स्टेटिस्टिक्स फॉर डाटा साइंस**- 1**

**प्रो. उषा मोहन**

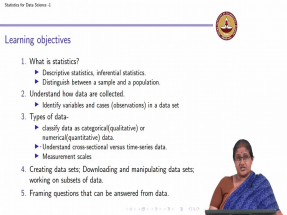
**प्रबंधन अध्ययन विभाग**

**भारतीय प्रौद्योगिकी संस्थान, मद्रास**

**लैक्चर -03 इंट्रोडक्शन एंड टाइप ऑफ डाटा पार्ट- 1**

**इस प्रथम सप्ताह में सबसे पहले आप इस पाठ्यक्रम को सीखने के उद्देश्यों को समझ ले, कि आप इस पाठ्यक्रम को क्यों सीख रहे हैं।**

**(स्लाइड समय देखें: 00:21)**



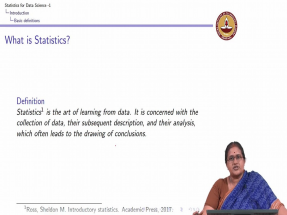
**सांख्यिकी या स्टेटिस्टिक्स क्या है? हम यहाँ आपको केवल स्टेटिस्टिक्स की दो मुख्य शाखाओं के बारे में संक्षेप में बताने जा रहे हैं, जो की इस बिन्दु पर आपके लिए प्रासंगिक हैं । हम आपको यहाँ बताएंगे कि डिसक्रिप्टिव स्टेटिस्टिक्स (descriptive statistics ) और अनुमानात्मक आंकड़ों (inferential statistics) से हम क्या समझते है।**

**जिस क्षण मैं अनुमानात्मक आँकड़ों के बारे में बात करती हूँ, मुझे आपसे सैंपल (नमूने) और पॉपुलेशन की धारणा को अवगत कराने की आवश्यकता पड़ती है। तो मैं अब आपसे इसका ही परिचय कराने जा रही हूं। इसके पश्चात हम यह समझने की कोशिश करेंगे कि हमें डाटा की आवश्यकता क्यों पड़ती है; हम इस बारे में भी समझेंगे कि डाटा कैसे एकत्र किया जाता है और हम इस बारे में भी बात करेंगे कि डाटा को डाटा सेट के रूप में कैसे व्यवस्थित किया जाता है।**

**एक बार हमारे पास डाटा सेट तैयार हो जाने के बाद हम डाटा को कैटगॉरिक्‌ल्‌ और न्यूमेरिकल (categorical and numerical) या क्रॉस-सेक्शनल (cross sectional) और टाइम सिरीज़ (time series) के रूप में वर्गीकृत करके डाटा के बारे में और अधिक समझेंगे। साथ ही साथ हम मेश़रमेंट स्केल (measurement scale) के बारे में भी चर्चा करेंगे।**

**अंत में, जैसा की मेरा मानना है कि कोई भी सांख्यिकीय विश्लेषण की कुंजी है कि आप अपने द्वारा एकत्र किए गए डाटा को समझे और एकत्र किए गए डाटा के आधार पर प्रश्नों को फ्रेम करे, तो इसके लिए हम अपने द्वारा एकत्र किए गए डाटा को समझने और डाटा के आधार पर प्रश्नों को फ्रेम करने के लिए खुद को प्रशिक्षित करने मे कुछ समय केंद्रित करेंगे। तो इस पाठ्यक्रम के पहले सप्ताह के लिए यही हमारे सीखने के उद्देश्य हैं।**

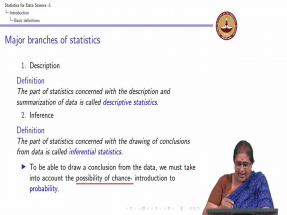
**(स्लाइड समय देखें: 01:36)**



**स्टेटिस्टिक्स क्या है? यदि आप वर्षों से विकसित स्टेटिस्टिक्स या सांख्यिकी की परिभाषाओं को देखें, तो आप आसानी से समझ सकेंगे कि स्टेटिस्टिक्स कि परिभाषा मे निरंतर परिवर्तन हुआ है और वह समय के साथ बदल रही है। जो शुरुआत मे केवल डाटा को सारांशित (summarize) करने के रूप में परिभाषित हुई थी, बाद में धीरे-धीरे, सांख्यिकी की परिभाषा, डाटा से अनुमान लगाने में परिवर्तित हुई और अब आज के समय मे बहुत सारा डाटा उपलब्ध होने के कारण, सांख्यिकी को डाटा से सीखने की कला के रूप में फिर से परिभाषित किया जा रहा है।**

**जब मैं कहती हूं कि डाटा से सीखना, तो इसमें यह शामिल है कि आप डाटा से कुछ जानकारी प्राप्त करना चाहते हैं। शेल्डन रॉस ने सांख्यिकी को डाटा से सीखने की कला के रूप में परिभाषित किया है । आप डाटा के संग्रह, डाटा के संग्रह के उपरांत डाटा के विवरण और विश्लेषण मे दिलचस्पी रखते हैं जो अन्तः हमे डाटा से निष्कर्ष निकालने की ओर ले जाता है। इसलिए सांख्यिकी और सांख्यिकीय विश्लेषण का मुख्य उद्देश्य वास्तव में डाटा के आधार पर निष्कर्ष निकालना है।**

**(स्लाइड समय देखें: 02:44)**

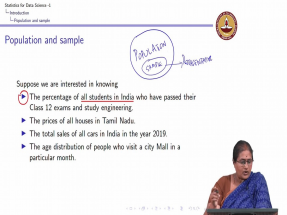


**भले ही स्टेटिस्टिक्स या सांख्यिकी की नई-नई शाखाएँ उत्पन्न हो रही हो और सांख्यिकी को नए शीर्षक भी दिए जा रहे हों लेकिन आप सांख्यिकी के वर्गीकरण को देखें तो आप मोटे तौर पर सांख्यिकी की शाखाओं को दो रूप में देख सकते हैं: एक तो आप डाटा का वर्णन कर रहे हैं जो सांख्यिकी का एक हिस्सा है और जो डाटा के विवरण और सारांश से संबंधित है, जिसे अधिक लोकप्रिय रूप से वर्णनात्मक सांख्यिकी या डिस्क्रिप्टिव स्टेटिस्टिक्स (descriptive statistics) शाखा के रूप में संदर्भित किया जाता है।**

**स्टेटिस्टिक्स या सांख्यिकी कि वह शाखा जो डाटा से निष्कर्ष निकालने से संबंधित होती है, उसे इन्फेरेंटीएल स्टेटिस्टिक्स (inferential statistics) कहा जाता है। इस शाखा का उपयोग आप डाटा से तर्क ढूंढने मे करते हैं। जब आप डाटा से तर्क ढूंढना या अनुमान लगाना चाहते हैं तो एक बहुत ही महत्वपूर्ण बात है और वो है, मौका की संभावना (possibility of chance)। क्योंकि जब आप डाटा से अनुमान लगा रहे होते हैं तो इस बात कि पूरी संभावना है कि आपके पास वह है ही नहीं है जो कि आप समझ रहे है कि आपके पास हैं।**

**तो इसलिए हम इस पाठ्यक्रम में प्रॉबबिलिटि विषय के परिचय (introduction to probability) को भी शामिल कर रहे हैं, ताकि आपको समझने में मदद मिले या आपको अगले पाठ्यक्रम के लिए तैयार होने में मदद मिल सके, जहां आप इन्फेरेंटीएल स्टेटिस्टिक्स (inferential statistics) के बारे में सीखेंगे।**

**(स्लाइड समय देखें: 04:11)**



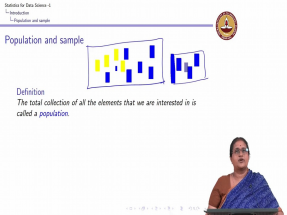
**जब हम मुख्य रूप से इन्फेरेंटीएल स्टेटिस्टिक्स (inferential statistics) के बारे में बात करते हैं तो हम डाटा से निष्कर्ष निकालने के बारे में बात करने की कोशिश करते हैं। यानि इन्फेरेंटीएल स्टेटिस्टिक्स (inferential statistics) को हम अनुमान की एक शाखा के रूप मे देखते हैं। यहाँ एक महत्वपूर्ण बात यह है कि जैसे कई बार आप भारत में उन सभी छात्रों के प्रतिशत के बारे में जानने में रुचि रखते हैं जिन्होंने अपनी बारहवीं कक्षा (12th) की परीक्षा उत्तीर्ण की है और इंजीनियरिंग का अध्ययन कर रहे हैं ; जैसे तमिलनाडु में सभी घरों की कीमतें; जैसे वर्ष 2019 में भारत में कारों की कुल बिक्री; किसी विशेष माह में शहर के मॉल में आने वाले लोगों का ऐज़ डिस्ट्रिबुशन यानि आयु वितरण।**

**तो इन सभी सवालों के जवाब पाने का एक तरीका यह है कि एक पूरी गणना कि जाए - आप जाएं और प्रत्येक व्यक्ति या आपकी रुचि की हर चीज पर डाटा एकत्र करें। उदाहरण के लिए, इस प्रश्न में आप भारत में सभी छात्रों के प्रतिशत के बारे में जानने में रुचि रखते हैं जिन्होने बारहवीं कि परीक्षा पास कि है और इंजीनियरिंग पढ़ रहे हैं, लेकिन बहुत जल्दी आप समझ जाएंगे कि इस प्रकार का डाटा प्राप्त करना बहुत आसान बात नहीं है।**

**अगर हम यह जानने में रुचि रखते हैं कि भारत में सभी छात्रों का प्रतिशत क्या है, तो इसके लिए अगर मैं एक डाटाबेस बनाना चाहती हूं और मैं उन सभी छात्रों का वास्तविक डाटा चाहती हूं जिन्होंने बारहवीं कक्षा पास की है तो और अगर मेरा इरादा सिर्फ यह जानना है कि आखिर किस तरह के लोग हैं जो अन्तः इंजीनियरिंग कर रहे हैं तो इसके लिए मैं जो एक मुख्य बात चाहूंगी वो यह कि भारत में सभी छात्रों के एक छोटे उपसमूह (subset) के साथ काम करना। भारत में सभी छात्रों का समूह वह है जिसे हम पॉपुलेशन के रूप में संदर्भित करते हैं। इसी पॉपुलेशन के एक छोटे सब सेट (subset) को सैंपल (sample) कहा जाता है। क्योंकि यह एक सब सेट (subset) है, इसलिए, मैं इसे एक सैंपल (sample) के रूप में रख रही हूँ।**

**इसी तरह अगर आप सभी घरों की कीमतों के बारे में जानना चाहते हैं तो आपको उन सभी घरों की कीमत के बारे में जानने की जरूरत नहीं है, जो किसी विशेष वर्ष में बेचे गए हैं; आप पूरी आबादी के एक छोटे उपसमूह के बारे में जानकारी ले सकते हैं। एक चीज जो आपको सैंपल (sample) एकत्र करने के बारे में ध्यान रखने की आवश्यकता है कि आपके द्वारा एकत्र किया गया सैंपल (sample) जितना संभव हो उतना पॉपुलेशन का प्रतिनिधित्व करे, आप कोशिश होनी चाहिए की सैंपल (sample) जितना संभव हो उतना प्रतिनिधि यानि रिप्रेजेन्टेटिव हो।**

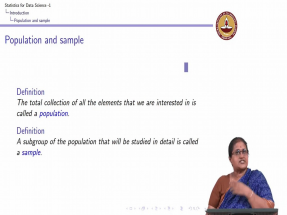
**(स्लाइड समय देखें: 06:42)**



**अब रिप्रेजेन्टेटिव सैंपल से हमारा क्या तात्पर्य है? उदाहरण के लिए, मैं पॉपुलेशन को परिभाषित करती हूं कि पॉपुलेशन उन सभी तत्वों का एक संग्रह है जिसमें हम रुचि रखते हैं। इसे समझने के लिए जैसे मैं यहाँ अलग-अलग रंग चुनती हूँ। मैं इस चुनाव के लिए किस टूल (tool) का उपयोग कर सकती हूं?.**

**तो मान लीजिए कि यह एक पॉपुलेशन (तत्वों का एक संग्रह) है और मैं यहां एक उपसमूह (subset) लेती हूं। मान लीजिए कि मैं एक सबसेट लेती हूं, यह एक सबसेट है, छोटा सेट यानि वास्तव में बड़े सेट का सबसेट, लेकिन हम तुरंत नोटिस करेंगे कि छोटे सेट में कोई पीला तत्व (रंग) नहीं है। इसलिए, मैं यह नहीं कह सकती कि यह छोटा सेट यानि सब सेट वास्तव में बड़े सेट का एक अच्छा प्रतिनिधित्व कर रहा है या अच्छा रिप्रेजेन्टेटिव है ।**

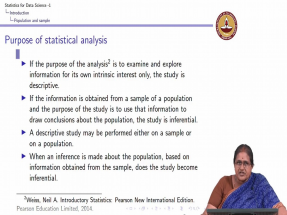
**(स्लाइड समय देखें: 07:39)**



**तो इसका तात्पर है कि सैंपल (sample) मूल रूप से पॉपुलेशन का एक उपसमूह (subgroup) है जिसका विस्तार से अध्ययन किया जाएगा।**

**यहाँ हमें पॉपुलेशन और सैंपल (sample) के बारे मे संक्षेप मे जितनी जानकारी की आवश्यकता थी हमने ले ली है। अब आगे आप जब अपना इन्फेरेंटीएल स्टेटिस्टिक्स (inferential statistics) का पाठ्यक्रम पढ़ेंगे तो आपको पॉपुलेशन और सैंपल (sample) की इस अवधारणा (concept) से अधिक विस्तार से परिचित कराया जाएगा। फिर भी हमें इस पाठ्यक्रम में पॉपुलेशन और सैंपल (sample) की अवधारणा की संक्षेप मे ही सही पर आवश्यकता क्यों है, अंततः जब भी हम समरी स्टेटिस्टिक्स (summary statistics) के बारे मे बात करते तो हमें यह समझने की आवश्यकता होती है कि समरी स्टेटिस्टिक्स (summary statistics) पॉपुलेशन के लिए है या सैंपल (sample) के लिए हैं। इस के बारे मे हम आने वाले समय में और अधिक जानेंगे।**

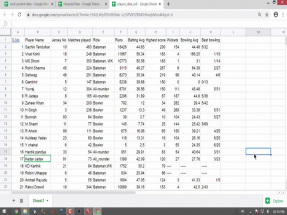
**(स्लाइड समय देखें: 08:15)**



**तो, सांख्यिकीय विश्लेषण यानि स्टेटिस्टिक्स एनालिसिस (statistics analysis) का उद्देश्य क्या है? आप डिसक्रिप्टिव स्टेटिस्टिक्स (descriptive statistics ) का प्रयोग कब करेंगे? और अनुमानात्मक स्टेटिस्टिक्स (inferential statistics) का प्रयोग कब करेंगे? यदि आपके विश्लेषण (analysis) का उद्देश्य केवल एकत्र कि गयी जानकारी के आंतरिक हित (intrinsic interest) के लिए जांच (examine) और अन्वेषण (explore) करना है, तो यह अध्ययन डिसक्रिप्टिव (descriptive) है।**

**इस बात से हमारा क्या तात्पर्य है? तो आइये मैं इसे एक डाटा सेट के माध्यम से आपके सामने प्रदर्शित करती हूं।**

**(स्लाइड समय देखें: 08:45)**



**यह फिर से एक और काल्पनिक डाटा सेट है जो सिर्फ क्रिकेट खिलाड़ियों के नाम दिखा रहा है। हम सभी इन क्रिकेट खिलाड़ियों- तेंदुलकर, कोहली, धोनी से अच्छी तरह से वाकिफ हैं। उन्होंने जो मैच खेले हैं, किस भूमिका में खेले हैं, कुल रन क्या हैं, बल्लेबाजी औसत क्या है, उच्चतम स्कोर क्या है, विकेट, गेंदबाजी औसत और सर्वश्रेष्ठ गेंदबाजी इत्यादि।**

**अब मान लीजिए कि मेरा उद्देश्य केवल यह समझना है कि बनाए गए कुल रन कितने हैं, बल्लेबाजी का औसत क्या है, सबसे अधिक बल्लेबाजी का औसत किसका है, किसने सबसे अधिक रन बनाए हैं, जिन्होंने सबसे अधिक मैच खेले हैं उनके बारे मे जानना, सबसे अधिक विकेट किसने लिए हैं- यदि ये सभी प्रश्न जो मैंने अभी-अभी उठाए हैं हमारी रुचि के हैं तो इनके उत्तर सीधे डाटा सेट से प्राप्त किए जा सकते हैं।**

**अगर मैं यहाँ किसी एक बल्लेबाज द्वारा बनाए गए रनों की संख्या को क्रमित करना चाहती हूँ या मैं यह भी जानना चाहती हूं कि किस बल्लेबाज ने कैसे रन बनाए हैं तो यह सब मैं केवल इस डाटा का वर्णन कर के कर सकती हूं। मुझे इस डाटा के बारे में और कुछ करने कि ज़रूरत नहीं है। तो इस मामले में मैं जो प्रश्न पूछ रही हूं वह मूल रूप से है। मेरे यहां जो उद्देश्य है वह यहां दी गई जानकारी की जांच और अन्वेषण करना है। इसलिए ये अध्ययन सिर्फ वर्णनात्मक यानि descriptive है। मैं यहाँ केवल इस डाटा सेट का वर्णन कर रही हूं जो यहां दिया गया है और यह अध्ययन वर्णनात्मक यानि डिसक्रिप्टिव कहलाएगा।**

**मान लीजिए अगर मैं इस डाटा सेट का उपयोग कर भी रही हूं तो एक बात जो हम इस डाटा सेट मे देखेंगे वो यह है की इस डाटा सेट मे सभी देशों के सभी क्रिकेटरों के बारे में संपूर्ण क्रिकेटिंग डाटा शामिल नहीं है।**

**यह डाटा पूरी पॉपुलेशन का सिर्फ सैंपल मात्र है। यह सिर्फ एक छोटा सा सैंपल है। मैं कह सकती हूं कि यह डाटा पिछले 5 या 10 वर्षों में भारतीय क्रिकेटिंग डाटा का सबसे अच्छा प्रतिनिधि सैंपल है या शायद यह डाटा पिछले एक दशक का हो सकता है। पर यह निश्चित रूप से एक सैंपल है, यह डाटा भारतीय क्रिकेट डाटा का एक सैंपल मात्र है।**

**यह डाटा पूरी पॉपुलेशन नहीं है जिसमें सभी बल्लेबाजों और सारे क्रिकेटरों को शामिल किया गया हो। लेकिन अगर मुझे इस डाटा को संक्षेप में प्रस्तुत करने में दिलचस्पी है और अगर मेरी अंतर्निहित (inherent) रुचि केवल इस डाटा को सारांशित करने के बारे में है तो मुझे केवल इस डाटा की वर्णनात्मक यानि डिसक्रिप्टिव स्टडि में दिलचस्पी होगी जिसके लिए हमारे पास वर्णनात्मक आँकड़े पर्याप्त हैं।**

**अब अगर मैं इस डाटा का उपयोग निष्कर्ष निकालने के लिए करूँ जैसे की उदाहरण के तौर पर यदि मैं यह जानना चाहती हूं कि कोई एक बल्लेबाज, बल्लेबाजी औसत के साथ- साथ क्या भूमिका निभाता है तो मुझे और अधिक जानकारी की आवश्यकता होगी और जिससे मैं भविष्य के लिए एक टीम चुन सकूँ। उदाहरण के लिए हम सभी आईपीएल (IPL) के लिए खिलाड़ियों की नीलामी और उनको कैसे चुना जाता है के बारे अच्छी तरह से जानते हैं। पर यहाँ अभी मैं इस डाटा का और अधिक वर्णन (description) करने में कोई दिलचस्पी नहीं रखती हूँ।**

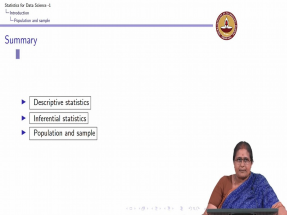
**मुझे ज़्यादा दिलचस्पी इस डाटा का उपयोग कर कुछ जानकारी इकट्ठा करने या कुछ अनुमान लगाने मे है जो की मैं आगे निर्णय लेने की प्रक्रिया में उपयोग करूंगी। इस प्रक्रिया मे हम एलिमंट ऑफ चान्स को शामिल करेंगे इससे हमे इस डाटा पर इन्फेरेंटीएल (inferential) स्टडि करने मे मदद मिलेगी।**

**वर्णनात्मक यानि डिसक्रिप्टिव स्टडि के दौरान अक्सर हमे यह समझने की आवश्यकता है कि क्या हमारे अध्ययन की प्रकृति केवल डिसक्रिप्टिव (descriptive) होगी या हम एक इन्फेरेंटीएल (inferential) स्टडि भी करना चाहते हैं।**

**जब हम (inferential) स्टडि की बात करते हैं, माफ कीजिएगा, जब हम डिसक्रिप्टिव स्टडि की बात करते हैं तो यह स्टडि या तो सैंपल (sample) पर या पॉपुलेशन पर की जा सकती है। चूँकि आगे आने वाली कक्षाओं में हम डिसक्रिप्टिव स्टडि के बारे में विस्तार से बात करेंगे तो हमें यहाँ यह बात समझ लेने की आवश्यकता है कि डिसक्रिप्टिव स्टडि सैंपल (sample) पर कि जा रही है या संपूर्ण पॉपुलेशन पर। इसी कारणवश हमने आपको इस स्तर पर सैंपल (sample) और पॉपुलेशन की धारणा से अवगत कराया।**

**हालाँकि, अपनी स्टडि मे हम अगर सैंपल (sample) के आधार पर पॉपुलेशन के बारे में अनुमान लगा रहे हैं तो यह अध्ययन इन्फेरेंटीएल होगा। इन्फेरेंटीएल स्टेटिस्टिक्स (inferential statistics) इस पाठ्यक्रम का दायरा नहीं है लेकिन फिर भी हम आपको संभाव्यता (probability) की अवधारणा से परिचित कराएंगे जो आपको इन्फेरेंटीएल स्टेटिस्टिक्स के लिए कार्यप्रणाली (methodology) विकसित करने में आगे मदद करेगी।**

**(स्लाइड समय देखें: 13:49)**



**यहाँ हम संक्षेप में कहे तो अब तक आपको ज्ञात हो जाना चाहिए कि स्टेटिस्टिक्ल एनालिसिस कि दो मुख्य शाखाएँ हैं डिस्क्रिप्टिव स्टेटिस्टिक्स और इन्फेरेंटीएल स्टेटिस्टिक्स। आप अपने अध्ययन के उद्देश्य के आधार पर या तो एक डिस्क्रिप्टिव स्टडि करते हैं या फिर एक इन्फेरेंटीएल स्टडि।**

**यदि आपका आंतरिक (intrinsic)उद्देश्य केवल आपके डाटा को संक्षेप में प्रस्तुत करना है तो आप एक डिस्क्रिप्टिव स्टडि के साथ जाएंगे। लेकिन यदि आपके अध्ययन का उद्देश्य भविष्य में अनुमान लगाना है या एक छोटे सबसेट का उपयोग करके एक बड़ी पॉपुलेशन के बारे में अनुमान लगाना है तो आप एक इन्फेरेंटीएल स्टडि के लिए जाएंगे। और इन्फेरेंटीएल स्टेटिस्टिक्स को समझने के लिए आपको पहले पॉपुलेशन और सैंपल (sample) की अवधारणा को समझने की जरूरत होगी।**

**GLOSSARY**

|  |  |  |
| --- | --- | --- |
| **ENGLISH WORD** | **WORD** | **MEANING** |
| **Descriptive Statistics** | **डिसक्रिप्टिव स्टेटिस्टिक्स** | **वर्णनात्मक आँकड़े** |
| **Inferential Statistics** | **अनुमानात्मक आंकड़ों** | **आनुमानिक आँकड़े** |
| **Cross Sectional** | **क्रॉस सेक्शनल** | **अनुप्रस्थ काट** |
| **Measurement Scale** | **मेश़रमेंट स्केल** | **नाप का पैमाना** |
| **Intrinsic** | **इंट्रिन्सिक** | **आंतरिक** |