Estadística I & Estadística Descriptiva

Coordinación de Estadística - UFME

Índice general

Presentación del curso

La asignatura Estadística I & Estadística Descriptiva, es el primer curso estadístico de la carrera de Ingeniería Comercial e Ingeniería en Control de Gestion respectivamente, los cuales, entregan las herramientas necesarias para realizar análisis descriptivos de datos, incluyendo formas simples de modelar relaciones entre variables con la intención de facilitar e iluminar la toma de decisiones económicas, de negocios, entre otros ámbitos. Al mismo tiempo, entrega los fundamentos básicos de la teoría de probabilidades para la modelación de fenómenos con base probabilística.

Esta asignatura aspira a enseñar estadística de forma aplicada, haciendo uso de herramientas modernas de programación, situando al estudiante en un rol de analista dentro de una unidad organizacional.

Modalidad de trabajo

Capítulo 1

Tópicos básicos de estadística

1.1. Conceptos

En esta sección repasaremos algunos conceptos claves de la estadística que están asociados a las ciencias cognitivas. Luego, se ahondará en las técnicas estadística básicas de visualización y tabulación de datos, para el estudio de estos.

1.1.1. Datos

El dato es la unidad básica de la estadística. Esta unidad es cualquier evento o hecho que no ha sido dotado de significado, es decir, es cualquier hecho del cual no se puede dar interpretación alguna [Brachman and Levesque, 2004].

Un ejemplo de este concepto, es cuando tratamos de responder la pregunta ¿por qué nos detenemos al caminar, cuando encontramos un semáforo en rojo? ¿Cuál es el dato? ¿Cuál es el significado?

1.1.2. Información

Información = Datos + Significado

Por otro lado, los datos existen independiente de quien observa, y cuando una persona adquiere datos y los dota de significado, estos se convierten en información [Brachman and Levesque, 2004]. Otra forma de entenderlo es:

Información = Datos + Reglas para decodificar

En el ejemplo anterior, el decodificador es la persona que va caminando, y el significado (reglas para decodificar) que le damos al semáforo al estar en rojo, viene de las reglas sociales que indican como actuar en determinadas situaciones.

En estadística, mediante el uso de distintas herramientas (gráficos, tablas, entre otras), dotaremos de significado a los datos, para así generar información de utilidad en distintos fenómenos de estudio.

1.1.3. Tipos de variables

El concepto de datos está fuertemente ligado a su naturaleza, es decir, el contexto de estudio que rodea los datos que se desean estudiar u observar. En este sentido, los datos están asociados a lo que llamamos variable ("naturaleza del dato"), las cuales, se pueden clasificar la siguiente manera:

- **Cualitativas** (Nominales y Ordinales): variables no numéricas que pueden o no llevar un orden, respectivamente.
- **Cuantitativas** (Discretas y Continuas): variables numéricas que pueden o no ser enumeradas, respectivamente.

Ejemplo: Determinar la clasificación de las siguientes variables: tiempo, dinero, altura, cantidad de vecinos en el lugar donde vivo, grado de conformidad (conforme, medianamente conforme, nada conforme) respecto a un servicio, color de pelo de un grupo de personas.

1.1.4. Población y Muestra

Los ingenieros y científicos constantemente están expuestos a la recolección de hecho o datos, tanto en sus actividades profesionales como en sus actividades diarias. La disciplina de estadística proporciona métodos para organizar y resumir datos y de sacar conclusiones basadas en la información contenida en datos.

Una investigación típicamente se enfocará en una colección bien definida de objetos que constituyen una **población** de interés. Cuando la información deseada está disponible para todos lo objetos de la población, se tienen lo que se llama un **censo**. Las restricciones de tiempo, dinero y otros recursos escasos casi siempre hacen que un censo sea infactible. En su lugar, se selecciona un subconjunto de la población, una **muestra**, de manera prescrita [Devore, 2008, pagína 2].

- **Población**: La población es el conjunto de todos los sujetos de interés en un estudio.
- **Muestra**: La muestra es un subconjunto de la población a través de los cuales el estudio recoge los datos.

A continuación, determine la población y muestra de los siguientes enunciados:

- Se realiza un sondeo para determinar los rubros con mayor inflación de venta de mercado en Santiago, para ello se estudia el rubro con mayor ingreso líquido de ventas, en algunas de las comunas de Santiago.
- La encuesta ENUSC elabora anualmente un informe respecto a la seguridad ciudadana, para ello, se contacta a una cantidad de personas deter-

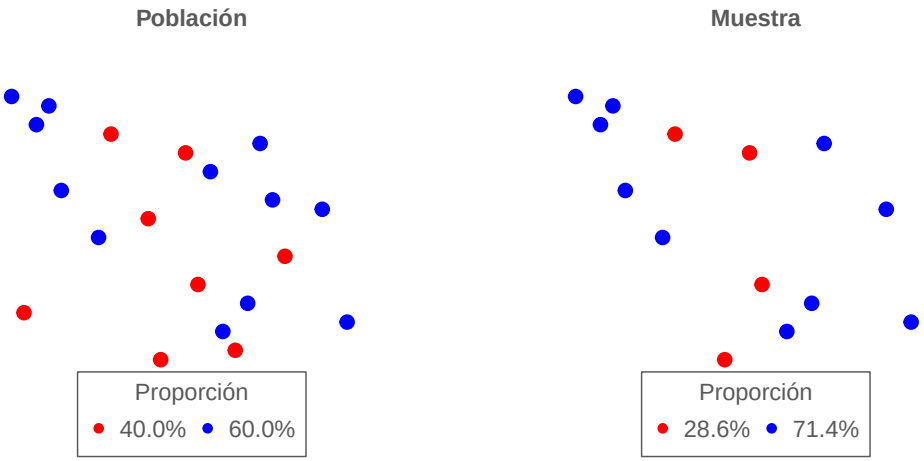
minadas de cada región del país, dando así, resultados a nivel nacional y regional.

1.1.5. Parámetros y Estadísticos

Ambos conceptos son utilizados de manera frecuente en distintos medios de comunicación, cometiéndose el error de tratarlos como sinónimos. Sin embargo, tienen definiciones totalmente distintas:

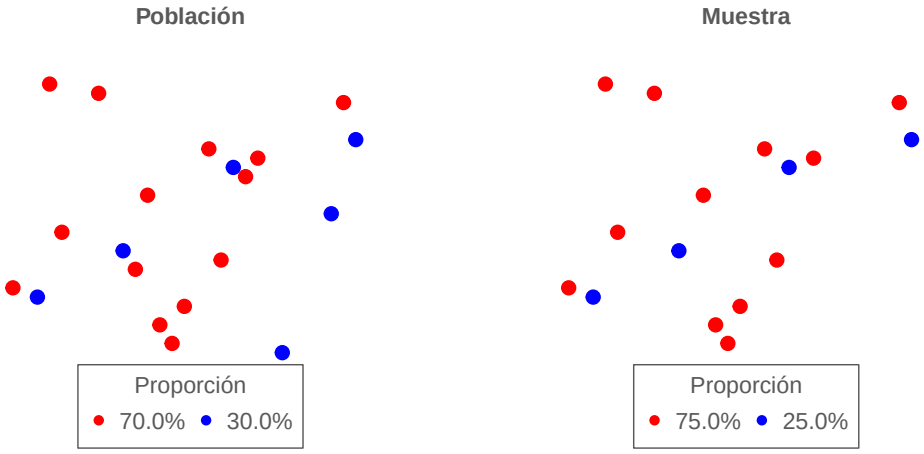
- **Parámetros**: característica numérica de resumen de la población.
- **Estadísticos**: característica numérica de resumen de la muestra.

Veamos el siguiente gráfico, distingamos el parámetro y estadístico correspondiente.



1.1.6. Estimador y Estimación

- **Estimador**: Un estimador es un estadístico usado para aproximar (incertidumbre) el valor de un parámetro. Usualmente no cambia la técnica entre la población y la muestra, por ejemplo, si deseo aproximar la proporción de bolitas rojas en la población, se usaría la proporción de bolitas rojas en la muestra.
- **Estimación**: Una estimación es el número que resulta de aplicar el estimador a una muestra particular. Esto difiera levemente de la definición anterior, ya que en términos estrictos, el estimador solo es la "fórmula", y la estimación es el valor resultante al aplicar la fórmula. Sin embargo, hoy en día es muy común encontrar textos en donde el estimador se considera tanto para la fórmula como para el valor obtenido.



¿Cuándo diríamos que una estimación es buena?

1.1.7. Variabilidad muestral

Efectivamente, al estimación de un parámetro está fuertemente determinada por la muestra con la que uno trabaja. La forma en la que se elige una muestra es azarosa, por lo que es imposible saber de antemano si la estimación será buena o mala respecto al parámetro (error de estimación). En estadística, la forma en la que se elige o genera una muestra puede llegar a ser muy compleja, siendo un tema que está fuera del alcance de este curso. Cabe mencionar que en todo momento la elección es "azarosa'', es decir, no podemos intencionarla en su totalidad.

El concepto detrás de esto es la variabilidad muestral, el cual, indica que dependiendo de la muestra que se obtenga de la población, esta se comportará distinto en relación al estadístico (igualmente para el valor del estimador: estimación).

¿Cuál es la proporción de círculos rojos en la población relfejada en la figura 1.1**?**

¿Qué podríamos inferir sobre el color predominante en la población en base a la muestra de la figura 1.2**?**

¿Y ahora? (Figura 1.3**) Diferentes muestras se comportan de manera diferente. Esto se denomina como variabilidad muestral**

1.1.8. Representatividad y sesgo de la muestra

Discusión: Ambos conceptos se usan con frecuencia en la vida cotidiana, y a su vez están mal empleados. El sesgo no es una propiedad de la muestra sino que del estimador (concepto avanzado de estadística).

Por otro lado, la representatividad no es un concepto válido matemáticamente (no existe tal definición).

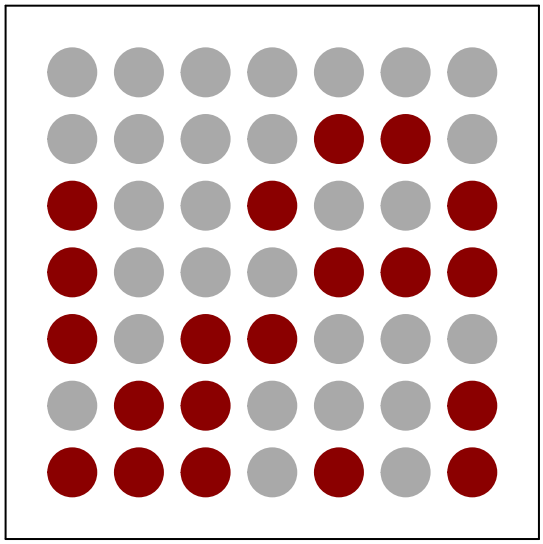


Figura 1.1: Población

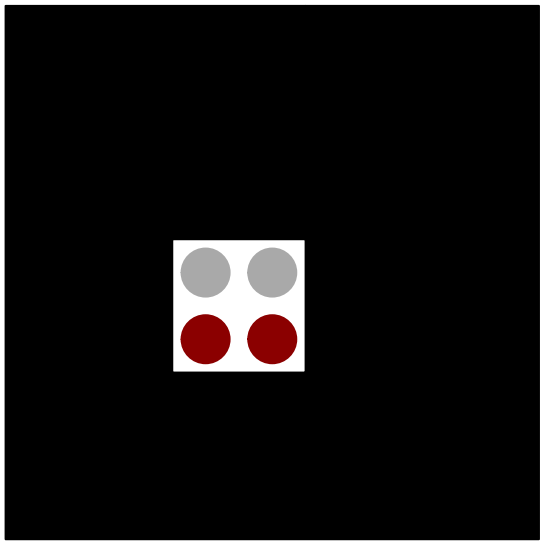


Figura 1.2: Muestra 1

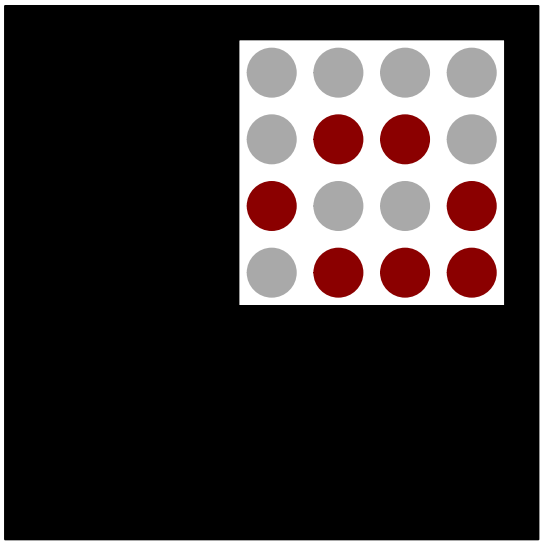


Figura 1.3: Muestra 2

1.1.9. Medidas de Tendencia Central (MTC)

- Media: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Mediana: La mediana de un conjunto de observaciones es el valor para el cual, cuando todas las observaciones se ordenan de manera creciente, la mitad de éstas es menor que este valor y la otra mitad.
- Moda: La moda de un conjunto de observaciones es el valor de la observación que ocurre con mayor frecuencia en el conjunto.

1.1.10. Medidas de Dipersión (MD)

- Varianza: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Desviación Estándar: $\sqrt{S}$

1.2. Gráficos descriptivos

La estadística descriptiva se divide en dos temas generales. En este apartado, se considera la representación de un conjunto de datos por medio de técnicas visuales. A continuación, se hará mención de algunas de las técnicas más útiles y pertinentes a la estadística de probabilidad. Para ello, usaremos la base de datos *rock* que viene incluida en R, la cual, contiene mediciones de 48 muestras de roca de un yacimiento de petróleo. Los las columnas de esta base son:

- area: área del espacio de poros, en píxeles de 256 por 256.
- peri: perímetro en píxeles.
- shape: perímetros dividido por la raíz cuadrada del área.
- perm: permeabilidad en mili-Darcies (unidad que se utiliza para cuantificar la capacidad de permeabilidad de un fluido a través de una roca; más información en este link).

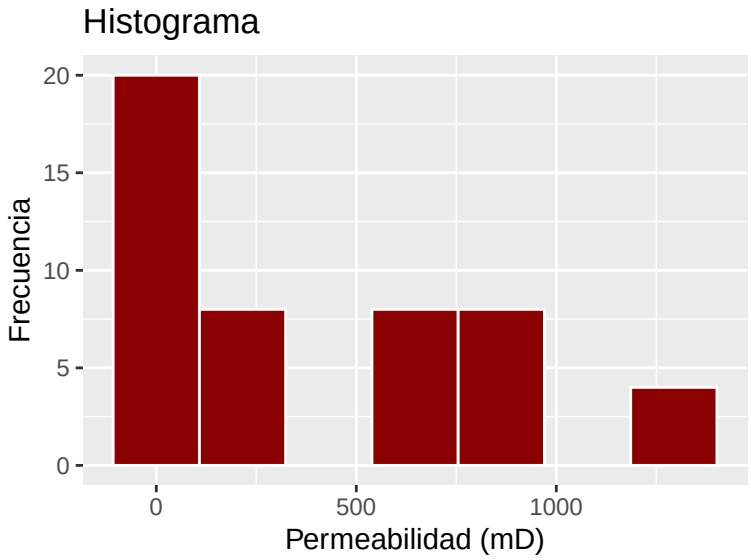
1.2.1. Histograma

Algunos datos numéricos se obtienen contando para determinar el valor de una variable (cuántas veces se repite un hecho), mientras que otro datos se obtienen tomando mediciones (peso, altura, tiempo de reacción). Usualmente, este tipo de gráfico se utiliza con datos continuos (aunque tiene una versión para datos discretos), para lo cual, se debe hacer lo siguiente [Devore, 2008, pagína 12]:

1. Subdividir los datos en **intervalos de clase** o **clases**, de tal manera que cada observación quede contenida en exactamente una clase. Para esto, se hace uso de la regla de Sturges, la cual, consiste en calcular la expresión $1 + \log_2(n)$, aproximando hacia el entero más próximo, dónde n corresponde a la cantidad de datos (existen otra variedad de técnicas).
2. Determinar la frecuencia y la frecuencia relativa de cada clase, es decir, cuántas observaciones hay en cada uno de los intervalos.
3. Se marcan los límite de clase sobre el eje horizontal del plano cartesiano.
4. Se traza un rectángulo cuya altura es la frecuencia absoluta (o relativa) correspondiente a cada intervalo de clase.

Para obtener el histograma en R, a partir de un conjunto de datos, se utiliza el siguiente código:

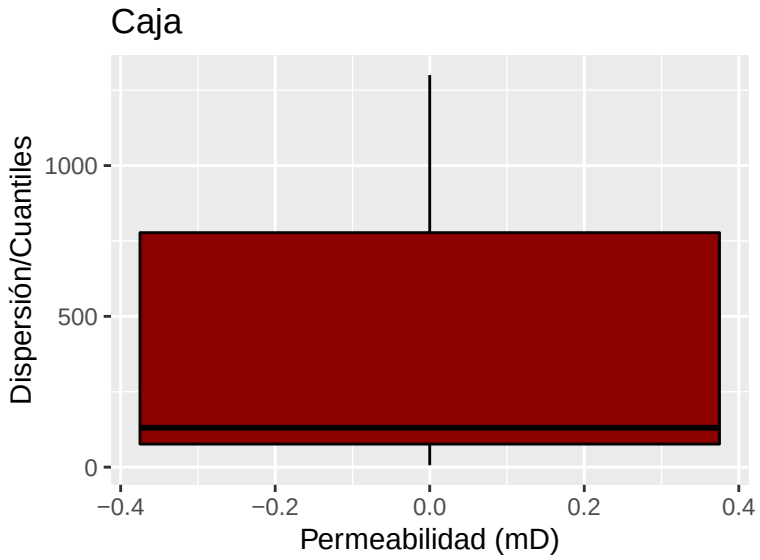
```
library(ggplot2)
datos = rock
ggplot(data = datos, aes(x = perm)) +
  geom_histogram(color = "white", fill = "darkred", bins = 7) +
  labs(title = "Histograma", x = "Permeabilidad (mD)", y = "Frecuencia")
```



1.2.2. Caja

Este gráfico se utiliza para evidenciar las medidas de posición conocidas como cuartiles.

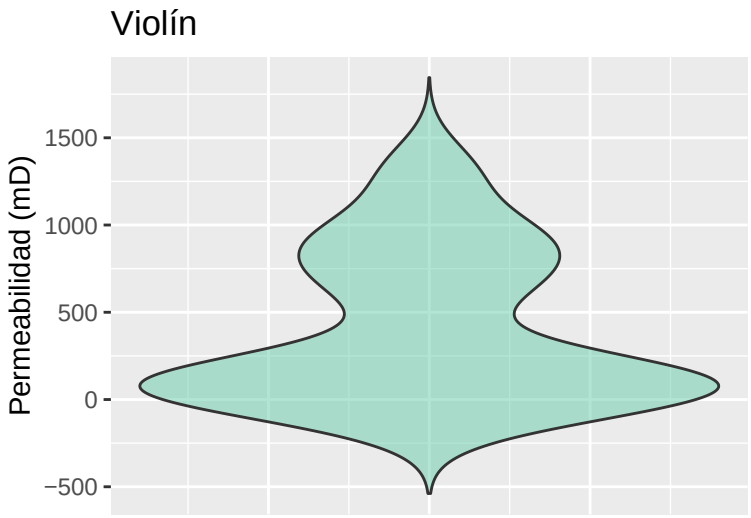
```
ggplot(data = datos, aes(y = perm)) +
  geom_boxplot(color = "black", fill = "darkred") +
  labs(title = "Caja", x = "Permeabilidad (mD)", y = "Dispersión/Cuantiles")
```



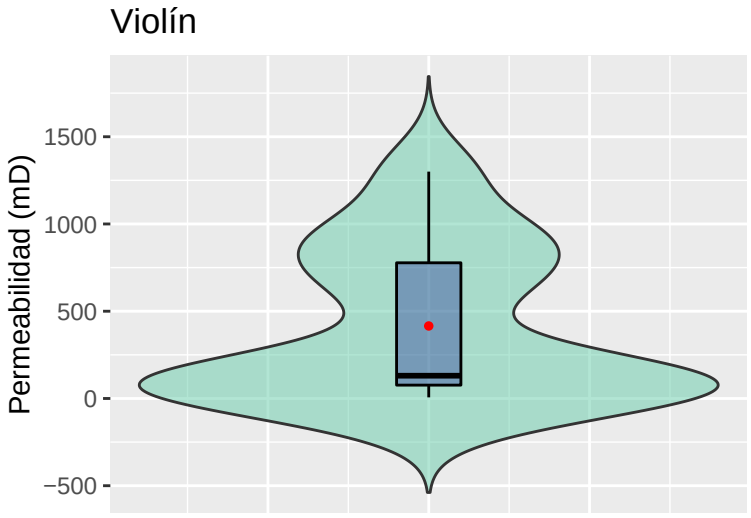
1.2.3. Violín

Gráfico utilizado para conocer la concentración de datos; es muy similar al histograma.

```
g = ggplot(data = datos, aes(x = 1, y = perm)) +
  geom_violin(trim = F, alpha = 0.5, fill = "aquamarine3") +
  labs(title = "Violín", x = "", y = "Permeabilidad (mD)") +
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank())
g
```



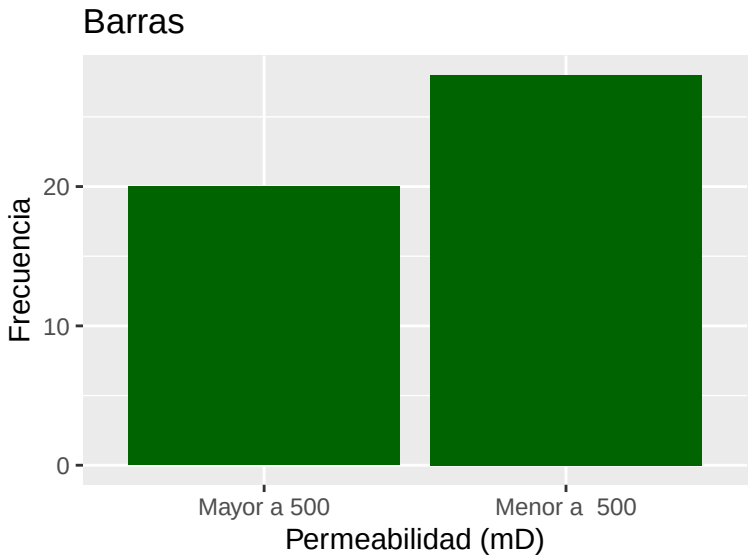

```
g + geom_boxplot(width = 0.1, color = "black",
                 alpha = 0.3, fill = "darkblue")  +
  stat_summary(fun = mean, geom = "point",
               size = 1, color = "red")
```



1.2.5. Barras

Gráfico por excelencia para visualizar la frecuencia de variables cualitativas.

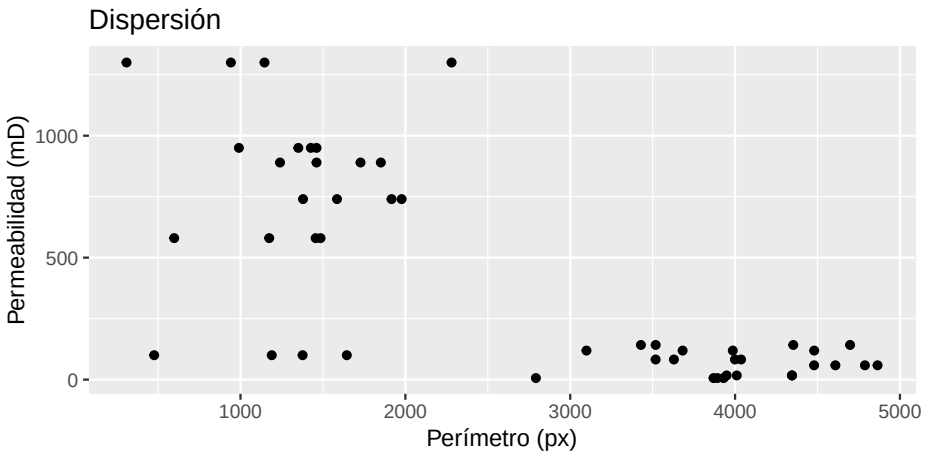
```
datos$Perm2 = ifelse(datos$perm > 500, "Mayor a 500", "Menor a  500")
ggplot(data = datos,aes(x = Perm2)) +
  geom_bar(fill = "darkgreen") +
  labs(title = "Barras", x = "Permeabilidad (mD)", y = "Frecuencia")
```



1.2.6. Dispersión

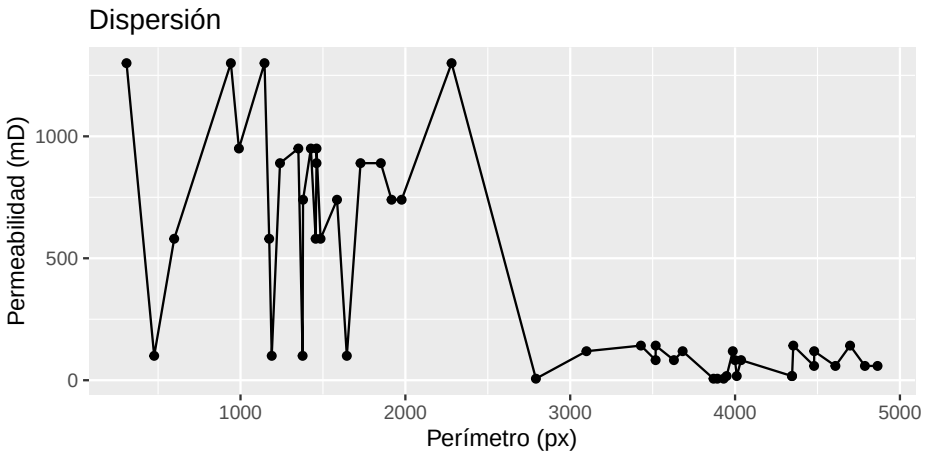
Gráfico que permite contrastar dos variables cuantitativas. En general, se usa para estudiar la asociación entre dos variables.

```
ggplot(data = datos, aes(x = peri, y = perm)) +
  geom_point() +
  labs(title = "Dispersión", x = "Perímetro (px)", y = "Permeabilidad (mD)")
```



1.2.7. Dispersión + líneas

```
ggplot(data = datos, aes(x = peri, y = perm)) +
  geom_point() +
  geom_line() +
  labs(title = "Dispersión", x = "Perímetro (px)", y = "Permeabilidad (mD)")
```



1.3. Ejercicios

Desarrollar el Taller 1.

Capítulo 2

Probabilidad y variables aleatorias.

Capítulo 3

Distribuciones muestrales y pruebas de hipótesis

Capítulo 4

Intervalos de confianza

4.1. Una media

4.1.1. Bajo distribución normal

4.1.2. Asintótico

4.2. Diferencia de medias

4.3. Compraración de varianzas

4.4. Diferencia de proporciones

Bibliografía

Ronald J. Brachman and Hector J. Levesque. *Knowledge representation and reasoning*. Morgan Kaufmann, Amsterdam ; Boston, 2004. ISBN 978-1-55860-932-7.

Jay L. Devore. *Probability and statistics for engineering and the sciences*. Thomson/Brooks/Cole, Belmont, CA, 7th ed edition, 2008. ISBN 978-0-495-38217-1 978-0-495-38223-2.