true?

$H_0: \mu_X = \mu_Y.$ $\quad$ $T \sim N(0, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$

$$P(T \geq t | H_0) \text{ (one-sided) or}$$
$$P(|T| \geq t | H_0) \text{ (two-sided)}$$

$$\mathbb{E}(T) = \mathbb{E}(\hat{\mu}_X - \hat{\mu}_Y) = \mathbb{E}(\hat{\mu}_X) - \mathbb{E}(\hat{\mu}_Y) = \mu_X - \mu_Y.$$

$$\mathrm{Var}(T) = \mathrm{Var}(\hat{\mu}_X - \hat{\mu}_Y) = \mathrm{Var}(\hat{\mu}_X) + \mathrm{Var}(\hat{\mu}_Y) = \frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{m}$$

Testing a difference of two means

observation is a sample from T.

- What we need to determine? $T = \hat{\mu}_X - \hat{\mu}_Y . \sim N\left(0, \underbrace{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}_{}\right)$

$\hat{\mu}_X' - \hat{\mu}_Y'$

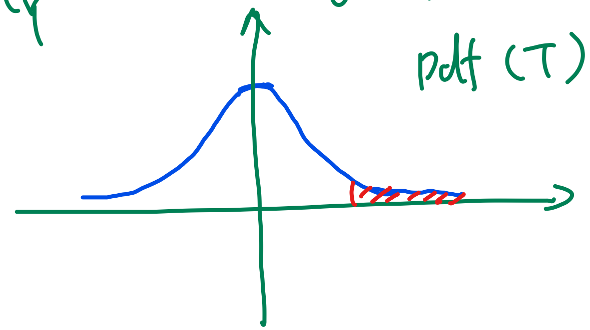
constructed from H_0.

Left side $P(T > \text{observed} \mid H_0) < 0.05$

? right side $P(T < \text{observed} \mid H_0)$

Two-side $P(|T| > \text{observed} \mid H_0)$

pdf (T)



- The observed difference t and some other information can be obtained from data:

$P(T > 0.4 \mid H_0)$

$P(T < -0.4 \mid H_0)$

$6.78 - 7.18 = -0.4$

$7.18 - 6.78 = 0.4$

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

Testing a difference of two means

What is the distribution of T?

- Given X_i, Y_k have the same, known variance?
- Given X_i, Y_k have the different but known variances?
- Given X_i, Y_k's variances are unknown but equal?
- Given X_i, Y_k' s variances are unknown and unequal?

$$T \sim N\left(0, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$$

$\sigma_X = \sigma_Y$

$$T \sim N\left(0, \sigma_X^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

Some preliminaries

- Let X_i be i.i.d. RV, what is the mean and variance of $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$?
- The linear combination of two independent Gaussian RVs. is a Gaussian RV.

Testing a difference of two means: **Different but known variances**

Given two data samples $\mathbf{X} = [X_0, X_1, \cdots, X_{n_X-1}]$ and $\mathbf{Y} = [Y_0, Y_1, \cdots, Y_{n_Y-1}]$, w[redacted]re all data [redacted]

$$T = \hat{\mu}_X - \hat{\mu}_Y .$$

Testing a difference of two means: **same, known variance**

Given two data samples $\mathbf{X} = [X_0, X_1, \cdots, X_{n_X - 1}]$ and $\mathbf{Y} = [Y_0, Y_1, \cdots, Y_{n_Y - 1}]$, where all data samples are assumed to be independent

Assume X_i, Y_k share the same variance σ^2.

$$T = \hat{\mu}_X - \hat{\mu}_Y.$$

Testing a difference of two means: **Unknown and ~~un~~equal variances**

Given two data samples $\mathbf{X} = [X_0, X_1, \cdots, X_{n_X-1}]$ and $\mathbf{Y} = [Y_0, Y_1, \cdots, Y_{n_Y-1}]$, where all data samples are assumed to be independent

Assume X_i, Y_k share the same variance σ^2.

- However, we don't know the variance.

T

estimate variance:

$$\hat{\mu}_X \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

$$\hookrightarrow T_v\left(\mu_X, \frac{\hat{\sigma}_X^2}{n}\right)$$

$$\sigma^2 = \sigma_X^2 = \sigma_Y^2 = \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2 + \sum_{j=1}^{m}(y_j - \hat{\mu}_Y)^2\right]$$

$$T \sim T_v\left(0, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

$$\text{dof} = n+m-2.$$

Testing a difference of two means: **Unknown and equal variances**

Given two data samples $\mathbf{X} = [X_0, X_1, \cdots, X_{n_X-1}]$ and $\mathbf{Y} = [Y_0, Y_1, \cdots, Y_{n_Y-1}]$, w[redacted]re all data [redacted]

- However, we don't know the variance for either?

$$T \sim T_v\left(\mu_X - \mu_Y = 0, \frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{m}\right)$$

dof :

Testing a difference of two means: **Unknown and unequal variances**

If we cannot assume that the variances are equal, then the distribution of the following normalized form will be approximately equal to a Student's t distribution:

$S_x = \hat{\sigma}_x$ $S_y = \hat{\sigma}_y$

$$\frac{T}{S_d} \sim t_\nu,$$

where

$$S_d = \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}.$$

The standard error of the difference of two sample means

Here, the value of ν must be determined from the sample standard errors of the means, which we denote by

$S_{\bar{x}}^2 = \frac{\hat{\sigma}_x^2}{n}$ $S_{\bar{y}}^2 = \frac{\hat{\sigma}_y^2}{m}$

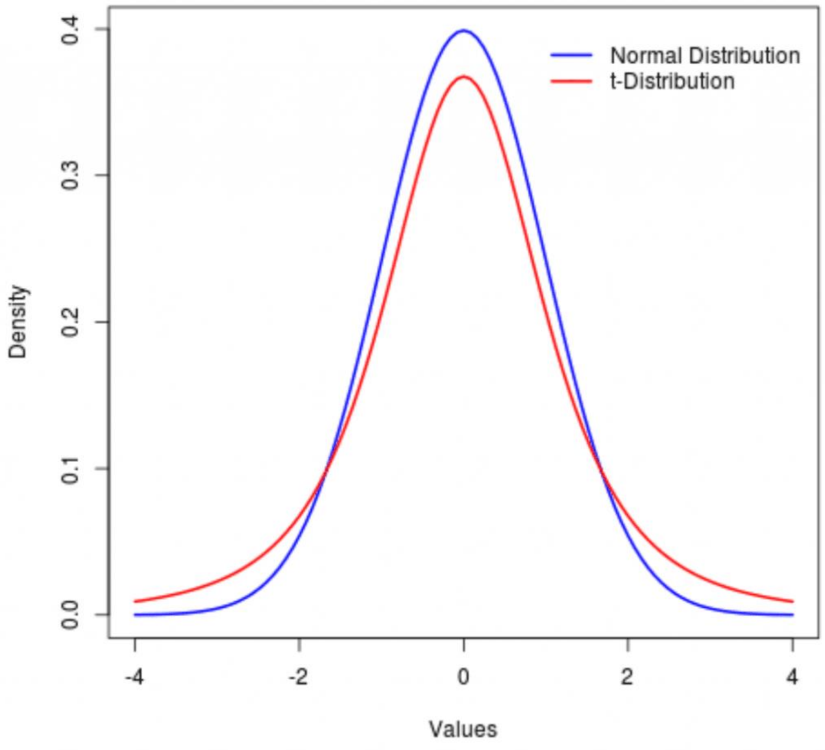
$$s_{\bar{x}}^2 = s_x^2/n_x, \text{ and}$$
$$s_{\bar{y}}^2 = s_y^2/n_x, \text{ and}$$

standard error of a single mean

Then ν is the largest integer that satisfies

$$\nu \le \frac{\left(s_{\bar{x}}^2 + s_{\bar{y}}^2\right)^2}{s_{\bar{x}}^4/(n_x - 1) + s_{\bar{y}}^4/(n_y - 1)}.$$

Testing a difference of two means: **Unknown and unequal variances**



- For sample sizes greater than 30, the differences between the t distribution and the normal distribution are negligible.

- In other words, we can assume, when the sample size > 30,

$$\frac{T}{S_d} \sim N(0,1)$$

Example

- Mother smoke related to baby weights.

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

- Observation: no knowledge about the underlying distribution: unknown and unequal variance

- The standard error of a single mean is:

- The standard error of the difference of two sample means:

Example (cont.)