

EEL 4930 Stats – Lecture 28

EEL 4930 Stats – Lecture 28

BINARY HYPOTHESIS TEST WITH KNOWN, EQUAL VARIANCES

- Now suppose we have samples from two populations

$$x_i, \quad i = 0, 1, \dots, m - 1$$

and

$$y_j, \quad j = 0, 1, \dots, n - 1$$

EEL 4930 Stats – Lecture 28

BINARY HYPOTHESIS TEST WITH KNOWN, EQUAL VARIANCES

- Now suppose we have samples from two populations

$$x_i, \quad i = 0, 1, \dots, m - 1$$

and

$$y_j, \quad j = 0, 1, \dots, n - 1$$

where x_i and y_j are observed values of random variables X_i and Y_j

EEL 4930 Stats – Lecture 28

BINARY HYPOTHESIS TEST WITH KNOWN, EQUAL VARIANCES

- Now suppose we have samples from two populations

$$x_i, \quad i = 0, 1, \dots, m - 1$$

and

$$y_j, \quad j = 0, 1, \dots, n - 1$$

where x_i and y_j are observed values of random variables X_i and Y_j , which are assumed to have common variance σ^2

known

- Let the averages of the data be

$$\bar{x} = \frac{1}{m} \sum_{i=0}^{m-1} x_i, \text{ and } \bar{y} = \frac{1}{n} \sum_{i=0}^{n-1} y_i,$$

and denote the true means of the distributions as μ_X and μ_Y , respectively

- Let the averages of the data be

$$\bar{x} = \frac{1}{m} \sum_{i=0}^{m-1} x_i, \text{ and } \bar{y} = \frac{1}{n} \sum_{i=0}^{n-1} y_i,$$

and denote the true means of the distributions as μ_X and μ_Y , respectively

- Note that if the number of samples from each population is relative large (≥ 10), then even if the original population does not have a Gaussian distribution, the averages will still be approximately Gaussian

- If $\bar{x} \neq \bar{y}$, how can we conduct a binary hypothesis test on whether the two populations have different means?

- If $\bar{x} \neq \bar{y}$, how can we conduct a binary hypothesis test on whether the two populations have different means?
- What is the null hypothesis?

The means are the same

$$\mu_x = \mu_y$$

- Under the null hypothesis, compute the difference in the sample averages and determine the probability that a difference that large would be observed under the null hypothesis

- Under the null hypothesis, compute the difference in the sample averages and determine the probability that a difference that large would be observed under the null hypothesis
- Thus, our test statistic is the difference in averages,

$$t = \bar{x} - \bar{y}$$

- Let $\hat{\mu}_X$ and $\hat{\mu}_Y$ be the sample means of random samples of sizes m and n from $f_X()$ and $f_Y()$, respectively.

- Let $\hat{\mu}_X$ and $\hat{\mu}_Y$ be the sample means of random samples of sizes m and n from $f_X()$ and $f_Y()$, respectively. We can view t as an instantiation of

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

- Let $\hat{\mu}_X$ and $\hat{\mu}_Y$ be the sample means of random samples of sizes m and n from $f_X()$ and $f_Y()$, respectively. We can view t as an instantiation of

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

- If $\mu_x = \mu_y = \mu$, then $E[\hat{\mu}_X] = E[\hat{\mu}_Y] = \mu$

sample mean
estimator is
unbiased

- Let $\hat{\mu}_X$ and $\hat{\mu}_Y$ be the sample means of random samples of sizes m and n from $f_X()$ and $f_Y()$, respectively. We can view t as an instantiation of

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

- If $\mu_x = \mu_y = \mu$, then $E[\hat{\mu}_X] = E[\hat{\mu}_Y] = \mu$
- Then (by linearity)

$$E[T] = E[\hat{\mu}_X - \hat{\mu}_Y] = E[\hat{\mu}_X] - E[\hat{\mu}_Y] \\ = \mu - \mu = 0$$

- We can compute the variance of T under the null hypothesis as

$$\text{Var}[T] = \text{Var}[\hat{\mu}_x - \hat{\mu}_y]$$

$$= \text{Var}[\hat{\mu}_x + (-\hat{\mu}_y)]$$

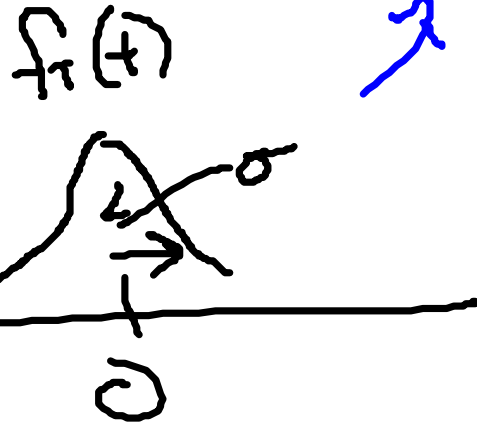
s.i.

$$= \text{Var}[\hat{\mu}_x] + \text{Var}[-\hat{\mu}_y]$$

$$= \text{Var}[\hat{\mu}_x] + (-1)^2 \text{Var}[\hat{\mu}_y]$$

$$= \frac{\sigma^2}{m} + \frac{\sigma^2}{n}$$

$$= \left[\frac{1}{m} + \frac{1}{n} \right] \sigma^2 \leftarrow \sigma_{\hat{T}}^2$$



- Finally, we can compute the probability of observing a difference in means as large as t . For convenience of discussion, assume $\bar{x} > \bar{y}$:

$$t = \text{observed difference} = \bar{x} - \bar{y} > 0$$

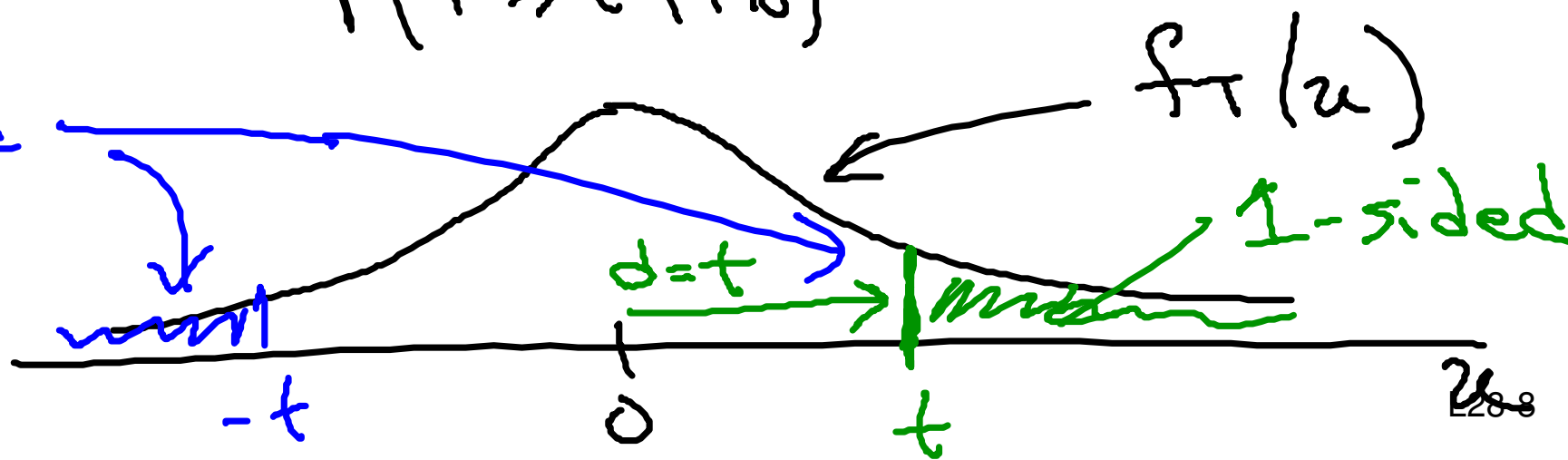
Hypothesis test:

What is $P(\text{see result as extreme under } H_0)$

1-sided

$$\rightarrow = P(T \geq t | H_0)$$

2-sided



1 sided:

$$P(T \geq t) = Q\left(\frac{t}{\sigma_T}\right)$$

$$= Q\left(\frac{t}{\sigma \sqrt{\frac{t}{3} + \frac{t}{n}}}\right)$$

2 sided:

$$P(|T| \geq t) = 2Q\left(\frac{t}{\sigma \sqrt{\frac{t}{3} + \frac{t}{n}}}\right)$$

HYPOTHESIS TESTS WITH UNKNOWN VARIANCES

- In many cases, the variance(s) of the underlying distributions are not known and must be estimated from the data

HYPOTHESIS TESTS WITH UNKNOWN VARIANCES

- In many cases, the variance(s) of the underlying distributions are not known and must be estimated from the data
- In this case, the underlying distribution is even more spread out from the mean than the Gaussian distribution.

HYPOTHESIS TESTS WITH UNKNOWN VARIANCES

- In many cases, the variance(s) of the underlying distributions are not known and must be estimated from the data
- In this case, the underlying distribution is even more spread out from the mean than the Gaussian distribution. More of the probability is in the tails

HYPOTHESIS TESTS WITH UNKNOWN VARIANCES

- In many cases, the variance(s) of the underlying distributions are not known and must be estimated from the data
- In this case, the underlying distribution is even more spread out from the mean than the Gaussian distribution. More of the probability is in the tails
- The first step is to determine how to estimate the variance. Any ideas?

$$x_0, x_1, \dots, x_{n-1} \leftarrow x_i (\mu_x, \sigma_x^2)$$

$$\overline{\sigma_x^2} = ?$$

$$\bar{x} = \frac{1}{n} \sum_{i=0}^{n-1} x_i$$

$$\sigma_x^2 = E[(X - \mu_x)^2]$$

$$(1) \quad \overline{\sigma_x^2} = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - \bar{x})^2$$

- In your Jupyter notebook, generate 10 Gaussian random variables with mean 10 and variance 100. Find the sample variance. Do this for 10000 simulation steps, during each of which you redraw the 10 random variables and estimate the sample mean and variance. Use the average of the sample variance over the 10000 simulations as an estimator of the true variance. What do you observe?

Estimator (1) is biased! Too low

Estimator (2):

$$s^2 = \overline{\sigma_x^2} = \frac{1}{n-1} \sum_{i=0}^{n-1} (x_i - \bar{x})^2 \text{ is } \underline{\text{unbiased}}$$

- If you want to understand this phenomenon more carefully, read the Wikipedia page on Bias of an Estimator: https://en.wikipedia.org/wiki/Bias_of_an_estimator

- If you want to understand this phenomenon more carefully, read the Wikipedia page on Bias of an Estimator: https://en.wikipedia.org/wiki/Bias_of_an_estimator
- If we use our unbiased estimator for the variance, then the distribution of

$$\frac{\hat{\mu} - \mu}{S_{n-1}/\sqrt{n}}$$

has a Student's t -distribution with $n - 1$ degrees of freedom



d.o.f.

nu

\rightarrow

$\sqrt{\quad}$

- The density and distribution functions for the Student- t distribution are shown on it's Wikipedia page: https://en.wikipedia.org/wiki/Student's_t-distribution

- The density and distribution functions for the Student- t distribution are shown on it's Wikipedia page: https://en.wikipedia.org/wiki/Student's_t-distribution
- Unlike the Gaussian distribution, the distribution function for Student's t -distribution is in closed form for several values of  

- Compare the density function of the Gaussian random variable with $\sigma^2 = 1$ with the Student's t random variable with different degrees of freedom

- Compare the density function of the Gaussian random variable with $\sigma^2 = 1$ with the Student's t random variable with different degrees of freedom
 - Why does it behave this way?