

Data collection, Analysis and Inference

Subject Code: CPE-RPE,

May 2021,
SRM Univeristy-AP, Andhrapradesh

Lecture- 7: Large Sample Tests and Small Sample Tests: Student's t-test

- Aim: To implement Student's t-test for small/large samples and infer the outcome on the hypothesis.

Suppose we wish to determine the following for students enrolled at a particular university:

- μ = (unknown) average height of all the students in the university
- p = (unknown) proportion of students whose height falls in the range 160-165 cm

In either case, we can't possibly survey the entire population (all university college students).

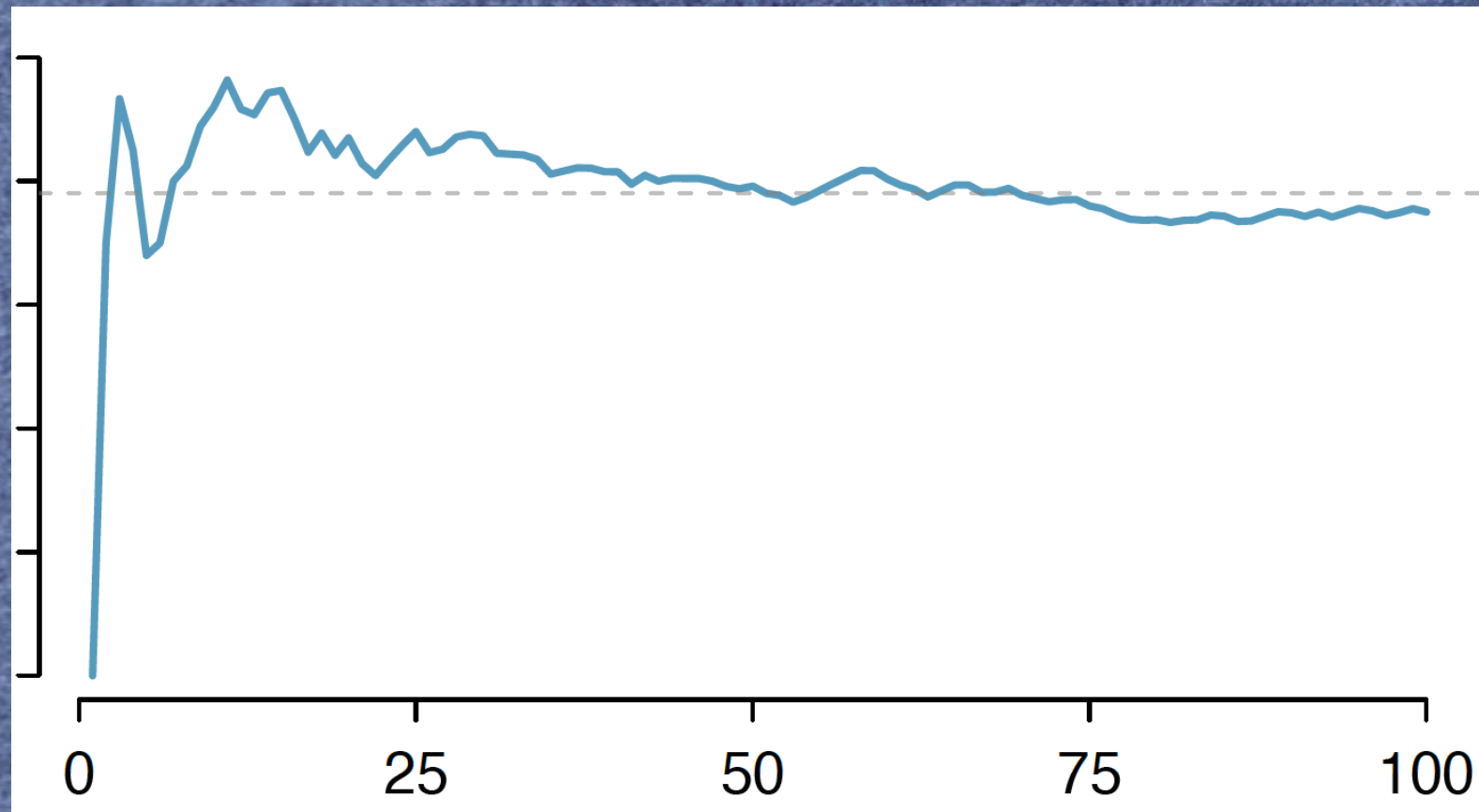
So, it is natural to take a random sample from the population, and use the resulting data to make generalizations about unknown population.

- This part of statistics is called **inferential statistics**

The estimates obtained from the sample data are called **point estimates (statistic)**.

- Estimates generally vary from one sample to another, and this sampling variation suggests our estimate may be close, but it will not be exactly equal to the parameter
- *Estimates get better as more data become available.*

We can see this by plotting a running mean



- A running mean is a sequence of means, where each mean uses one more observation in its calculation than the mean directly before it in the sequence.
- The running mean tends to approach the true population average (dotted line) as more data become available

Standard error (SE)


- If the variability between an estimate to another is small then that estimate is '*probably*' very accurate.
- If it varies widely from one sample to another, then we should not expect our estimate to be very good

This variation is measured by standard error (SE) of the corresponding estimate/statistic.

Sampling distribution of statistics

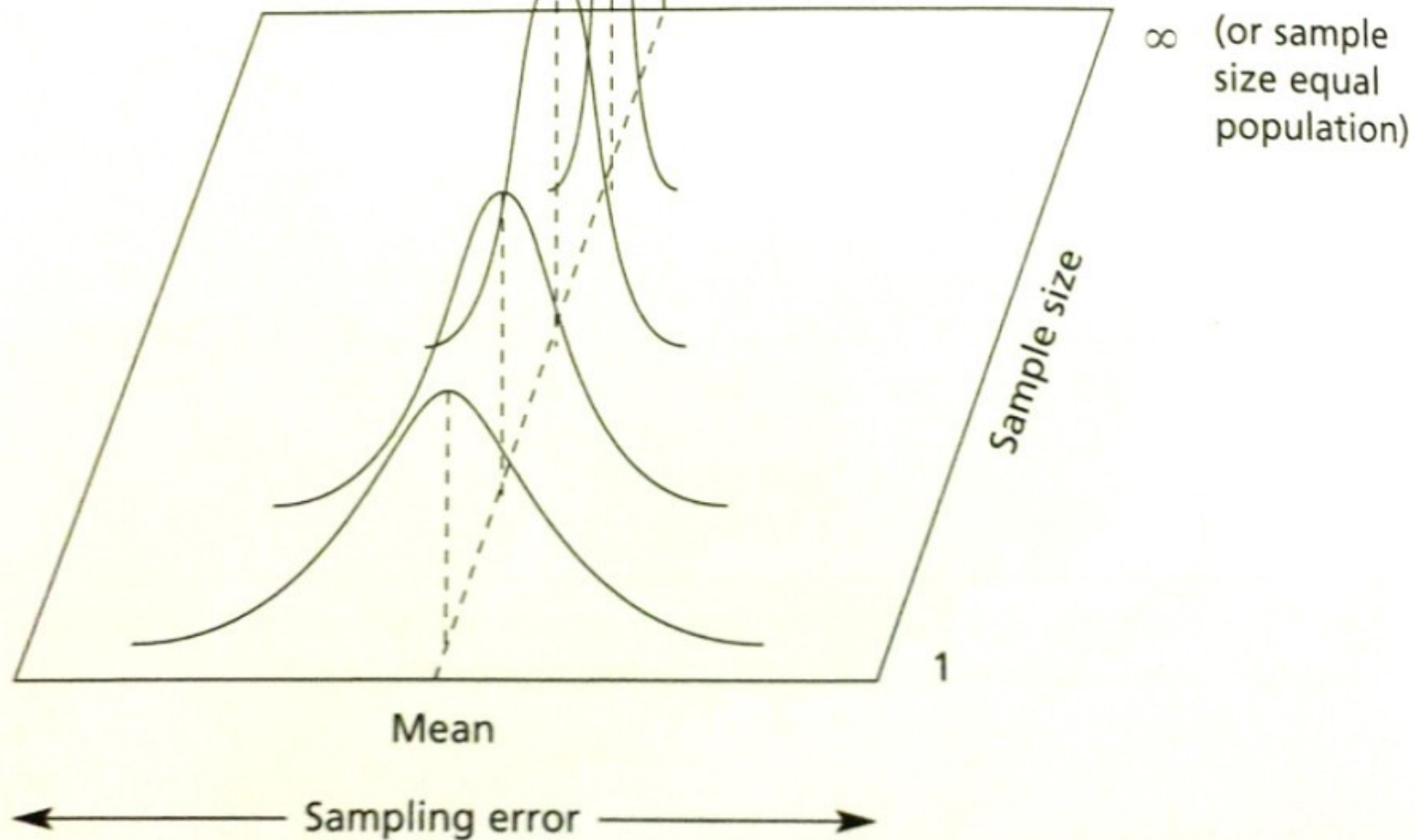
- Suppose we are interested in estimating the unknown parameter θ (may be mean, median, proportion etc.) of a population.
- From the population, start picking up random samples of fixed size, say n . That is, we pick a random sample x_1, x_2, \dots, x_n from the population.
- Calculate statistic which estimates θ . Let us denote it by $\hat{\theta}_1$.
- Replace the sample and pick another random sample of same size and do the same calculation to get $\hat{\theta}_2$

- Repeating this procedure, we get t
sampling distribution of the statistic $\hat{\theta}$

 **Standard error** of $\hat{\theta}$ is the standard deviation of the sampling distribution of $\hat{\theta}$

- To estimate the population mean μ look at the corresponding sampling distribution of sample mean \bar{X} based on samples of size n .
- Note: Sample mean is an unbiased estimate of population mean, we have $E[\bar{X}] = \mu$,
“Which means that the mean of the sampling distribution of sample mean is the population mean itself!”

Sampling distribution of the statistic μ



☞ What about standard deviation of the sampling distribution of sample mean?

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n^2} [\text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)] \quad (\text{i.i.d!}) \\ &= \frac{1}{n^2} (n \cdot \sigma^2) = \frac{\sigma^2}{n}\end{aligned}$$

where σ is the population standard deviation.

Clearly, for a given σ , as the sample size n increases, the error goes to zero and we get better estimates!

☞ The standard error of sample mean ($SE(\bar{X})$) is $\frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation and n is the sample size

Example: It is found that average weight of people residing in a particular city is 62 Kg with a standard deviation of 5.2 Kg.

If we choose a random sample of 75 people, what is probability that mean weight of the chosen sample is between 61 Kg and 64 Kg?

- To answer such questions we should know the actual distribution of sample means! How do we get to know about the distribution?

Answer: Central limit theorem!

CLT says that if $n > 30$, $X_1 + X_2 + \dots + X_n$ will be Normal with parameters $n\mu$ and $n\sigma^2$

That is, for $n > 30$,

- sample mean $\bar{X} = 1/n \sum_{i=1}^n X_i$
will be Normal with parameters $(\mu, \sigma^2/n)$

Hence, for $n > 30$, any probability involving \bar{X} can be computed using the standard normal table!

Caution: This approach is valid only if $n > 30$, that is, if the sample size is greater than 30

Example: It is found that average weight of people residing in a particular city is 62 Kg with a standard deviation of 5.2 Kg.

If we choose a random sample of 75 people, what is probability that mean weight of the chosen sample is between 61 Kg and 64 Kg?

Solution : The variable of interest, X , is the weight of people residing in the city

We are given that population mean and population SD are 62Kg and 5.2 Kg respectively

That is, $\mu = E[X] = 62$ and $\sigma = SD(X) = 5.2$

The question asks for probability involving the sample mean \bar{X} computed from random samples of size $n = 75$

Since $n > 30$, by CLT, we may assume that

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

- That is, $\bar{X} \sim N(62, (5.2)^2/75)$

$$\begin{aligned} P\{61 < \bar{X} < 64\} &= P\left\{\frac{61 - 62}{(5.2/\sqrt{75})} < Z < \frac{64 - 62}{(5.2/\sqrt{75})}\right\} \\ &= P\{-1.67 < Z < 3.33\} \\ &= \Phi(3.33) - \Phi(-1.67) \\ &= \Phi(3.33) - (1 - \Phi(1.67)) \\ &= 0.9996 - (1 - 0.9525) \\ &= 0.9521 \end{aligned}$$

CENTRAL LIMIT THEOREM FOR THE SAMPLE MEAN

When we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ , the sampling distribution of \bar{x} will be nearly normal with

$$\text{Mean} = \mu$$

$$\text{Standard Error (SE)} = \sigma/\sqrt{n}$$

Population proportion

- Till now we were talking about sampling distribution of sample mean, it's standard error etc., CLT, helped us to answer questions about probabilities involving sample mean
- What would be the standard error in estimation of population proportion with a sample of size n ?
- How Sampling distribution of sample proportion looks like?

Example: It is known that 45% of students have their heights in the range 160-165 cm. If random sample of 75 students is chosen,

what is the probability of proportion of students with height in the range 160-165 cm is greater than 40%?

- Here the variable of interest is the “proportion” of students with height in the range 160-165 cm

It is given that population proportion $p = 0.45$

and the random sample is of size 75 students.

- Let the proportion \hat{p} be the students with height in the range 160-165 cm in this sample.

Distribution of the sample proportion (\hat{p})

To answer the question about the probability, we need to know the distribution of \hat{p} .

- If the number of successes is S , then $\hat{p}=S/75$
- S is a random variable which records the number of successes in 75 trials! Thus, $S \sim \text{Bin}(75, p=0.45)$
- But Binomial probability computation is expensive, so seek the help from its close friend !
- Mean of S is $np = 75 \times 0.45 = 33.75$ and variance is $np(1 - p) = 18.5625$

By CLT, S can be approximated by normal distribution only if both np and $n(1 - p) > 5$.

- With such an approximation,

$$\hat{p} = S/n \sim N(p, p(1-p)/n)$$

- Thus, $\hat{p} \sim N(0.45, 0.0033)$, Now answer the question asking for probability of $\hat{p} > 0.4$

$$\begin{aligned} P\{\hat{p} > 0.4\} &= P\left\{\frac{\hat{p} - 0.45}{\sqrt{0.0033}} > \frac{0.4 - 0.45}{\sqrt{0.0033}}\right\} \\ &= P\{Z > -0.87\} \\ &= 1 - \Phi(-0.87) \\ &= \Phi(0.87) = 0.8078 \end{aligned}$$

Summary

Sampling distribution of sample mean

- ▶ From a population with mean μ and variance σ^2 , we draw random samples of size n
- ▶ Since sample mean (\bar{X}) varies from sample to sample, we are interested in the distribution of sample mean, which we call as **sampling distribution** of \bar{X}
- ▶ The standard deviation of sampling distribution of \bar{X} is termed as **standard error (SE)** of \bar{X}
- ▶ The mean of sampling distribution of \bar{X} will be the population mean μ and $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- ▶ CLT \implies for $n > 30$, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Summary

Sampling distribution of sample proportion

- ▶ We are interested in a proportion of population p and we draw random samples of size n to estimate it
- ▶ Different samples give different sample proportions (\hat{p}) and we looked at sampling distribution of \hat{p}
- ▶ The standard deviation of sampling distribution of \hat{p} is termed as **standard error (SE)** of \hat{p}
- ▶ The mean of sampling distribution of \hat{p} will be the population proportion p and $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$
- ▶ $n\hat{p} \sim \text{Bin}(n, p)$
- ▶ By CLT, for $np > 5$ and $n(1-p) > 5$, $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

Statistical inference

One job of a statistician is to make statistical inferences about populations based on samples taken from the population

- *Statistical inference is to make a decision about a parameter based on the sample data and it's Standard Error.*
- For instance,
 - ▶ a car dealer advertises that its new small truck gets 35 miles per gallon, on average
 - ▶ a bike company claims that 6% of employees in a city travel to their work place by riding their bikes
- ▶ a tutoring service claims that its method of tutoring helps 90% of its students get an A or a B

- ▶ a company says that women managers in their company earn an average of \$60,000 per year
- ▶ a university claims that 30% of the students stay on campus
- A statistician will make a decision about these claims (called **null hypothesis**)
- This process is called hypothesis testing
- A hypothesis test involves collecting sample data and evaluating it.

Then, the statistician makes a decision as to
“whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis”.

The actual test begins by considering two hypotheses

They are called the null hypothesis and the alternative hypothesis. These hypotheses contain opposing viewpoints

- H_0 : The null hypothesis: It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt
- H_a : The alternative hypothesis: It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0

- Since the null and alternative hypotheses are contradictory, we must examine evidence to decide if we have enough evidence to reject the null hypothesis or not
- The evidence is in the form of sample data.
- There are two options for a decision

They are “reject H_0 ” if the sample information favors the alternative hypothesis or “do not reject H_0 ” or “decline to reject H_0 ” if the sample information is insufficient to reject the null hypothesis

- The null statement must always contain some form of equality ($=$, \leq or \geq)
- Always write the alternative hypothesis, typically denoted with H_a , using less than, greater than, or not equals symbols, i.e., (\neq , $>$ or $<$).

- Example-1: We want to test whether the mean GPA of students in colleges is different from 5.0 (out of 10.0). The null and alternative hypotheses are:
 - ▶ $H_0 : \mu = 5.0$
 - ▶ $H_a : \mu \neq 5.0$
- Example-2: A quality control expert at a factory that paints car parts. He knows that 20% of parts have an error in their painting.

He recommended a change in the painting process and he wants to see if this error rate had changed. The null and alternative hypotheses will be:

- ▶ $H_0 : p = 0.2$
- ▶ $H_a : p \neq 0.2$

Example-3: According to a very large poll in 2015, about 90% of homes in California had access to the internet. Market researchers want to test if that proportion is now higher. The null and alternative hypotheses in this case will be

► $H_0 : p = 0.9$

► $H_a : p > 0.9$

Example-4: We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

► $H_0 : \mu \geq 5$

► $H_a : \mu < 5$

- Hypothesis can be one-tailed or two-tailed depending on the alternative hypothesis H_a
- We call a hypothesis to be one-tailed if the alternative hypothesis states that a parameter is larger or smaller ($>$ or $<$) than the null hypothesis value
- It is two-tailed if it states that the parameter is different (\neq) from the null value

Example-1: We want to test whether the mean GPA of students in colleges is different from 5.0 (out of 10.0). The null and alternative hypotheses are:

► $H_0 : \mu = 5.0$

► $H_a : \mu \neq 5.0$

Since H_a has \neq , the hypothesis is two-tailed

Example :

► $H_0 : p = 0.9$

► $H_a : p > 0.9$

Now H_a has $>$. So the hypothesis is one-tailed.

Example

► $H_0 : \mu \geq 5$

► $H_a : \mu < 5$

Since H_a has $<$, the hypothesis is one-tailed

Test of Significance and P-values

A popular method of testing hypothesis is by using 'P-values'

This way of testing hypothesis using P-value is usually termed as 'tests of significance'

- Every significance test starts by fixing a '**significance level**' β in $(0, 1)$
- Significance level $\beta = 1 - \alpha/100$, α being the confidence level.
- Default significance level is 0.05 (with the default confidence level 95%).

- P-value is “the probability that sample mean is more extreme to population mean than \bar{x} given that the null hypothesis is true”.

That is,

$$\text{P-value} = P\{X \text{ is more extreme to } \mu \text{ than } \bar{x} \mid \mu = 45\}$$

- What does “more extreme to μ than \bar{x} ” mean?

Example: A local pizza store knows the mean amount of time it takes them to deliver an order is 45 minutes after the order is placed. The manager has a new system for processing delivery orders, and they want to test if it changes the mean delivery time.

They take a random sample of 15 delivery orders and find their *mean delivery time is 48 minutes* with a sample standard deviation of 10 minutes.

- First thing to note from the problem is that the manager wishes to test a claim about “mean” delivery time

setup the hypotheses

► $H_0 : \mu = 45$

► $H_a : \mu \neq 45$

We are assuming that H_0 is true in the P-value definition. Hence the population mean is $\mu = 45$

More extreme to μ than $\bar{x} = 48$ means that “our sample mean \bar{X} should be more *farther* from the population mean $\mu = 45$ than $\bar{x} = 48$ ”.

- *Farther* in which direction?

This is where importance of one-tailed or two-tailed comes into play.

Since the alternative hypothesis is two-tailed,
“*farther*” in our case can be on both the sides.

$$\begin{aligned}\text{Thus, P-value} &= P \left\{ \left\{ \bar{X} - \mu > \bar{x} - \mu \right\} \cup \left\{ \bar{X} - \mu < -(\bar{x} - \mu) \right\} \right\} \\ &= P \left\{ \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} > \frac{\bar{x} - \mu}{SE(\bar{X})} \right\} \cup \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} < -\frac{\bar{x} - \mu}{SE(\bar{X})} \right\} \right\} \\ &= P \left\{ \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} > \frac{48 - 45}{SE(\bar{X})} \right\} \cup \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} < -\frac{48 - 45}{SE(\bar{X})} \right\} \right\} \\ &= P \left\{ \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} > \frac{3}{SE(\bar{X})} \right\} \cup \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} < -\frac{3}{SE(\bar{X})} \right\} \right\} \\ &= P \left\{ \left\{ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{3}{\sigma/\sqrt{n}} \right\} \cup \left\{ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < -\frac{3}{\sigma/\sqrt{n}} \right\} \right\}\end{aligned}$$

as $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation

- The problem here is that we do not know the population SD σ
- An estimate for the population SD would be s
We have $(\bar{x}-\mu)/s/\sqrt{n} = 3/2.582=1.162$
- Hence P-value will be

$$P \left\{ \left\{ \frac{\bar{X} - \mu}{s/\sqrt{n}} < -1.162 \right\} \cup \left\{ \frac{\bar{X} - \mu}{s/\sqrt{n}} > 1.162 \right\} \right\}$$

Since population sd is unknown, the distribution of

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

will be either $N(0, 1)$ or Student's t -distribution with $(n - 1)$ degrees of freedom depending on the sample size n .

- *It turns out that for large sample size (degrees of freedom), t -scores will be close to z -scores*

Hence, instead of going for different approaches for different sample sizes, we will be using t -distribution irrespective of sample size

- That is, $(\bar{X} - \mu)/s/\sqrt{n}$ follows Student's t-distribution with df $v = n - 1 = 14$.

$$\begin{aligned} \text{P-value} &= P \left\{ \left\{ \frac{\bar{X} - \mu}{s/\sqrt{n}} < -1.162 \right\} \cup \left\{ \frac{\bar{X} - \mu}{s/\sqrt{n}} > 1.162 \right\} \right\} \\ &= 2F_{14}(-1.162) \\ &= 2 \times 0.1323 = 0.2646 \end{aligned}$$

- **Decision:**
 - ▶ if $\text{P-value} < \beta$ (significance level), we reject the null hypothesis in favour of alternative hypothesis
 - ▶ If $\text{P-value} \geq \beta$, we do not reject null hypothesis *or we say that sample do not provide enough evidence against the null hypothesis.*

Decision:

- ▶ if $P\text{-value} < \beta$, reject the null hypothesis
 - ▶ If $P\text{-value} \geq \beta$, do not reject null hypothesis
-
- In our example, since no significance level is specified, we take $\beta = 0.05$ and compare it with the P-value 0.2646
 - Since $0.2646 > 0.05$, *we do not reject the null hypothesis* and conclude that the sample does not provide enough evidence to say that the average delivery time is different from 45 minutes

Example: Fernanda runs a large bowling league. She suspects that the league average score is greater than 150 per game. She takes a random sample of 36 game scores from the league data.

The scores in the sample have a mean of 156 and a standard deviation of 30. Conduct a test of significance with $\beta = 0.10$ to decide on Fernanda's suspicion.

- The hypotheses would be:
 - ▶ $H_0 : \mu \leq 150$
 - ▶ $H_a : \mu > 150$
- Here, since there is $>$ in the alternative hypothesis, the test is going to be a one-tailed calculation of P-value

“more extreme” in this case would be in only one direction!

We assume μ to be the 'border' value from null hypothesis. That is, $\mu = 150$

Hence the P-value in this case is

$$P\left\{\frac{\bar{X} - \mu}{s/\sqrt{n}} > \frac{\bar{x} - \mu}{s/\sqrt{n}}\right\} = P\left\{\frac{\bar{X} - \mu}{s/\sqrt{n}} > \frac{156 - 150}{30/\sqrt{36}}\right\} = P\left\{\frac{\bar{X} - \mu}{s/\sqrt{n}} > 1.2\right\}$$

Now we have

$(\bar{X} - \mu)/s/\sqrt{n}$ following Student's t-distribution
with $df = n - 1 = 35$

Thus, P-value =
$$P\left\{\frac{\bar{X} - \mu}{s/\sqrt{n}} > 1.2\right\} = P\left\{\frac{\bar{X} - \mu}{s/\sqrt{n}} < -1.2\right\}$$

$$= F_{35}(-1.2) = 0.1191$$

As P-value $0.1191 > \beta = 0.10$, we do not reject the null hypothesis!

That is at 0.10 significance level, the sample data does not provide enough evidence against the null hypothesis of the league average score being 150 per game

Test of significance for single population mean

☞ We will be given a sample of size n (hence its mean \bar{x} and SD s), and significance level β (default is 0.05)

Step-1: Setup the hypotheses and note down if the test is one-tailed or two-tailed

Step-2: Compute $\frac{s}{\sqrt{n}}$

Step-3: Note down the value of μ from null hypothesis and compute the '**test statistic**' $t = \left| \frac{\bar{x} - \mu}{SE} \right|$

Step-4: Compute P-value as below:

- ▶ For two-tailed test, $P\text{-value} = 2F_{n-1}(-t)$
- ▶ For one-tailed test, $P\text{-value} = F_{n-1}(-t)$

where F_{n-1} is the distribution function of a Student's t -distributed random variable with $df = n - 1$

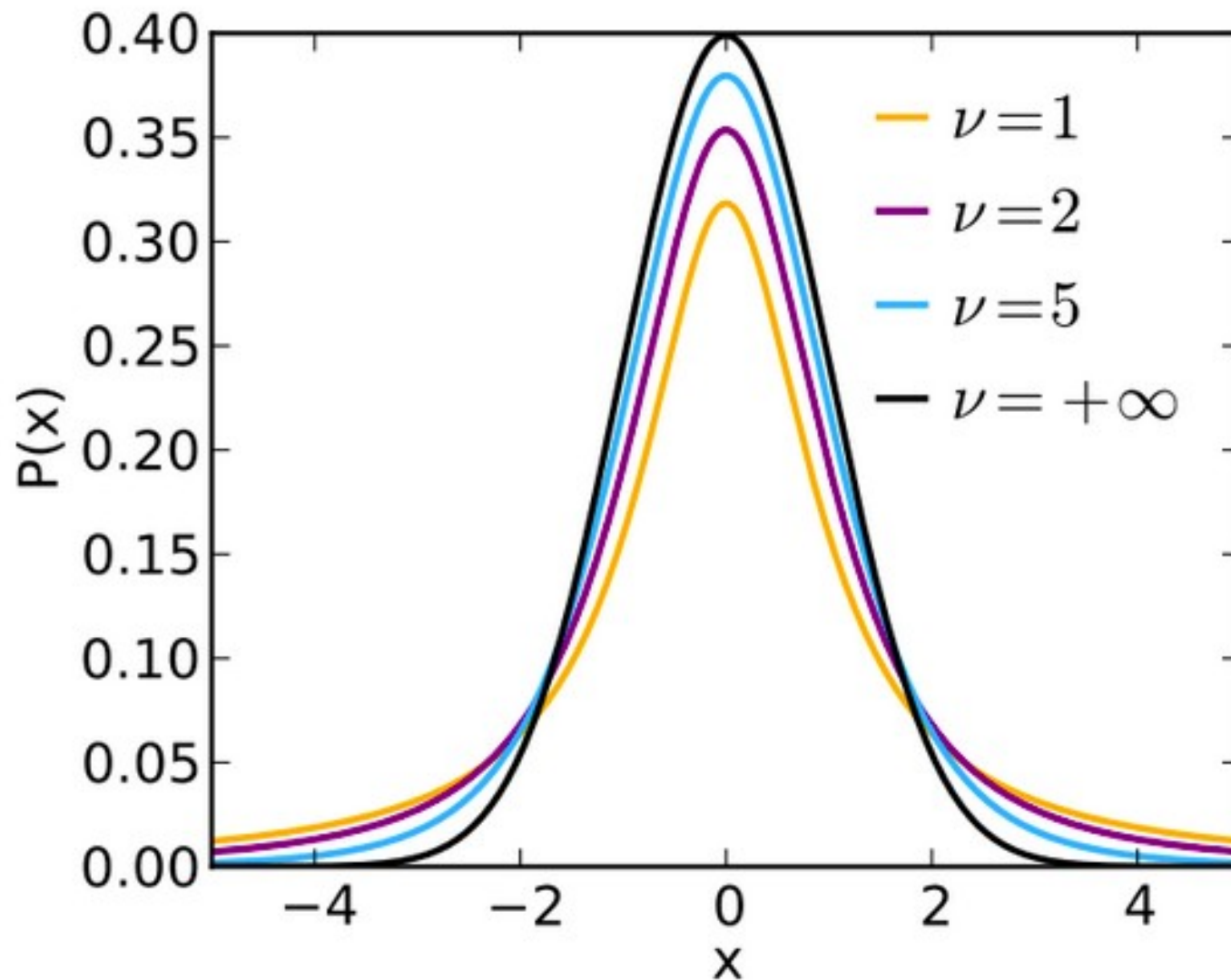
Step-5: Take a decision

- ▶ If P-value $< \beta$ (significance level), we reject the null hypothesis
- ▶ If P-value $\geq \beta$, we do not reject null hypothesis

★ Since we are using the Student's t-distribution as the '*reference distribution*' in the above test, the test is usually referred to as '*t-test for single population mean*'

Student's t

Probability density function



Probability density function

Student's t-distribution has the probability density function given by

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where ν is the number of degrees of freedom and Γ is the gamma function.

Types of errors in inference

Hypothesis tests are not flawless as we can make a wrong decision in statistical hypothesis tests based on the data.

- A Type I Error (false positive) is rejecting the null hypothesis when H_0 is actually true.
- A Type 2 Error (false negative) is failing to reject the null hypothesis when the alternative is actually true.

There are four possible scenarios, which are summarized in following table:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

There are tools in hypothesis testing which quantify these types of errors

- At a particular significance level β it turns out that

$$P(\text{Type I error}) = \beta$$
- The probability of not committing a Type II error is called the power of a hypothesis test
- That is, $\text{Power} = 1 - P(\text{Type II error})$

$$= 1 - P(\text{not rejecting } H_0 \mid H_0 \text{ is false})$$

The Truth

Test
Score:

	Has the disease	Does not have the disease
Positive	True Positives (TP) a	False Positives (FP) b
Negative	False Negatives (FN) c	True Negatives (TN) d

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

Sensitivity

$$\frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity

$$\frac{\text{TN}}{\text{TN} + \text{FP}}$$

Or,

$$\frac{a}{a + c}$$

$$\frac{d}{d + b}$$

Statistically significant

- In tests of significance, if the P-value is less than the significance level (typically 0.05), then we conclude that results are statistically significant (to reject the null hypothesis)
- Results are said to be statistically significant when the difference between the hypothesized population parameter and observed sample statistic is large enough to conclude (by P-value) that it is unlikely to have occurred by chance

END