Histograms and density curves

What's in our toolkit so far?

- Plot the data: histogram (or stemplot)
- Look for the overall pattern and identify deviations and outliers
- Numerical summary to briefly describe center and spread

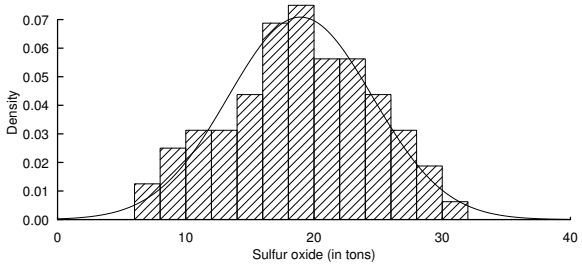
A new idea:

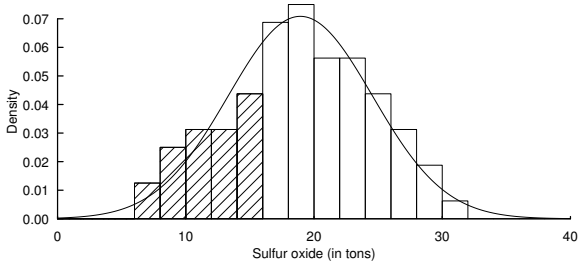
> If the pattern is sufficiently regular, approximate it with a smooth curve.

Any curve that is always on or above the horizontal axis and has total are underneath equal to one is a **density curve**.

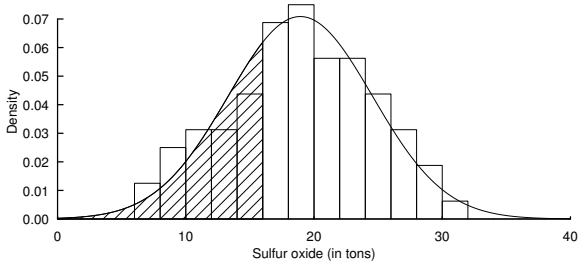
- Area under the curve in a range of values indicates the proportion of values in that range.
- Come in a variety of shapes, but the "normal" family of familiar bell-shaped densities is commonly used.
- Remember the density is only an approximation, but it simplifies analysis and is generally accurate enough for practical use.

Examples

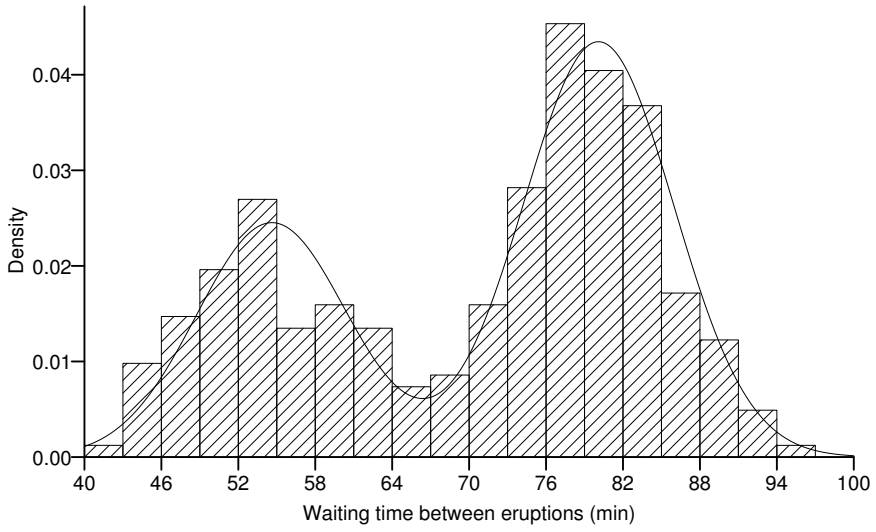




Shaded area of histogram: 0.29



Shaded area under the curve: 0.30



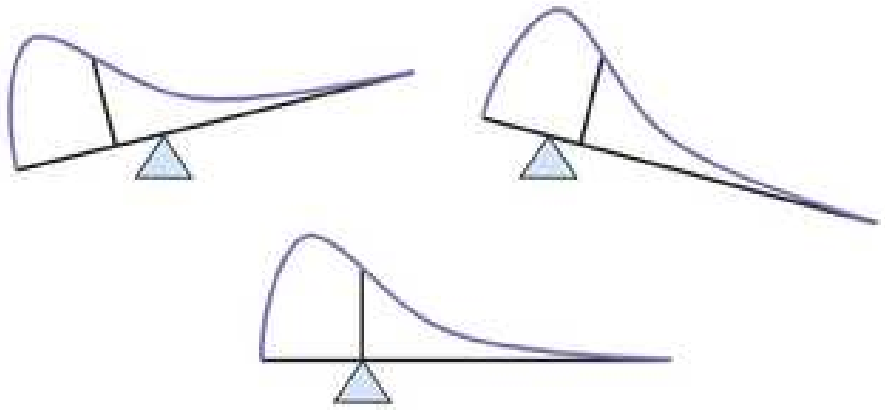
Median and mean of a density curve

Median:

The equal-areas point with 50% of the "mass" on either side.

Mean:

The balancing point of the curve, if it were a solid mass.



Note:

- The mean and median of a symmetric density curve are equal.
- The mean of a skewed curve is pulled away from the median in the direction of the long tail.

The mean and standard deviation of a density are denoted μ and σ, rather than $\bar{x}$ and s, to indicate that they refer to an idealized model, and not actual data.

Normal distributions: $\mathcal{N}(\mu, \sigma)$

The normal distribution is

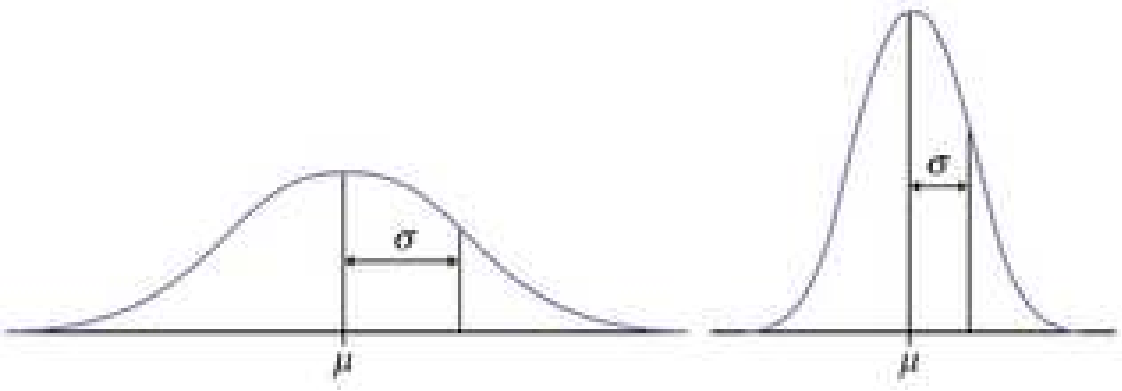
- symmetric,
- single-peaked,
- bell-shaped.

The density curve is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(X-\mu)^2 \Big).$$

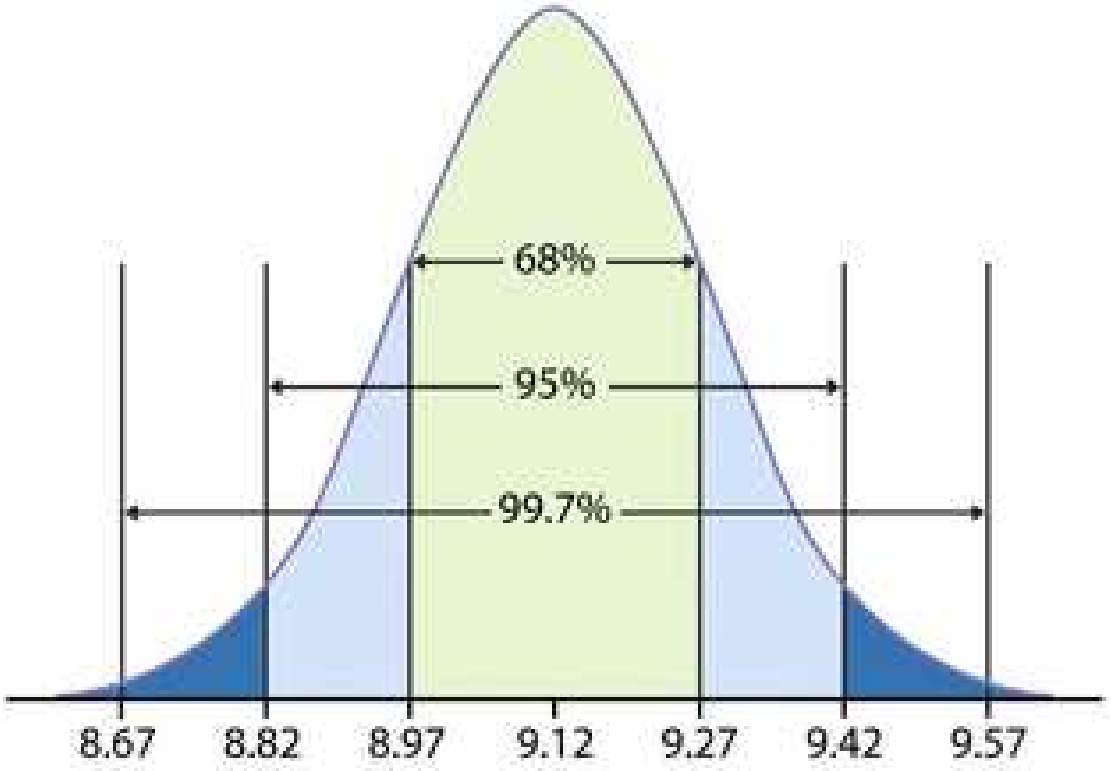
It is determined by two parameters μ and σ:

- μ is the mean (also the median)
- σ is the standard deviation



Note: The point where the curve changes from concave to convex is σ units from μ in either direction.

The 68-95-99.7 rule



- About 68% of the data fall inside $(\mu - \sigma, \mu + \sigma)$.
- About 95% of the data fall inside $(\mu - 2\sigma, \mu + 2\sigma)$.
- About 99.7% of the data fall inside $(\mu - 3\sigma, \mu + 3\sigma)$.

Example

Scores on the Wechsler Adult Intelligence Scale (WAIS) for the 20 to 34 age group are approximately $N(110, 25)$.

- About what percent of people in this age group have scores above 110?

- About what percent have scores above 160?

- In what range do the middle 95% of all scores lie?

Standardization and z-scores

Linear transformation of normal distributions:

$$X \sim \mathcal{N}(\mu, \sigma) \quad \Rightarrow \quad a\,X + b \sim \mathcal{N}(a\,\mu + b, a\,\sigma)$$

In particular it follows that

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

$\mathcal{N}(0, 1)$ is called **standard normal distribution**.

For a real number x the standardized value or ***z*-score**

$$z = \frac{x - \mu}{\sigma}$$

tells how many standard deviations x is from μ, and in what direction.

Standardization enables us to use a standard normal table to find probabilities for any normal variable.

For example:

- What is the proportion of $N(0, 1)$ observations less than 1.2?
- What is the proportion of $N(3, 1.5)$ observations greater than 5?
- What is the proportion of $N(10, 5)$ observations between 3 and 9?

Normal calculations

Standard normal calculations

1. State the problem in terms of x.
2. Standardize: $z = \frac{x-\mu}{\sigma}$.
3. Look up the required value(s) on the standard normal table.
4. Reality check: Does the answer make sense?

Backward normal calculations

We can also calculate the values, given the probabilities:

If MPG $\sim \mathcal{N}(25.7, 5.88)$, what is the minimum MPG required to be in the top 10%?

"Backward" normal calculations

1. State the problem in terms of the probability of being **less** than some number.
2. Look up the required value(s) on the standard normal table.
3. "Unstandardize," i.e. solve $z = \frac{x-\mu}{\sigma}$ for x.

Example

Suppose $X \sim \mathcal{N}(0, 1)$.

- $\mathbb{P}(X \leq 2) = ?$
- $\mathbb{P}(X > 2) = ?$
- $\mathbb{P}(-1 \leq X \leq 2) = ?$
- Find the value z such that
  - $\mathbb{P}(X \leq z) = 0.95$
  - $\mathbb{P}(X > z) = 0.99$
  - $\mathbb{P}(-z \leq X < z) = 0.68$
  - $\mathbb{P}(-z \leq X < z) = 0.95$
  - $\mathbb{P}(-z \leq X < z) = 0.997$

Suppose $X \sim \mathcal{N}(10, 5)$.

- $\mathbb{P}(X < 5) = ?$
- $\mathbb{P}(-3 < X < 5) = ?$
- $\mathbb{P}(-x < X < x) = 0.95$

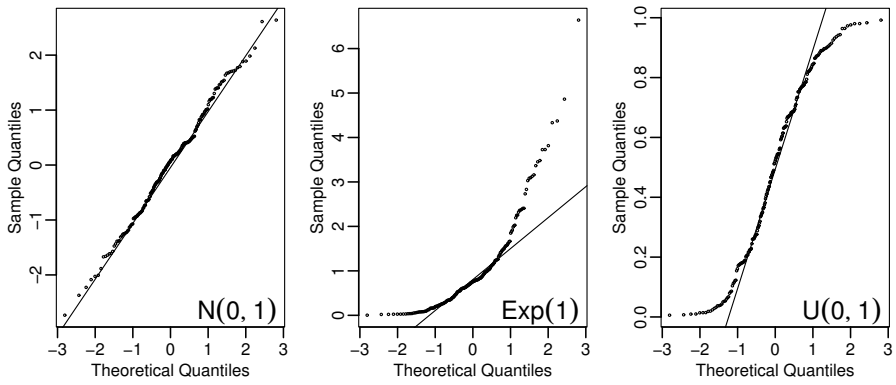
Assessing Normality

How to make a normal quantile plot

1. Arrange the data in increasing order.
2. Record the percentiles $(\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n})$.
3. Find the z-scores for these percentiles.
4. Plot x on the vertical axis against z on the horizontal axis.

Use of normal quantile plots

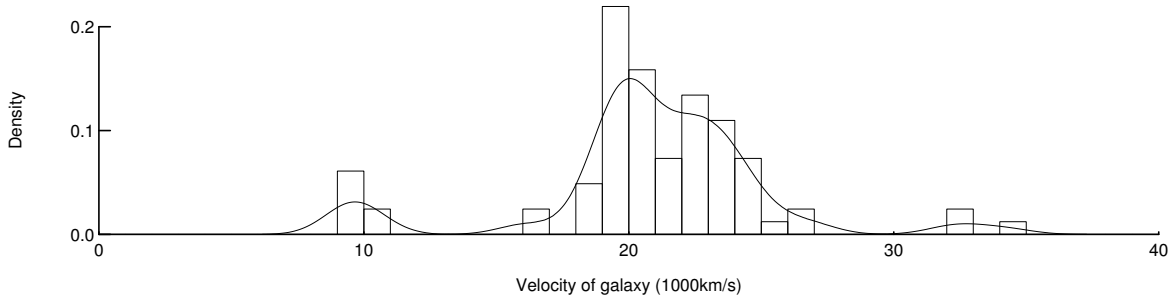
- If the data are (approximately) normal, the plot will be close to a straight line.
- Systematic deviations from a straight line indicate a nonnormal distribution.
- Outliers appear as points that are far away from the overall patter of the plot.



Density Estimation

The normal density is just one possible density curve. There are many others, some with compact mathematical formulas and many without.

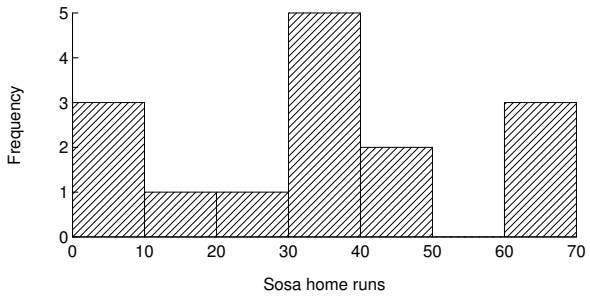
Density estimation software fits an arbitrary density to data to give a smooth summary of the overall pattern.



Histogram

How to scale a histogram?

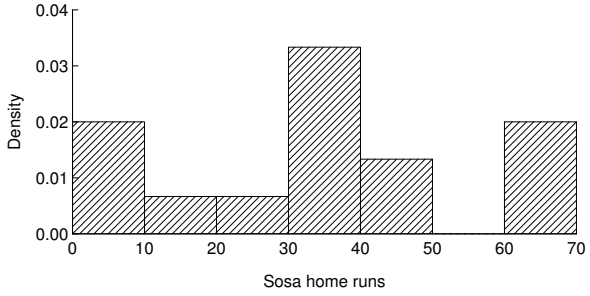
- Easiest way to draw a histogram:
  - ◇ qqually spaced bins
  - ◇ counts on the vertical axis



Disadvantage: Scaling depends on number of observations and bin width.

- Scale histogram such that area of each bar corresponds to proportion of data:

$$\text{height} = \frac{\text{counts}}{\text{width} \cdot \text{total number}}$$

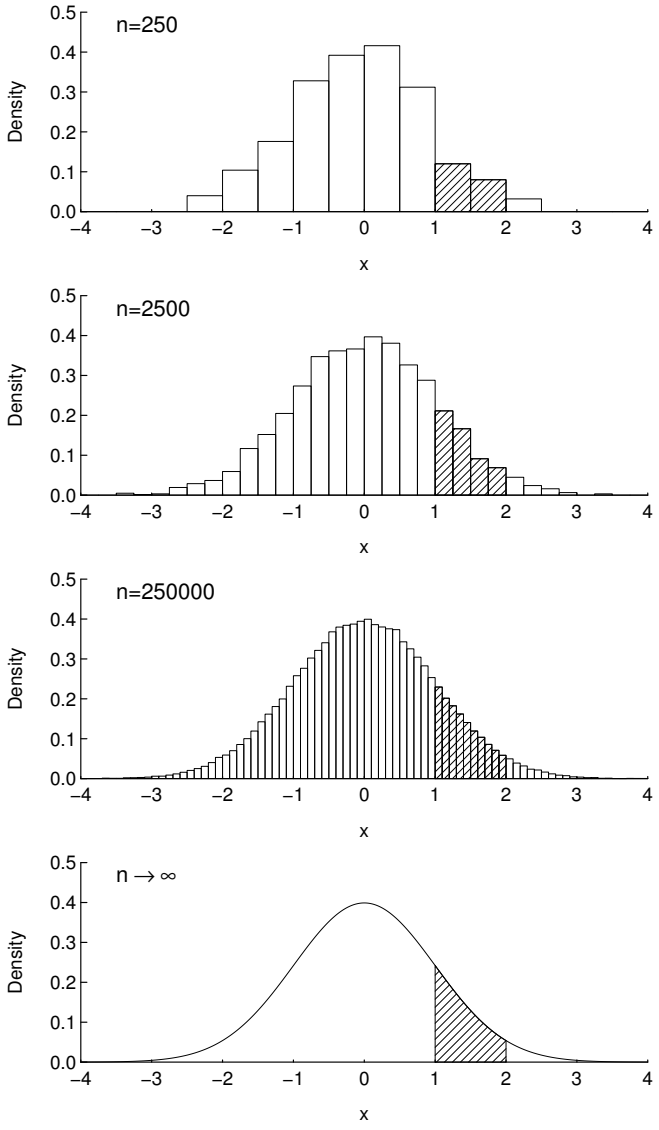


Proportion of data in interval $(0, 10]$:

$$\text{height} \cdot \text{width} = 0.02 \cdot 10 = 0.2 = 20\%$$

Since $n = 15$ this corresponds to 3 observations.

Density curves



Proportion of data in (1,2]:

$$\frac{\#\{x_i : 1 < x_i \leq 2\}}{n}$$

$\downarrow n \to \infty$

$$\int_1^2 f(x)\,dx$$

Probability that a new observation X fall into $[a, b]$

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx = \lim_{n\to\infty} \frac{\#\{x_i : 1 < x_i \leq 2\}}{n}$$