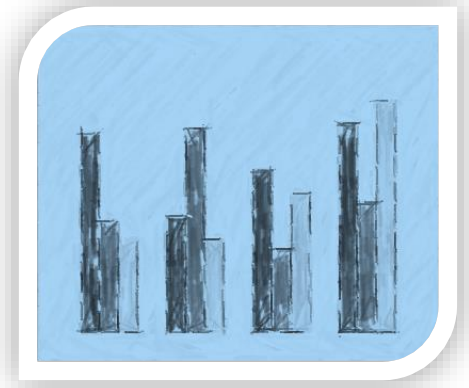
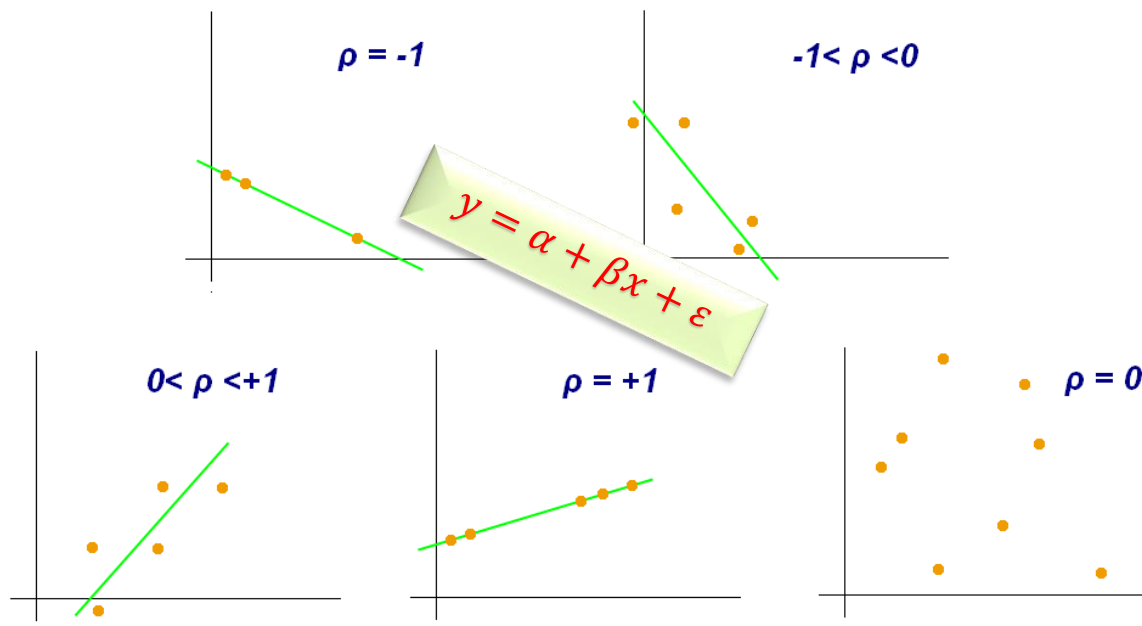


Pobreza y Educación en Puerto Rico



Análisis de Datos y Probabilidad
Grado 10-12

Repositorio Virtual para la Enseñanza de Estadística y Probabilidad en Escuela Superior (RepASA)



Introducción

En esta actividad los estudiantes podrán estudiar la relación entre dos variables y cuantificar dicha relación usando el coeficiente de correlación lineal de Pearson. También, ajustarán un modelo lineal usando el método de mínimos cuadrados, y aprenderán a identificar e interpretar los coeficientes del modelo. El uso de tecnología (Excel) para graficar y ajustar el modelo será visto en esta actividad.

Objetivos de Aprendizaje

Luego de finalizar la actividad, el estudiante será capaz de:

1. C-determinar la correlación entre dos variables numéricas.
2. P-interpretar la pendiente y el intercepto de un modelo lineal simple.
3. A-apreciar el valor de la estadística en las investigaciones.

Estándares y Expectativas (PR Common Core)

Estándar de Contenido: **Análisis de Datos y Probabilidades**

ES.E.44.1 Interpreta la inclinación (razón de cambio) y el punto de corte (término constante) de un modelo lineal según el contexto de los datos.

ES.E.44.2 Distingue entre correlación y causalidad. Determina la correlación entre dos variables numéricas con o sin tecnología.

Materiales

- Una computadora con a Excel y acceso a internet (opcional).
- Plantilla de la actividad.
- Calculadora.

Tiempo Estimado: 50 minutos.

Repositorio Virtual para la Enseñanza de Estadística y Probabilidad en Escuela Superior (RepASA)

Este material se distribuye gratuitamente para uso en los salones de clase. Su venta está prohibida. Su desarrollo fue posible gracias al apoyo de la *American Statistical Association (ASA)*, Capítulo de Puerto Rico de la ASA y el proyecto AFAMaC-Matemáticas de la Universidad de Puerto Rico – Mayagüez.

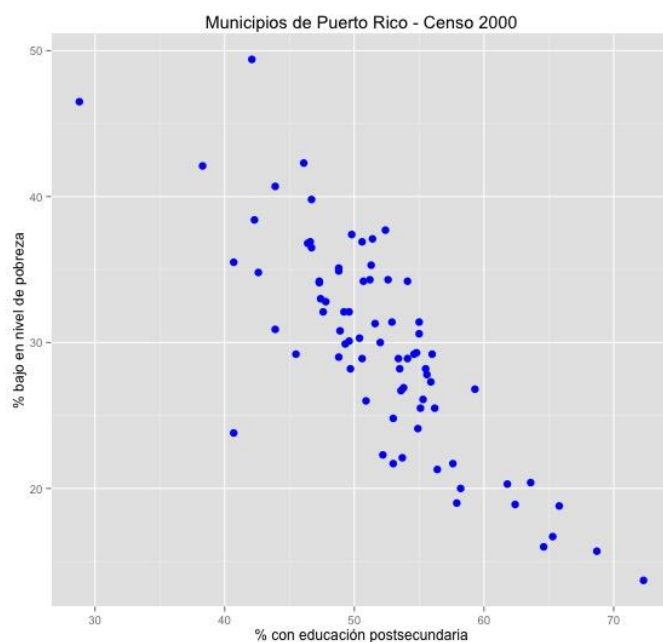
Disponible en la página web: <http://pegasus.uprm.edu/~pedro.torres/RepASA>

Actividad

El gran líder sudafricano, Nelson Mandela, dijo: **“La educación es el arma más poderosa para cambiar el mundo”**. Es bien conocido que la educación es una buena forma de combatir la pobreza. Ha habido conferencias y foros a nivel mundial sobre el papel de la educación en la reducción de la pobreza. En Puerto Rico se estima que cerca del 42% de las personas viven bajo el nivel de pobreza a pesar de que los niveles de alfabetismo son cercanos al 92%¹, según datos recientes. En esta actividad vamos a estudiar la asociación entre la pobreza y el nivel educativo en PR usando los datos publicados por la Junta de Planificación de PR². La base de datos de esta actividad contiene el porcentaje de personas mayores de 16 años bajo el nivel de pobreza y el porcentaje de personas con grado postsecundario en los 78 municipios de la isla, según el censo del 2000. Vamos a determinar si los pueblos con personas más educadas tienden a tener menores niveles de pobreza. Se espera que un pueblo más educado tenga oportunidades de acceder a mejores empleos, y sobre todo, tenga ciudadanos más preparados para la toma de decisiones.

Discusión

1. La siguiente gráfica muestra los datos del porcentaje de la población civil mayor de 16 años bajo nivel de pobreza versus el porcentaje de la población civil mayor de 16 años con educación postsecundaria para cada uno de los pueblos de Puerto Rico, según el censo del 2000.



“Nunca consideres el estudio como una obligación, sino como una oportunidad para penetrar en el bello y maravilloso mundo del saber”.

Albert Einstein

“Erradicar la pobreza no es acto de caridad, es un acto de justicia”.

Nelson Mandela

- a) ¿Cuál es el % de pobreza esperado para un pueblo de PR con un 50% de sus habitantes con educación postsecundaria?

De acuerdo con el diagrama de dispersión, un pueblo con un 50% de personas con educación postsecundaria se espera que tenga alrededor de un 30% de personas bajo el nivel de pobreza. Aquí es importante que enfatice en la palabra “espera” ya que, como se puede ver en el diagrama, hay pueblos con niveles de educación de 50% y sus niveles de pobreza podrían variar de 27% a 37%, aproximadamente. Es decir, dado un 50% de nivel educativo, se pueden encontrar varios pueblos con diferentes niveles de pobreza pero en promedio tienen un nivel de pobreza de 30%.

- b) Describa la relación que existe entre las variables.

El diagrama de dispersión indica que a medida que el % de personas con educación postsecundaria aumenta el % de nivel de pobreza tiende a disminuir. Es muy común que la frase anterior se sustituya por “a medida que el % de personas con educación postsecundaria aumenta el % de nivel de pobreza disminuye”. Note que esta última frase implica que si el nivel educativo aumenta entonces la pobreza tiene que disminuir. Esto es cierto para un modelo matemático más no para los datos como se visualizan en el diagrama de dispersión. Por ejemplo, note que hay un pueblo con un nivel educativo igual a 50% y un nivel de pobreza de 30%. Sin embargo, tal como indica el siguiente gráfico, hay otros pueblos con un % más alto de educación postsecundaria (~55%) pero sus niveles de pobreza están alrededor de 35%. Se observa que aumentó el nivel de educación y la pobreza aumentó comparado al pueblo con un nivel educativo de 50%. Es decir, un aumento en el nivel educativo no necesariamente implica una disminución en la pobreza. Justamente por este resultado es importante insistir en el concepto de “en promedio”.



Correlación: El grado con el cual dos o más variables cuantitativas muestran una tendencia a variar conjuntamente. Dos variables cuantitativas pueden tener correlación positiva si tienden a aumentar o disminuir conjuntamente, correlación negativa si una de ellas aumenta y la otra tiende a disminuir, o no estar correlacionadas.

Coefficiente de correlación (de Pearson): Cuantifica cuán fuerte es la relación lineal entre dos variables cuantitativas. El coeficiente de correlación varía entre -1 y 1. Muchas veces se denota por la letra r o la letra griega ρ (rho).

$r = 1$: Relación lineal exacta y positiva.

$0 < r < 1$: Correlación positiva.

$r = 0$: No existe relación lineal.

$-1 < r < 0$: Correlación negativa.

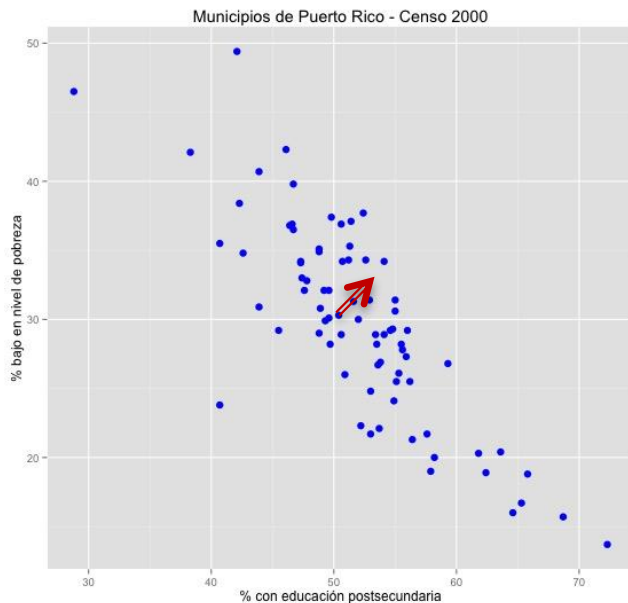
$r = -1$: Relación lineal exacta y negativa.

Karl Pearson (1857-1936). Matemático inglés y bioestadístico. Hizo grandes contribuciones a la estadística matemática, biometría y meteorología. (Fuente: Wikipedia)



¹ Fuente: <http://www.bgfpr.com/economy/documents/PREconomicFactSheet-Dec2014.pdf>

² Fuente: http://www.jp.gobierno.pr/Portal_JP/Portals/0/Formularios/PICA/PERFILDEMOGRAFICO.pdf




Los estudiantes quizás digan que la relación o asociación es inversa o negativa. Lo importante es insistir en una interpretación estadística más no matemática como se discutió anteriormente. Algo importante para comentar de este diagrama es que a medida que aumenta el % de personas con educación postsecundaria la variabilidad de los niveles de pobreza disminuye. Los puntos tienen una forma de V o embudo. Hasta cierto punto, la relación entre las dos variables parece lineal pero en los valores grandes de % de educación la relación parece un poco más curvilínea.

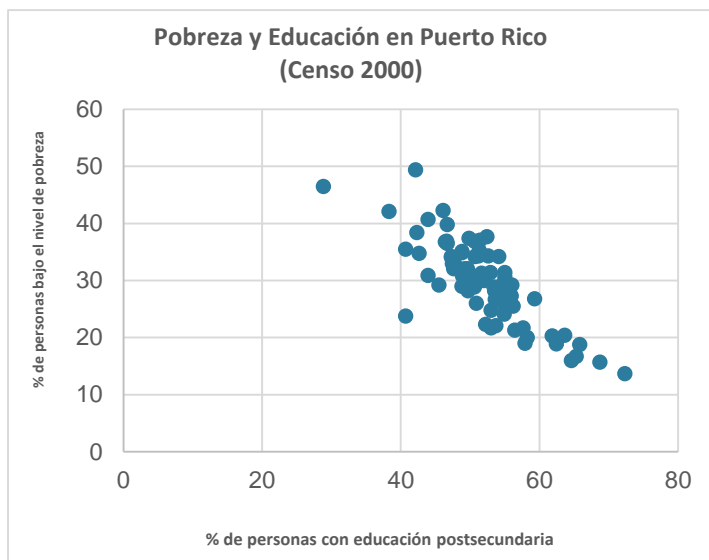
- c) ¿Qué coeficiente de correlación lineal asignaría a la relación entre estas variables? ¿Es la correlación débil, moderada o fuerte? Comente su respuesta.

A primera vista, quizás sea difícil decir un valor en particular. Sabemos del ejercicio anterior que el valor del coeficiente de correlación lineal debe ser negativo. Una posibilidad es dar un intervalo. Sin hacer cálculos, se podría decir que el coeficiente de correlación podría estar entre -0.8 y -0.5. Como la dispersión de los puntos alrededor de una línea recta que pasa por el conjunto de puntos no parece ser tan alta, podríamos decir que la correlación es moderada. ¿Cómo sería un diagrama de dispersión con correlación débil, moderada y fuerte?

- Utilice la tabla de datos de la actividad para estudiar la relación entre estas dos variables. Copie y pegue estos datos a Excel o baje el archivo que se publica con la actividad. Asegúrese que los datos ocupen tres columnas únicamente.

- a) Use Excel para dibujar un diagrama de dispersión similar al que muestra el ejercicio 1). Los datos se encuentran en la Tabla 1 de la actividad. Para hacer la gráfica seleccione las columnas B y C del archivo PobrezaPR.xls. Luego, vaya a "Insert" y, en la sección "Charts", haga clic en "Scatter". 

La gráfica que hagan los estudiantes debe quedar similar a la que se muestra a continuación. En particular, se debe prestar atención a que tenga los títulos en los ejes y en la gráfica. Es posible que algunos estudiantes grafiquen los ejes invertidos. Es decir, la pobreza en el eje X y la educación en el eje Y. Si esto sucede, aproveche para hablar del significado de variable dependiente (Y) e independiente (X). En nuestro caso, estamos interesados en ver cómo cambia la pobreza si hay cambios en el nivel educativo de los pueblos. Por esta razón, consideramos la pobreza como la variable dependiente. Sin embargo, es razonable pensar que el nivel educativo también se afecta por la pobreza pero ese no es nuestro interés en esta actividad.

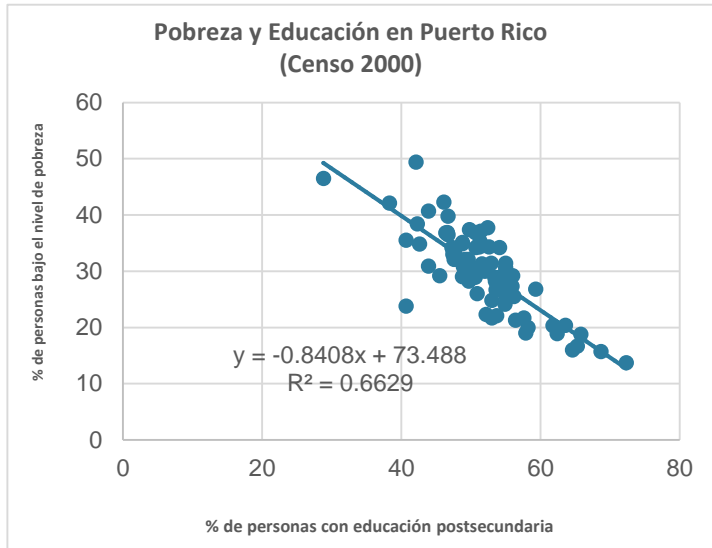


- b) Calcule el coeficiente de correlación lineal y compare el resultado con su respuesta en la parte 1b). Ubíquese en cualquier celda vacía de Excel e ingrese la siguiente fórmula =CORREL(B:B, C:C). Oprima Enter.

El resultado de usar la fórmula de Excel es el siguiente -0.8141. Es decir, el valor estimado del coeficiente de correlación entre el % de personas bajo nivel de pobreza y el % de personas con educación postsecundaria es -0.81. Este valor es relativamente cercano a -1, lo cual indica una correlación alta. Note que este valor está cercano a uno de los valores mencionados en la parte 1b).

3. Ajuste un modelo de regresión lineal. Sustituya cada término de la ecuación $y = mx + b$ por el término correspondiente de la regresión lineal. ¿Qué ecuación se obtiene? Una vez salga la gráfica, busque el ícono “Trendline” (también puede hacer clic derecho sobre los puntos) y seleccione “Linear Trendline”. Luego, oprima “More Options” y marque “Display Equation on chart” y “Display R-squared value on chart”.

Siguiendo las instrucciones anteriores, se obtiene una gráfica similar a la siguiente.



El **método de mínimos cuadrados** es un método de estimación de los coeficientes en un modelo de regresión lineal.

Note que la ecuación del modelo de regresión lineal es: $y = -0.8408x + 73.488$. En términos de las variables del problema original la ecuación es:

$$\% \text{ de pobreza} = -0.8408 * \% \text{ de educación} + 73.488$$

- a) Interprete el coeficiente de determinación del modelo (R^2).

De acuerdo a la gráfica anterior, el coeficiente de determinación del modelo es igual a 0.6629. Esto indica que aproximadamente el 66.29% de la variabilidad del % de personas bajo el nivel de pobreza en los pueblos es explicado por el % de personas con educación postsecundaria. Otras variables no incluidas en el modelo podrían explicar el 33.71% restante.

- b) ¿Cuál es la pendiente? Realice su interpretación en el contexto del problema.

La pendiente del modelo es igual a -0.8408 . La interpretación sería que por cada punto porcentual adicional de personas con educación postsecundaria se espera que el % de personas bajo el nivel de pobreza disminuya en aproximadamente 0.8% . En otras palabras, por cada 10% adicional de personas con educación postsecundaria se espera que el % de personas bajo el nivel de pobreza disminuya en aproximadamente 8% .

- c) Use el modelo de regresión lineal obtenido en la parte 3b) para obtener la estimación puntual del % de pobreza esperado para un pueblo de Puerto Rico donde el 25% de la población mayor de 16 años realiza estudios de postsecundaria.

Usando la ecuación del modelo $y = -0.8408x + 73.488$ debemos reemplazar x por 25 (recuerde que las unidades de la variable son %):

$$y = -0.8408(25) + 73.488 = 52.468.$$

De acuerdo al modelo, un pueblo con un 25% de personas con educación postsecundaria se espera que tenga un 52% de personas bajo el nivel de pobreza. Un ejercicio adicional sería identificar este punto (25% , 52%) en la gráfica para ver que coincide con la línea del modelo. Volvemos a insistir que en la práctica un pueblo podría tener un nivel de pobreza mayor o menor a 52% . Pero si se observaran varios pueblos o el mismo pueblo en distintas ocasiones con el mismo nivel de educación se espera que en promedio tengan ese nivel de pobreza.

- d) Use el modelo de regresión lineal obtenido en la parte 3b) para obtener la estimación puntual del nivel de pobreza esperado para un pueblo de Puerto Rico donde el 96% de la población mayor de 16 años realiza estudios postsecundarios. ¿Existe algún inconveniente en hacer esta predicción? ¿Por qué?

Si sustituimos x por 96 en la ecuación del modelo $y = -0.8408x + 73.488$ tenemos que $y = -0.8408(96) + 73.488 = -7.2288$.

Esto es, se estima que el porcentaje de personas bajo el nivel de pobreza es -7.229 . Claramente, este número no tiene sentido ya que es negativo. El problema aquí no es matemático ya que los cálculos están bien hechos. El problema aquí es que estamos usando el modelo para predecir un valor de pobreza con un nivel educativo muy lejano a los datos observados (un pueblo con un nivel

educativo muy alto). En otras palabras, estamos extrapolando el modelo a un valor lejano de los datos de educación observados. A pesar de que observamos una tendencia lineal en la relación entre las dos variables para pueblos con niveles educativos entre 10% y 50%, es claro que nuestros datos y, por lo tanto, nuestro modelo no debería asumir que dicha relación prevalece para pueblos con niveles educativos muy lejanos a este intervalo.

Adaptaciones e ideas adicionales

Otros recursos:

1. Ingrese a la siguiente página web. Esta página muestra una aplicación en la cual, al variar el coeficiente de correlación y/o el tamaño de una muestra, la gráfica presentada cambia de acuerdo a estas especificaciones dadas.

<http://www.stat.berkeley.edu/~stark/Java/Html/Correlation.htm>

Seleccione un tamaño de muestra de 200 y cambie el valor del coeficiente de correlación lineal a los valores indicados en la siguiente tabla. Para cada uno de estos, ¿qué tipo de relación observa (positiva, negativa o ninguna)? ¿Cuán fuerte es la relación lineal (débil, moderada o fuerte)?

r	Positiva	Negativa	Ninguna	Débil	Moderada	Fuerte
-0.8		X				X
-0.25		X		X		
-0.5		X			X	
0			X	X		
0.1		X		X		
0.6		X			X	
0.75		X				X

2. Ingrese a la siguiente página web del [Instituto de Estadísticas de PR](#) (ver dirección de internet en la nota de pie de página³). Allí podrá encontrar gráficas dinámicas que muestran la relación de varias variables relacionadas a los pueblos de PR a través del tiempo. Estas gráficas dinámicas son de bastante interés para visualizar cambios en la relación de dos variables a través del tiempo.

Estudie la relación de las variables que se indican en la siguiente tabla. Para cada par de variables y año diga qué tipo de relación observa (por ejemplo: lineal débil, no lineal).

Eje X	Eje Y	Año	Tipo de relación
Años de escolaridad (Mediana)	Familias bajo nivel de pobreza (%)	1977	Se observa una relación inversa: a mayores años de escolaridad menor nivel de pobreza. Esta relación parece ser fuerte.
Años de escolaridad (Mediana)	Tasa de desempleo (%)	1963	La tasa de desempleo no parece cambiar cuando cambia la mediana de los años de escolaridad. Esto es señal de que las variables no están correlacionadas.

Use otros pares de variables que sean de interés para estudiar su relación.

³ <http://www.estadisticas.gobierno.pr/iepr/Publicaciones/Proyectosespeciales/GraficasDinamicasMunicipales.aspx>

Tabla 1. Datos de pobreza (% de personas bajo el nivel de pobreza) y nivel educativo (% de personas con educación postsecundaria) en los 78 pueblos de Puerto Rico (Censo 2000).

Pueblo	Postsecundaria	Pobreza	Pueblo	Postsecundaria	Pobreza
Adjuntas	43.9	40.7	Juncos	54.1	28.9
Aguada	50.6	36.9	Lajas	46.6	36.9
Aguadilla	55	30.6	Lares	46.1	42.3
Aguas Buenas	50.6	28.9	Las Marías	38.3	42.1
Aibonito	53.6	26.7	Las Piedras	52.2	22.3
Añasco	43.9	30.9	Loíza	49.6	30.1
Arecibo	55.5	28.2	Luquillo	54.8	29.3
Arroyo	49.2	32.1	Manatí	55.3	26.1
Barceloneta	45.5	29.2	Maricao	28.8	46.5
Barranquitas	51.4	37.1	Maunabo	47.8	32.8
Bayamón	65.3	16.7	Mayagüez	55.6	27.8
Cabo Rojo	55.9	27.3	Moca	46.7	36.5
Caguas	63.6	20.4	Morovis	51.2	34.3
Camuy	49.6	32.1	Naguabo	48.8	29
Canóvanas	49.7	28.2	Naranjito	48.9	30.8
Carolina	64.6	16	Orocovis	42.1	49.4
Cataño	57.6	21.7	Patillas	49.3	29.9
Cayey	53.5	28.2	Peñuelas	48.8	34.9
Ceiba	53	21.7	Ponce	59.3	26.8
Ciales	49.8	37.4	Quebradillas	47.3	34.1
Cidra	54.9	24.1	Rincón	42.6	34.8
Coamo	52	30	Río Grande	55.1	25.5
Comerío	46.4	36.8	Sabana Grande	53.8	26.9
Corozal	52.9	31.4	Salinas	50.7	34.2
Culebra	40.7	23.8	San Germán	53.4	28.9
Dorado	58.2	20	San Juan	65.8	18.8
Fajardo	53.7	22.1	San Lorenzo	50.4	30.3
Florida	47.6	32.1	San Sebastián	48.8	35.1
Guánica	42.3	38.4	Santa Isabel	47.3	34.2
Guayama	56.2	25.5	Toa Alta	62.4	18.9
Guayanilla	52.6	34.3	Toa Baja	56.4	21.3
Guaynabo	72.3	13.7	Trujillo Alto	68.7	15.7
Gurabo	61.8	20.3	Utua	52.4	37.7
Hatillo	54.1	34.2	Vega	50.9	26
Hormigueros	57.9	19	Vega	56	29.2
Humacao	53	24.8	Vieques	40.7	35.5
Isabela	47.4	33	Villalba	51.3	35.3
Jayuya	46.7	39.8	Yabucoa	54.6	29.2
Juana Díaz	51.6	31.3	Yauco	55	31.4