ENMES 155 – Estadística I

Presentación para la Cátedra Sincrónica 1

Semana del 8 al 12 de marzo, 2021

Eduardo Engel

Primera versión: Marzo 5, 2021.

Esta versión: Marzo 11, 2021.

Los conceptos de la semana

Problema de Nivel 2

Ventaja del local y pandemia

Los conceptos de la semana

Problema de Nivel 2

Ventaja del local y pandemia

CONCEPTOS CLAVE VIDEO 1

Probabilidades

Estadística

Se combinan para proveer metodologías para:

- Decidir si regularidades observadas son reales o producto del azar.
- Decidir si asociaciones observadas entre dos variables son causales.

De asociación a causalidad

Para establecer causalidad:

- Ideal: EAC.

Experimento aleatorizado controlado (EAC):

- Tratamiento
- Grupo tratado y grupo control
- Asignación de tratados y controles al azar

¿Qué hacer cuando no se cuenta con un EAC o no es posible realizar uno?

El experimento de John Snow:

- ¿Fue un EAC?
- ¿Cómo estableció causalidad?
- ¿Qué hubiese pasado si John Snow analiza la evidencia a nivel de comunas de Londres?

Conceptos claves Video 2

Tipos de datos: corte transversal, serie de tiempo, panel/longitudinal.

Tipos de variables: categórica (nominal, ordinal, dummy) y cuantitativas (discreta, continua).

Medidas de localización: media, mediana, moda y media podada.

Medidas de dispersión: varianza, desviación estándar, rango y rango intercuartil.

Robustez: una medida de localización (o dispersión) es robusta si el error que introduce una observación aberrante (outlier) es acotado.

Visualización de los datos.

Medidas de localización y dispersión: Lecciones

- No existe una medida de localización que sirva en todas las situaciones, se recomienda examinar los datos y varias medidas y comparar. Lo mismo vale para medidas de dispersión.
- Uno de los criterios a tener en cuenta al momento de elegir medidas de localización es cuán robustas son. Pero no es el único, el promedio no es robusto pero, al menos intuitivamente, usa mucho más la información disponible que la mediana.
- Las medias podadas proveen un compromiso entre uso de la información y robustez.

Los conceptos de la semana

Problema de Nivel 2

Ventaja del local y pandemia

Los conceptos de la semana

Problema de Nivel 2

Ventaja del local y pandemia

INFLUENCIA MÁXIMA

Los valores observados de 12 cartas de un mazo, del cual se excluyen los ases, J,Q y K, ordenados de menor a mayor, son:

2, 2, 2, 3, 3, 4, 5, 5, 8, 8, 9, 10.

Como los valores que pueden tomar las observaciones de la muestra pertenecen al conjunto $\{2,3,4,5,6,7,8,9,10\}$, no podemos usar el concepto de robustez visto en clases (si una outlier puede llevar a un error arbitrariamente grande o no) para capturar cuán robusto es una medida de localización que describe la localización de la muestra.

Por eso, en este problema exploramos un nuevo concepto de robustez para una medida de localización que sirve cuando la variable de interés toma un número pequeño de valores. Este concepto, que llamaremos **influencia máxima**, se define como el mayor cambio en la medida de localización, en valor absoluto, si el valor de una observación de la muestra cambia a otro valor posible.

Para la muestra anterior, calcule la influencia máxima para:

- La media.
- La mediana.
- La media podada con $k = 1$.

Además de la media, mediana y media podada (con $k = 1$) consideramos la moda. Como vimos en el Video 2, la moda se define como aquel valor que aparece más veces en la muestra. Por ejemplo, en los datos del comienzo de este problema, la moda es 2.

- ¿La moda le parece una medida de localización robusta? Justifique, intuitivamente, su respuesta.
- Para la muestra anterior, calcule la influencia máxima de la moda.
- Explique por qué el concepto de robustez visto en clases no sirve en este caso. Algo se insinuó en el enunciado, vaya más allá de eso.

Respuestas lámina anterior

Para la muestra anterior, calcule la influencia máxima para:

- La media. **Resp**: $2/3$. Se obtiene pasando un 2 a 10 o viceversa. Si se pasa un 2 a 10, la media pasa de 6,1 a 6,9.
- La mediana. **Resp**: 1. Se obtiene pasando un 8, 9 o 10 a un 2. La mediana pasa de 4,5 a 3,5.
- La media podada con $k = 1$. **Resp**: $4/5$. Se obtiene pasando un 2 a un 10. La media podada pasa de 4,9 a 5,7.
- La moda. **Resp**: 6. Se obtiene pasando un 2 a un 8.

En este ejemplo, la moda es la menos robusta de las medidas de localización. Intuitivamente, el motivo es que la moda puede "saltar mucho" cuando cambia una observación, hay ejemplos donde puede pasar del menor al mayor valor posible. En cambio, las otras medidas de localización no saltan tanto.

Los conceptos de la semana

Problema de Nivel 2

Ventaja del local y pandemia

Los conceptos de la semana

Problema de Nivel 2

Ventaja del local y pandemia

Ventaja del local en deportes: Antecedentes

Existe bastante evidencia sobre la ventaja de ser local en los deportes.

Un ejemplo que no requiere datos:

- Al dueño de casa le suele ir mejor en el mundial en que es local que en otros mundiales: Chile y el mundial de futbol de 1962

La lámina siguiente muestra evidencia sistemática: 6 campeonatos para casi 40 países.

EVIDENCIA: VENTAJA DE LOS LOCALES: 6 CAMPEONATOS ANTERIORES A 12/2002

Country	Played	Won	Drawn	Lost	H
Albania	786	328	135	103	
Bosnia-Herzegovina	1156	760	215	181	
Bulgaria	1324	816	228	280	
Serbia & Montenegro	1032	518	204	210	
Romania	1704	1033	314	357	
Macedonia FYR	834	499	146	189	
Croatia	876	489	198	189	
Czech Republic	1440	739	398	303	
Ukraine	1324	710	303	311	
Slovakia	1320	701	299	320	
Greece	1646	889	338	419	
Portugal	1530	774	404	352	
France	1910	930	551	429	
Georgia	984	529	198	257	
Poland	1332	674	336	322	
Italy	1836	882	541	413	
Azerbaijan	482	295	79	138	
Spain	2362	1150	638	584	
Slovenia	1152	584	274	294	
Switzerland	792	373	237	182	
Germany	1836	909	465	462	
Austria	900	446	213	241	
Turkey	1836	913	428	495	
Russia	1506	720	401	385	
Hungary	1322	645	328	349	
Netherlands	1836	897	438	501	
Belgium	1836	884	442	510	
England	2280	1031	623	606	
Belarus	1294	625	260	409	
Faroe Islands	360	178	63	119	
Iceland	540	240	142	158	
Israel	1116	510	264	342	
Sweden	1092	479	296	317	
Norway	1092	512	224	356	
Cyprus	1092	527	194	371	
Finland	1188	622	238	429	
Moldova	826	373	191	262	
Ireland	920	394	261	275	
Scotland	1176	517	288	371	
Wales	1190	550	240	400	
Denmark	1188	509	315	364	
Malta	270	124	50	96	
Armenia	520	247	79	194	
Northern Ireland	1012	413	278	321	
Lithuania	946	404	216	326	
Luxembourg	396	165	95	136	
Estonia	560	243	115	202	
Latvia	360	248	91	221	

¿Cómo medimos la ventaja del local en un campeonato?

Queremos un **indicador** que capture la ventaja del local en un conjunto de partidos.

Algunas opciones:

- Porcentaje de partidos (donde hubo ganador) que ganaron los locales.
- Diferencia promedio de goles entre locales y visitas.
- Fracción de puntos adjudicados que ganaron los locales.

Criterios para comparar:

- Robustez.
- Uso de información.

Los autores del trabajo que consideramos antes usan la tercera opción.

DATOS DE LOS 6 CAMPEONATOS ANTERIORES A DICIEMBRE DE 2002

Country	Played	Won	Drawn	Lost	Home advantage
Albania	786	528	155	103	78.9%
Bosnia-Herzegovina	1156	760	215	181	76.7%
Bulgaria	1324	816	228	280	71.5%
Serbia & Montenegro	1032	618	204	210	71.2%
Romania	1704	1033	314	357	71.1%
Macedonia FYR	834	499	146	189	69.7%
Croatia	876	489	198	189	68.5%
Czech Republic	1440	739	398	303	66.7%
Ukraine	1324	710	303	311	66.3%
Slovakia	1320	701	299	320	65.6%
Greece	1646	889	338	419	65.3%
Portugal	1530	774	404	352	65.1%
France	1910	950	531	429	65.0%
Georgia	984	529	198	257	64.8%
Poland	1332	674	336	322	64.4%
Italy	1836	882	541	413	64.2%
Azerbaijan	482	265	79	138	63.9%
Spain	2362	1160	638	564	63.9%
Slovenia	1152	584	274	294	63.7%
Switzerland	792	373	237	182	63.4%
Germany	1836	909	465	462	63.3%
Austria	900	446	213	241	62.4%
Turkey	1836	913	428	495	62.3%
Russia	1506	720	401	385	62.2%
Hungary	1322	645	328	349	62.2%
Netherlands	1836	897	438	501	61.7%
Belgium	1836	884	442	510	61.1%
England	2280	1051	623	606	60.7%
Belarus	1294	625	260	409	58.9%
Faroe Islands	360	178	63	119	58.7%
Iceland	540	240	142	158	58.3%
Israel	1116	510	264	342	58.2%
Sweden	1092	479	296	317	58.2%
Norway	1092	512	224	356	57.7%
Cyprus	1092	527	194	371	57.6%
Finland	1389	622	338	429	57.6%
Moldova	826	373	191	262	57.3%
Ireland	930	394	261	275	57.0%
Scotland	1176	517	288	371	56.8%
Wales	1190	550	240	400	56.8%
Denmark	1188	509	315	364	56.7%
Malta	270	124	50	96	55.5%
Armenia	520	247	79	194	55.4%
Northern Ireland	1012	413	278	321	55.0%
Lithuania	946	404	216	326	54.5%
Luxembourg	396	165	95	136	54.0%
Estonia	560	243	115	202	53.9%
Latvia	560	248	91	221	52.5%

Posibles razones

- El aliento del público.
- La presión del público sobre el equipo visitante.
- Fatiga de viajar.
- Sesgos del árbitro.
- Conocer mejor la cancha.

Ilustraciones anecdóticas:

- El caso de los camarines del equipo visitante en la Bombonera en Buenos Aires.
- Copa Libertadores y árbitros y guardalineas de tres países distintos para dificultar el pago de sobornos.

¿Cuánto importa cada uno de los factores anteriores?

Un experimento natural

La pandemia ofrece una muy buena oportunidad para determinar la relevancia de algunos de los factores que podrían explicar la ventaja del local, pues los partidos se han jugado sin público, de modo que los primeros tres factores no han estado presentes.

Consideramos el fútbol.

Se dice que la pandemia provee un **experimento natural** para conocer mejor los determinantes de la ventaja del local:

- Tratados: partidos sin público.
- Controles: partidos con público.
- Resultado del experimento: ganador, número de puntos, número de goles, número de tarjetas rojas (o amarillas), etc.

Posibles controles:

- Partidos del mismo campeonato pre-pandemia
- Partidos de campeonatos anteriores
- Ventajas y desventajas de las dos opciones

Causalidad:

- Discutir el VAR como eventual factor de confusión

Un estudio concreto

Un estudio reciente analiza los resultados del campeonato de primera división del fútbol alemán para la temporada 2019-2020, la cual se inició con partidos con público (223 partidos) pero, producto de la pandemia, terminó con partidos sin público (83 partidos). También incluye las estadísticas de la temporada 2018-2019.

La tabla que sigue muestra el número y la fracción de partidos que ganaron, empataron y perdieron los locales bajo las dos modalidades.

Modalidad	Ganador (número)			Ganador (porcentaje)			Total puntos		
	Local	Empate	Visita	Local	Empate	Visita	Local	Visita	Fracción local
Con público	96	49	78	43.0%	22.0%	35.0%	337	283	54,4%
Sin público	27	20	36	32.5%	24.1%	43.4%	101	128	44,1%
2018-2019	138	73	95	45.1%	23.9%	31.0%	487	358	57,6%

Comentarios

- Los puntos obtenidos por los locales pasaron de 57,6% (o 54,4%), dependiendo de los controles que se usen, a 44,1%.
- Las diferencias observadas (la ventaja de local en partidos con público y la ventaja de la visita en partidos sin público), ¿son reales o podrían ser producto del azar?
- También podría ser que una es real y la otra producto del azar. De hecho, nuestros prejuicios se verían confirmados si solo la primera es real.

Esta es una de las tareas fundamental de la estadística, que veremos en detalle a lo largo del curso:

- Determinar cuándo los hallazgos que sugieren los datos son **reales** y cuándo podrían ser producto del **azar**.

Volveremos sobre esta aplicación en cátedras futuras.

Temas que se cubrieron en esta aplicación

Temas de la semana:

- Experimento natural: tratados, controles, factores de confusión
- ¿Qué se necesita para pasar de asociación a causalidad?
- Uso de estadisticas descriptivas: media, mediana, moda.
- Regularidades observadas: ¿Reales o producto del azar?

Temas adicionales:

- Cómo pasar de una pregunta a una estrategia empírica para responder la pregunta
- Cómo elegir un indicador apropiado para medir el efecto de interés.

ENMES 155 – Estadística I

Presentación para la Cátedra Sincrónica 1

Semana del 8 al 12 de marzo, 2021

Eduardo Engel

Primera versión: Marzo 5, 2021.

Esta versión: Marzo 11, 2021.