INF-280: Estadística Computacional
Ayudantía No. 1
Análisis de Datos

Ignacio Cea Diego Quezada

1. Existe una discusión respecto a qué medida de tendencia central es más representativa a la hora de analizar los ingresos de la población de un país. Suponga que se tiene la siguiente muestra de ingresos mensuales (en miles de pesos chilenos) ordenada de menor a mayor:

300	320	320	335	350	350	400	400	410	420	
420	450	500	600	620	650	680	750	850	1000	1500

   Se sabe que su media y su mediana son, respectivamente, 553.6 y 420.

   (a) Si se agrega el dato 3000 a la muestra, ¿cómo varía la media y la mediana?

   (b) ¿A qué se debe la notable diferencia entre la variación de la media y la variación de la mediana al agregar el dato 3000?

   (c) ¿Qué puede concluir respecto al uso de la media versus la mediana como un valor representativo de los ingresos de una población?

2. El agarre se aplica para producir fuerzas superficiales normales que comprimen el objeto que se quiere aferrar. Los ejemplos incluyen a dos personas dándose la mano, o una enfermera apretando el antebrazo del paciente para detener el sangrado. Se incluyen los siguientes datos sobre la fuerza de prensión en Newton para una muestra de 42 individuos:

16	18	18	26	33	41	54	56	66	68	87	91	95
98	106	109	111	118	127	127	135	145	147	149	151	168
172	183	189	190	200	210	220	229	230	233	238	244	259
294	329	403										

   (a) Calcule el rango y el rango intercuartílico ($\boldsymbol{Q}_{\alpha=0.5}$). ¿Qué indican cada uno de estos estadísticos?

   (b) ¿Qué tan grande o pequeña tiene que ser una observación para calificar como valor atípico/valor extremo? ¿Hay valores atípicos?

   (c) ¿Cuánto podría disminuir la observación 403, actualmente la más grande, sin afectar $\boldsymbol{Q}_{0.5}$?.

   (d) ¿Qué puede decir respecto a las medidas de forma (asimetría y curtosis)?

3. Una compañía quiere analizar entre sus integrantes el aumento del tiempo dedicado al trabajo en el actual escenario de teletrabajo, medido como porcentaje (Y), junto con los años de antigüedad (X) que lleva en la empresa. La siguiente tabla muestra los datos de aquellos colaboradores:

Y	95	90	95	95	80	80	85	90	80	85	95	95
X	1	3	3	3	1	2	3	3	3	1	2	3

Y	90	85	95	80	80	85	90	90	85	85	90	80
X	3	3	1	1	3	1	2	1	2	3	1	3

y la respectiva tabla de contingencia **incompleta**:

X/Y	80	85	90	95	
1	2	2	2		8
2	1			1	
3	3		3		12
		6	6		24

(a) Complete la tabla de contingencia.

(b) ¿En cuánto varían los años de antigüedad cuando los colaboradores han aumentado un 95 % el tiempo dedicado al trabajo?.

(c) Estudie la variabilidad del aumento de tiempo dedicado a trabajar, estratificado por los años de antigüedad de los colaboradores, es decir, obtenga la varianza total. ¿Qué puede comentar respecto a la heterogeneidad del comportamiento evidenciado por los colaboradores?.

(d) ¿Las variables X e Y son estadísticamente dependientes?. De ser así, y basado en un modelo lineal, ¿qué porcentaje de la variabilidad de una de las variables puede ser explicado por la otra?.

4. **(Bonus track)** Un sensor de temperatura envía cada segundo una nueva medición de temperatura t_i $[K]$. Interesa almacenar en cada instante el promedio de temperaturas $\bar{t}_n$ por segundo considerando los n datos enviados por el sensor.

   Lamentablemente no se tiene acceso a todas las mediciones t_i pues sería muy costoso su almacenamiento. En cada instante solo se conoce el número de n datos enviados, el nuevo dato t_n y la media histórica $\bar{t}_{n-1}$ considerando los primeros $n-1$ datos.

   (a) Defina la función $\bar{t}_n = u(\bar{t}_{n-1}, t_n, n)$ encargada de actualizar la media cada segundo usando el nuevo dato t_n.

   (b) En el instante τ se identifica que el sensor de manera sistemática agregaba 30 Kelvin a cada lectura de temperatura debido a un desajuste. ¿Es posible recuperar la media correcta? Explique.

   (c) Las mediciones del sensor no están exentas de errores, además, sabemos que la media es un estadístico muy sensible a outliers. Modifique la función definida en (a) para que las mediciones históricas tengan un mayor peso en el cálculo de la actualización de la media, de forma que, nuevas mediciones no tengan el peso necesario para desestabilizar el valor de la media.