Estadística en los tests de usabilidad. Menos miedo y más remangarse

Statistics in usability tests: Roll up your sleeves and have no fear

Mari-Carmen Marcos

Marcos, Mari-Carmen (2015). "Estadística en los tests de usabilidad. Menos miedo y más remangarse". *Anuario ThinkEPI*, v. 9, pp. 260-263.

http://dx.doi.org/10.3145/thinkepi.2015.61

Publicado en *IweTel* el 26 de septiembre de 2014

Resumen: Dentro del proceso del diseño de sitios web, la evaluación de su usabilidad es una fase clave para asegurar el éxito. De entre los métodos de evaluación, el test con usuarios es el que mayor información proporciona sobre posibles problemas de rendimiento, entendido como la eficacia y la eficiencia con la que los usuarios realizan las tareas en la Web. En este contexto de evaluación y de medición de rendimiento, la estadística es una herramienta que proporciona rigurosidad a los resultados de los tests, y permite obtener de ellos evidencia para tomar mejores decisiones de diseño.

Palabras clave: Estadística; Tests con usuarios; Experiencia de usuario; UX; Usabilidad; Métricas.

Abstract: Within a website design process, usability evaluation is a key step to ensure success. Among the various evaluation methods, user testing provides information about potential performance problems, such as the effectiveness and efficiency with which users perform tasks on the web. In this context of evaluation and performance measurement, statistics is a tool that provides validity and reliability to the tests' results, and delivers the evidence needed to make better design decisions.

Keywords: Statistics; User testing; User experience; UX; Usability; Metrics.

Para qué medir

Desde pequeños nos medimos con los adultos para comprobar cuánto hemos crecido. Durante la etapa escolar se miden nuestras aptitudes con exámenes y pruebas que muestran la evolución personal y nuestra posición dentro del grupo de la clase. En el entorno laboral se mide nuestra productividad y la consecución de nuestros objetivos. Las empresas miden sus productos con los de otras empresas de su competencia para conocer en qué posición se encuentran y encontrar puntos de mejora. Medimos para conocer, para gestionar y para controlar. Y sobre todo medimos con el fin de mejorar, porque sólo midiendo podremos tomar buenas decisiones.

Tanto en la creación de nuevos productos –sitio web, intranet, aplicación- como en su rediseño, el papel de la evaluación es crítico. Habitualmente se barajan varios diseños de interfaz, pero sólo uno será el elegido. Poner a prueba las opciones y testearlas con un grupo de usuarios ya es un gran avance. Es un hecho que muchos productos todavía se lanzan sin testear, y después se pagan las consecuencias de un diseño inadecuado y las facturas de los futuros rediseños.

Las formas en que se puede medir la calidad de un producto son varias. Si el producto se encuentra accesible a través de una pantalla, uno de los aspectos que a menudo se mide es su *usabilidad*, es decir, la capacidad que ese producto tiene para ser usado por su público objetivo de una forma eficiente, eficaz y satisfactoria (norma *ISO/IEC 9241*).

Desde la aparición de la informática se han realizado estudios de calidad de las interfaces de estos productos, pero fue a raíz de la aparición de internet y en particular de la *world wide web* cuando se dio a conocer de forma masiva entre los diseñadores de interfaces lo que se conoce

como diseño centrado en el usuario (*user-centered design*, UCD). El UCD pone en el centro del diseño a los usuarios, más allá de otras consideraciones como el diseño visual o las limitaciones tecnológicas, con la finalidad de crear interfaces que se adecuen a la forma en que las personas interactúan con la información.

"Medimos para conocer, para gestionar y para controlar. Y sobre todo medimos para mejorar, porque sólo midiendo podremos tomar buenas decisiones"

Si se ha conseguido convencer a los responsables del producto de la necesidad de testear las interfaces con un grupo de usuarios, es el momento de que entre en acción la estadística. Sí, esa palabra que seguramente preferimos evitar, que nos provoca pensamientos negativos del tipo "No tengo ni idea" o "Para qué complicarse la vida".

Nos equivocamos. Primero, al nivel que necesitaremos nosotros, no es difícil. De verdad. Y segundo: sí, sí hace falta aplicar estadística a los resultados porque sólo así podremos estar seguros de tomar buenas decisiones y podremos demostrarlo a quienes tienen la última palabra sobre el producto.

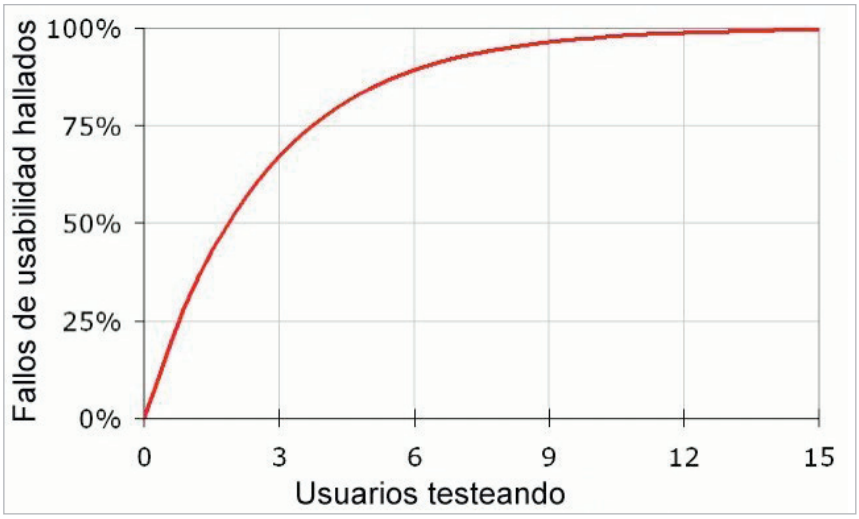
Ahora que ya vemos, o deberíamos ver, por qué es importante recurrir a la estadística, revisemos las ventajas de testear con usuarios.

Por qué testear con usuarios

Es cierto que para medir, y en particular para medir la usabilidad de una interfaz, existen distintas técnicas, y no todas incluyen la participación de usuarios. La evaluación heurística es una de las más conocidas dentro de las llamadas técnicas de inspección. Está basada en una lista de indicadores de calidad consensuados por los evaluadores, y su aplicación consiste en revisar una interfaz y anotar en qué grado cumple con esos indicadores. Como paso previo para una evaluación de usabilidad puede ser muy útil, pero en cambio tiene tres grandes limitaciones:

- la subjetividad de los evaluadores;
- la forma de cuantificar los errores detectados;
- la ausencia de un contexto de uso.

Ejemplo de métricas



Como en otras actividades, aquí se cumple la ley de Pareto: a partir de determinado número de usuarios el número de nuevos fallos descubiertos es ya despreciable. La simplificación de la ley conduce a la popular regla del 80/20: aproximadamente un 20% de usuarios permite descubrir el 80% de los fallos.

Los métodos que incorporan a usuarios son los de observación, donde los evaluadores son meros espectadores, y los actores son los usuarios del producto. Aunque pueden realizarse tests con usuarios en su contexto habitual, es más común llevarlos a cabo en un laboratorio donde los evaluadores tengan a su alcance los medios para tomar las medidas requeridas (grabación de la sesión, grabación de la pantalla del usuario, sala de observación para los *stakeholders* (todos

los implicados o interesados en el producto testeado) dispositivos de seguimiento de la mirada (*eyetracking*), software de seguimiento del ratón, etc.), y donde se tenga control de todas las variables que pueden interferir en la realización de las tareas (correcto funcionamiento de software y hardware, realización de las tareas sin interrupciones…).

En el contexto de la usabilidad y la UX (experiencia de usuario) se suelen diferenciar dos tipos de estudios con usuarios según su finalidad:

- los que tienen por objeto detectar problemas de usabilidad en un interfaz, y que se realizan con pocos participantes (10 personas suelen ser suficientes), sin aplicar análisis cuantitativo;
- los que comparan dos o más interfaces, por ejemplo dos sitios web que se dedican al mismo tipo de negocio, o dos versiones de una aplicación móvil, para saber cuál es mejor en función de las variables que quieran medirse.

"El grado de fiabilidad de los resultados así obtenidos es tan alto como el de dificultad para presentar los resultados a los responsables del producto"

Las variables que se miden de forma cuantitativa suelen corresponder al rendimiento objetivo (*performance*) de los usuarios:

- eficacia: nivel de éxito en la resolución de las tareas;
- eficiencia: facilidad para resolver las tareas, medida en tiempo, en número de clics, rendimiento, ahorro, etc.

Pero también se usan otras variables de tipo subjetivo como:

- comprensión y recuerdo del producto;
- satisfacción con el producto.

Para poder obtener datos fidedignos que nos hagan decantar por una u otra interfaz de las analizadas se requiere que participe un número mayor de usuarios que en los estudios de carácter cualitativo, ya que para poder saber si existen diferencias entre una interfaz y otra se aplicará un análisis estadístico, que por definición será cuantitativo.

Tabla 1. Tests de comparación de promedios entre dos grupos de datos

	Datos paramétricos	Datos no paramétricos
Medidas independientes	T-test de medidas independientes	*Test de Mann-Whitney*
Medidas repetidas	T-test de medidas repetidas	*Test de Wilcoxon*

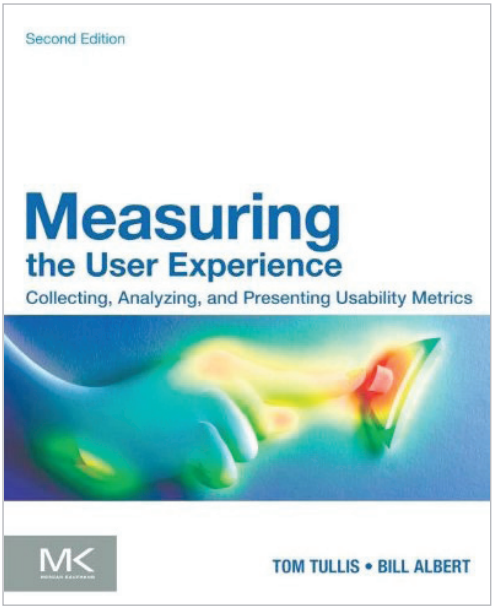
Estadística para medidas de usabilidad en 5 pasos

Sin ánimo de convertir esta nota en un manual, resumo en 5 pasos lo que a mí me ha servido para lanzarme a aplicar estadística a mis estudios con usuarios.

a) Planificar

La forma en que se planifica la sesión de test con usuarios condiciona que podamos posteriormente aplicar tests estadísticos "con sentido" y que den resultados fiables. Algunas claves:

- Muestra: planificar qué perfil de usuarios participará.
- Tareas: definir qué tareas realizarán, en qué orden y cómo se anotará el comportamiento de los usuarios mientras las realizan.
- Sopesar si conviene que todos los participantes usen las distintas variantes de la interfaz que se testea (diseño de medidas repetidas), o si es preferible dividir a los participantes y que cada subgrupo pruebe una de las alternativas de interfaz planteadas (diseño de medidas independientes). En función de la decisión que tomemos deberemos posteriormente aplicar unos u otros tests estadísticos para analizar los resultados.
- Definir cómo son las variables que se van a estudiar (las dependientes): nominales o numéricas. Para estudios comparativos de rendimiento en dos interfaces en el contexto de la usabilidad es conveniente que las variables objeto de estudio sean numéricas, ya que eso permite realizar tests que comparan promedios, por ejemplo comparar el tiempo que tardan los usuarios en realizar cierta tarea en dos versiones de una interfaz.



El libro de **Albert** y **Tullis** es como la biblia de las métricas de la usabilidad. Morgan Kaufmann, 320 pp. ISBN: 978 0 12 415781 1

b) Intuir: estadística descriptiva

La estadística descriptiva, como indica su nombre, describe los datos recogidos, por ejemplo el promedio de determinado dato y cuánto se desvían los datos de ese promedio (la desviación típica). Ojear estos primeros datos nos ayuda a entender cómo es la muestra de la que hemos tomado las medidas.

En esta fase es primordial identificar si los datos de la muestra son paramétricos, es decir, si siguen la esperada curva de normalidad (*campana de Gauss*) o no. De ello dependerá el tipo de test que se aplicará para comparar promedios. Estos tests nos dan la respuesta: *Shapiro-Wilk* y *Kolmogorov-Smirnof*. Están disponibles en todos los paquetes de análisis estadístico.

c) Aplicar

En estudios de rendimiento, lo más común es comparar dos muestras, normalmente para determinar si es mejor la interfaz x o la interfaz y. Para poder aplicar un test estadístico de comparación de promedios necesitaremos que la variable objeto de estudio sea numérica: número de participantes que ha logrado terminar con éxito una tarea, tiempo que han tardado, número de clics que han hecho. En los estudios de recuerdo, de comprensión y de satisfacción, que son más cualitativos, se pueden transformar las respuestas en números para darles un tratamiento estadístico con técnicas de comparación de promedios.

El test que deberá aplicarse dependerá de dos factores:

- si se ha hecho un diseño de medidas independientes o repetidas;
- si los datos son paramétricos o no.

Conocer estos 4 tests (tabla 1) nos puede solucionar la mayoría de los casos.

d) Interpretar

Precisamente si se ha aplicado alguno de los tests anteriores es porque se quiere estar seguro de si hay una interfaz mejor que la otra en cuanto a alguna de las métricas analizadas. Para eso contamos con la significancia estadística, que nos dice si las diferencias encontradas en los datos de una y otra interfaz son suficientes para poder afirmar que hay una diferencia real, es decir, que no se debe al azar que puede introducir haber trabajado con unos participantes u otros. Este dato nos lo da el llamado *p-valor*, que de forma consensuada por la ciencia se estima que debe ser menor de 0,05 para poder afirmar que efectivamente existen diferencias estadísticamente significativas. En ocasiones los promedios habrán indicado que hay diferencias entre el rendimiento de los usuarios en una y otra interfaz, pero el *p-valor* nos dirá si esa diferencia es tal o era un espejismo, por decirlo de alguna forma.

e) Presentar

Una vez hecho todo esto, ya sabemos qué opción de interfaz generará mejores resultados en cuanto a las métricas analizadas. Pero atención, el grado de fiabilidad de estos resultados es tan alto como el de dificultad para presentar los resultados a los responsables del producto. Por ello debemos equilibrar la rigurosidad de la información que presentamos con la facilidad para que sea entendida por quienes no dominan los términos estadísticos. Usar tablas que marquen datos clave y gráficos que ayuden a visualizar los resultados ayudará. Si no somos capaces de presentar nuestro estudio de una forma comprensible y convincente, de poco habrá servido el esfuerzo.

Mari-Carmen Marcos
Departamento de Comunicación
Universitat Pompeu Fabra, Barcelona
mcarmen.marcos@upf.edu