MODULE 4 :

Thống kê cơ bản cần cho ML và ra quyết định dựa trên dữ liệu, nhận biết ước tính vs suy luận, thống kê tham số vs phi tham số, các phân phối phổ biến, thống kê tần số vs Bayes.

Ước tính (estimation): chỉ đưa ra giá trị tham số (ví dụ trung bình từ mẫu).

Trung bình = tổng giá trị / số quan sát.

Suy luận thống kê (statistical inference): muốn hiểu cả phân phối dân số, gồm ước tính trung bình + sai số chuẩn (standard error).

Sai số chuẩn đo lường khoảng cách trung bình từ mỗi giá trị đến trung bình.

ML vs suy luận thống kê: gần như giống nhau về cơ bản: dùng dữ liệu mẫu để suy ra phân phối dân số thực tế.

Một số mô hình ML chỉ quan tâm tới dự đoán, không cần biết hết thông số phân phối.

Tham số vs Phi tham số & các phân phối phổ biến là tìm ra quá trình tạo dữ liệu cơ bản.

Mô hình tham số: Có số lượng tham số hữu hạn, dựa trên giả định cụ thể về phân phối dữ liệu. Ước lượng tham số thường dùng Maximum Likelihood Estimation (MLE) → tìm giá trị tham số làm xác suất dữ liệu mẫu lớn nhất.

Mô hình phi tham số: Không dựa vào giả định phân phối cụ thể → “distribution-free”. Thường cần nhiều dữ liệu hơn.

Thống kê tần số (Frequentist): Dựa trên quan sát lặp lại hướng tới giới hạn khi số mẫu lớn. Cốt lõi: tự tin dựa vào kích thước mẫu và dữ liệu thực tế, không dùng thông tin ngoài.

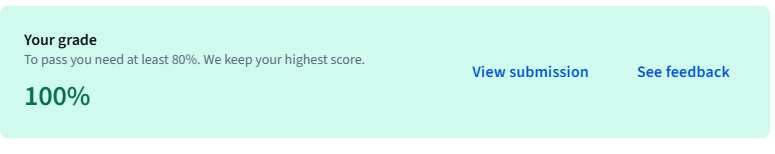
Thống kê Bayesian: Các tham số có phân phối xác suất không cố định. Cho phép kết hợp niềm tin trước (prior) và cập nhật khi có dữ liệu mới sẽ ra posterior distribution. Càng nhiều dữ liệu → phân phối tham số càng hẹp → dự đoán chắc chắn hơn.

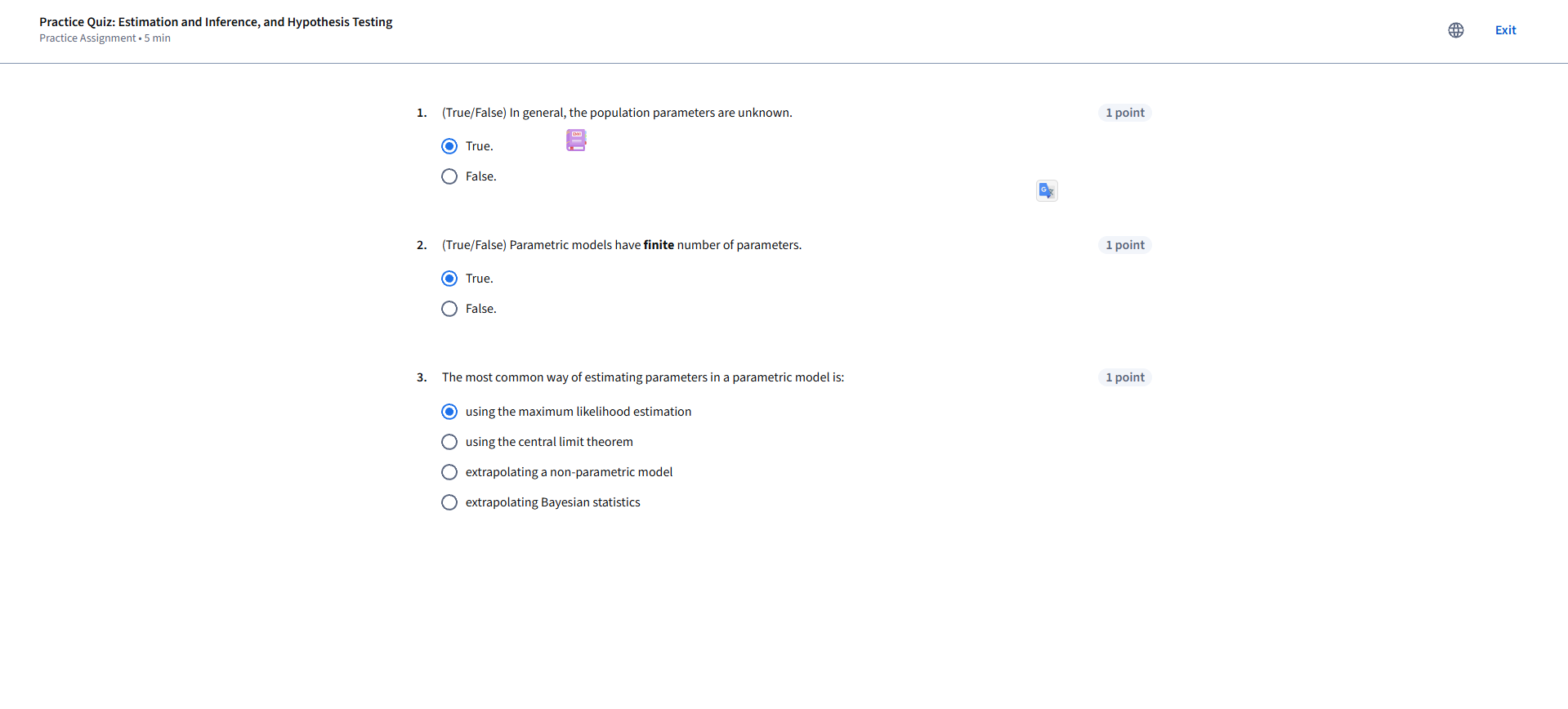
Chú ý Vẫn dùng các công thức tương tự như Frequentist, nhưng giải thích khác:

Frequentist xác suất mẫu bao phủ tham số dân số.

Bayesian xác suất phân bố tham số dựa trên dữ liệu + prior.

Quiz : Practice Quiz: Estimation and Inference, and Hypothesis Testing





Kiểm tra giả thuyết (Hypothesis Testing) là tuyên bố về tham số.

Có hai loại giả thuyết:

Null hypothesis (H0): giả thuyết vô hiệu, giá trị cụ thể.

Alternative hypothesis (H1): giả thuyết thay thế, giá trị khác hoặc lớn hơn/bé hơn H0.

Dễ hiểu là cái này phản bác cái kia

Quy trình kiểm tra giả thuyết (Frequentist):

Dùng dữ liệu mẫu quyết định từ chối H0 hay không

Thường nói: chỉ từ chối null, không bao giờ “chấp nhận H1” tuyệt đối.

Thực tế: nếu thống kê thử nghiệm đủ mạnh, H1 được coi là đúng.

Cách tiếp cận Bayesian: Không có ranh giới quyết định cố định. Tính xác suất posterior của H0 và H1 dựa trên dữ liệu.

Lỗi loại I & II trong kiểm tra giả thuyết Giả thuyết vô hiệu (H0): giả thuyết ban đầu, thường là “không có hiệu ứng” / giá trị cụ thể.

Giả thuyết thay thế (H1): giá trị khác H0, hiệu ứng thực sự tồn tại.

Ví dụ ném xu : H0 công bằng do 50 50 phần tram

H1 ko công bằng

Lỗi loại I (Type I): Từ chối H0 sai cách. Nghĩa là: H0 đúng nhưng dữ liệu mẫu khiến chúng ta quyết định từ chối.

Lỗi loại II (Type II):Chấp nhận H0 sai cách. Nghĩa là: H1 đúng nhưng dữ liệu mẫu khiến chúng ta không từ chối H0.

Power (Công suất) của bài kiểm tra: = 1 – typerII xác suất bác bỏ H0 khi H1 đúng. Tăng độ mạnh dễ từ chối H0 có thể tăng lỗi Type I. Làm khó bác bỏ H0 giảm lỗi Type I tăng lỗi Type II.

Thống kê thử nghiệm (Test Statistic): Giá trị tính toán từ dữ liệu mẫu, dùng để xác định xem có nên từ chối giả thuyết vô hiệu (H0**)** hay không.

Phân phối null (Null Distribution): Phân phối của thống kê thử nghiệm giả sử H0 đúng.

Giúp xác định vùng từ chối và vùng chấp nhận.

Vùng từ chối (Rejection Region), Vùng chấp nhận (Acceptance Region)

Mức ý nghĩa và giá trị P trong kiểm tra giả thuyết:

Mức ý nghĩa là ngưỡng xác suất mà tại đó chúng ta sẽ từ chối giả thuyết null (H0)

Chọn trước khi chạy thử nghiệm để tránh P-hacking.

Mức α phổ biến: 0.1, 0.05, 0.01, 0.001

Giá trị P (P-value) Nếu P ≤ α từ chối H0. Nếu P > α không từ chối H0.

F-Statistics, đa so sánh, và Bonferroni Correction

F-Statistics trong hồi quy tuyến tính F-statistic**:** đo lường mức độ tổng thể mà mô hình hồi quy cải thiện dự đoán so với chỉ dùng giá trị trung bình.

Quy tắc:

Nếu p-value của F-statistic nhỏ từ chối H0 ít nhất một β có ảnh hưởng.

Nếu p-value lớn không từ chối H0 các biến có thể không mang thông tin bổ sung.

Lỗi loại I khi thực hiện nhiều kiểm tra

Bonferroni Correction Mục đích**:** kiểm soát tổng xác suất lỗi loại I khi chạy nhiều test.

Mối tương quan vs. quan hệ nhân quả

Mối tương quan (correlation) tức là X và Y thay đổi cùng nhau không nhất thiết chỉ ra nguyên nhân và kết quả.

Quan hệ nhân quả (causation): Thay đổi X thực sự gây ra thay đổi Y.

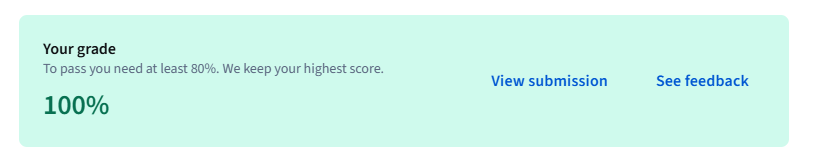
X và Y có thể tương quan vì nhiều lý do: X → Y (nguyên nhân trực tiếp), Y → X (nguyên nhân ngược) ,Biến gây nhiễu Z ảnh hưởng cả X và Y,Ngẫu nhiên / mối tương quan giả mạo( chả liên quan gì nhau )

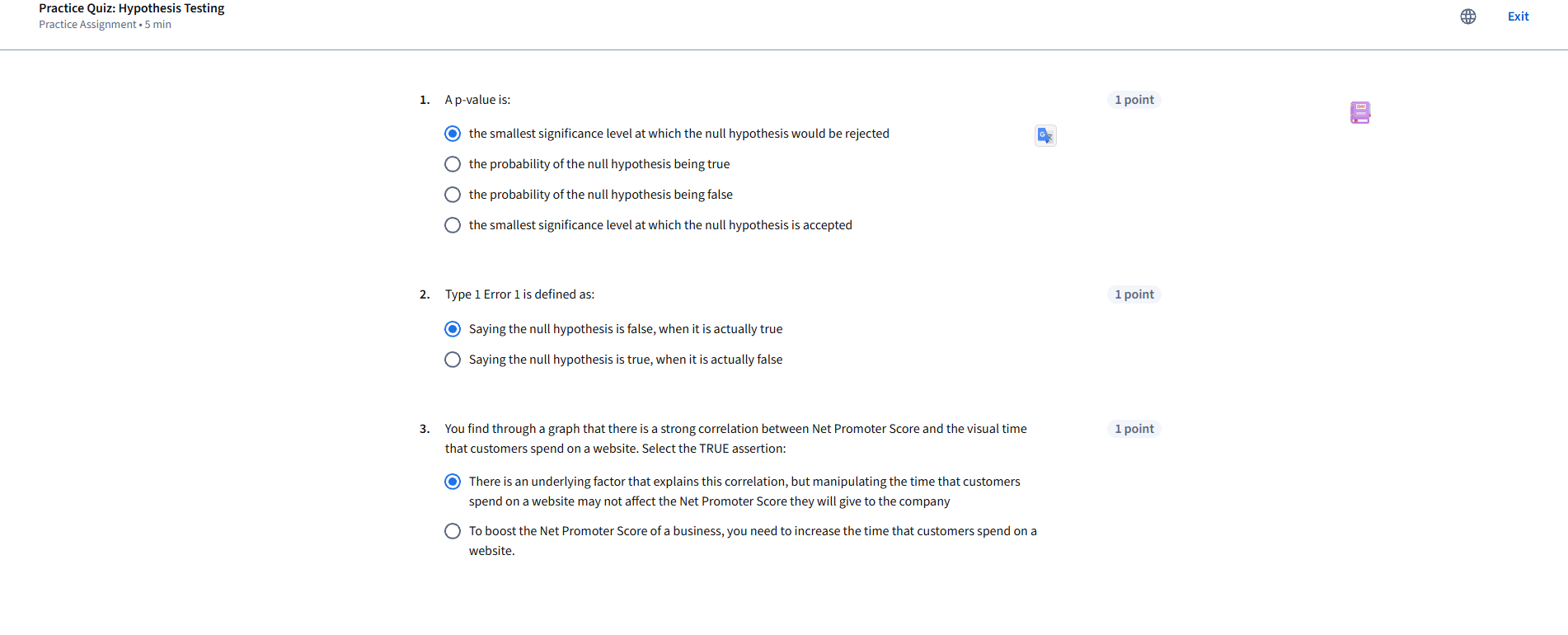
Biến gây nhiễu (Confounders) :Biến ngoài Z ảnh hưởng cả X và Y → tạo mối tương quan giả mạo.

Điều quan trọng**:** khi muốn xác định nhân quả, phải kiểm soát biến gây nhiễu.

Mối tương quan giả mạo (Spurious correlation) mối tương quan cao không có nghĩa là có cơ chế nhân quả hợp lý.

Quiz : Practice Quiz: Hypothesis Testing





Graded Quiz: Module 4 - Inferential Statistics and Hypothesis Testing

