Data describing

mean: $\frac{1}{N}\sum_{i=1}^{N} x_i$

> p27

- scaling: `mean(kx)=kmean(x)`
- translating: `mean(x+c)=mean(x)+c`
- $\sum_{i=1}^{N}(x - mean(\{x\})) = 0$
- sum of squared distances of data points to *mean* is minimized
- affect strongly by outlier

standard deviation: $std(\{x\}) = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(x - mean(\{x\}))^2}$

> p29

- when std is small, most data tend to close to mean
- $std(\{kx_i\}) = k \cdot std(\{s_i\})$
- scalable
- there are at most $\frac{1}{k^2}$ data points lying k or more standard deviations away from the mean.
- there must be at least one data item that is at least one standard deviation away from the mean
- referred as scale parameter

variance: $var(\{x\}) = \frac{1}{N}(\sum_{i=1}^{i=N}(x_i - mean(\{x\}))^2)$

> p31

- translating
- $var(k) = 0$, where k is a constant
- $var(\{kx\}) = k^2 var(\{x\})$

median: another use of a mean, less affect by outlier

- scalable
- translating

interquartile range:

> p34

The interquartile range of a dataset $\{x\}$ is $iqr(\{x\}) = percentile(\{x\}, 75) - percentile(\{x\}, 25)$

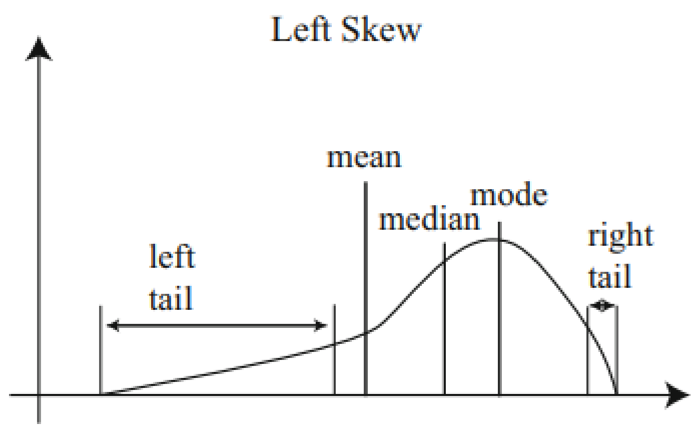
- estimate how spread the data is, regardless the affect by outlier
- $iqr(x + c) = iqr(x)$$
- $iqr(kx) = |k| \cdot iqr(x)$

graph

histogram:

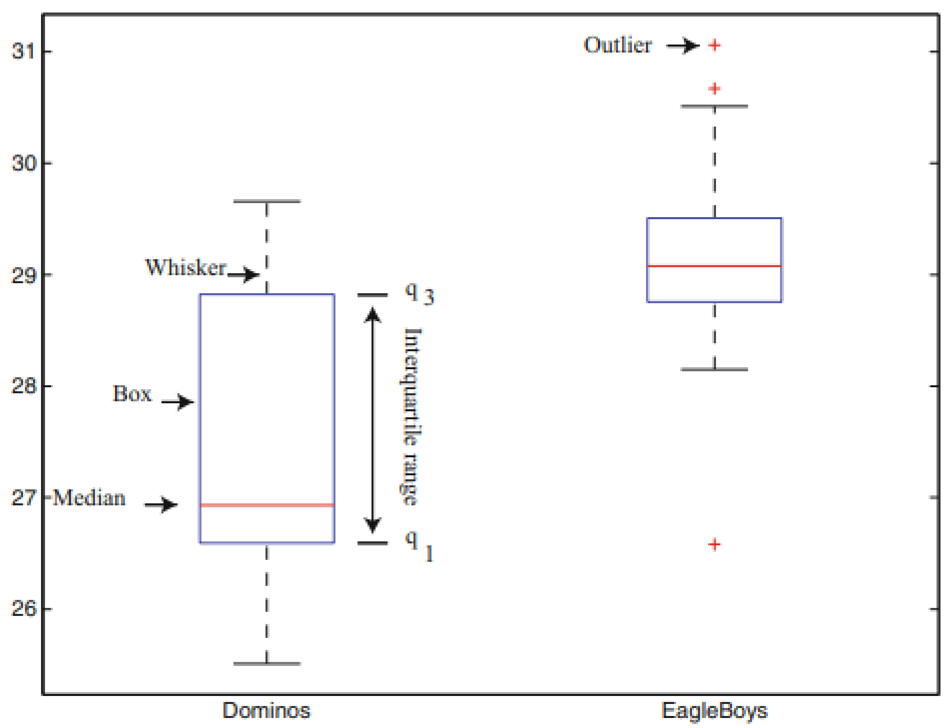
> p35

- bar chart vs histogram: bar char is for category while histogram for quantitative data
- uni/multi modal: unimodal has one peak, multimodal has many, bimodal has two
- skew: symmetric, left skew, right skew, left skew refer to its tail is long on left



box plot:

A box plot is a way to plot data that simplifies comparison



outlier: data item that are larger than $q_3 + 1.5(q_3 - q_1)$ or smaller than $q_1 - 1.5(q_3 - q_1)$

whisker: non-outlier data

standardized coordinate

> p37

coordinate with normalized data

$$\hat{x_i} = \frac{x_i - mean(\{x\})}{std(\{x\})}$$

- mean of standard coordinate is equal to 0
- standard deviation is equal to 1
- for many kinds of data, histograms of these standard coordinates look the same, which is the **standard normal curve**, given by:

$$y(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

- data in standard coordinate is called the normal data

correlation:

$$corr(\{(x, y)\}) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$$

- range from -1 to 1, the larger (absolute value), the better predict
- sign represent positive/negative correlation
- 0 means no correlation,1 means $\hat{x}_i = \hat{y}_i$
- $corr(\{(x, y\}) = corr(\{y, x\})$
- The value of the correlation coefficient is not changed by translating the data.
- Scaling the data can change the sign, but not the absolute value
- $corr(\{ax_i + b, cy_i + d\}) = sign(a \cdot c) corr(\{x_i, y_i\})$

predict:

> p62

1. Transform the data set into standard coordinates
2. Compute the correlation r
3. predict $\hat{y_0} = r\hat{x_0}$
4. transform back into original coordinate

- Rule of Thumb: The predicted value of y goes up by r standard deviations when the value of x goes up by one standard deviation.
- root mean square error: $\sqrt{1 - r^2}$

probability

> p70

outcome: what we expect from the experiment, every run of the experiment produces exactly one of the set of possible outcomes

sample space: the set of all outcomes, which we usually write Ω

event: event is a set of outcomes, usually write as sets, for example, ε

- $P(\Omega) = 1$
- $P(\emptyset) = 0$
- denote A_j as a set of disjoined event, that is $A_i \cap A_j = \emptyset$ where $i \neq j$, we have:

$$P(\cap_i A_i) = \sum_i P(A_i)$$

combination:

> p74

regardless the order, number of outcome when select k from N

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

probability calculating:

$$P(A) + P(A^c) = 1$$
$$P(A - B) = P(A) - P(A \cap B)$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

> application:
>
> **Worked example 3.14 (Dice)** You flip two fair six-sided dice, and add the number of spots. What is the probability of getting a number divisible by 2, but not by 5?
>
> **Solution** There is an interesting way to work the problem. Write $\mathcal{D}_n$ for the event the number is divisible by n. Now $P(\mathcal{D}_2) = 1/2$ (count the cases; or, more elegantly, notice that each die has the same number of odd and even faces, and work from there). Now $P(\mathcal{D}_2 - \mathcal{D}_5) = P(\mathcal{D}_2) - P(\mathcal{D}_2 \cap \mathcal{D}_5)$. But $\mathcal{D}_2 \cap \mathcal{D}_5$ contains only three outcomes (6, 4, 5, 5 and 4, 6), so $P(\mathcal{D}_2 - \mathcal{D}_5) = 18/36 - 3/36 = 5/12$

Conditional probability

> P84

the probability that B occurs given that A has definitely occurred. We write this as $P(B|A)$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

- $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$

independent:

Two events A and B are independent if and only if $P(A \cap B) = P(A)P(B)$

In other form, if two events are independent, $P(A|B) = P(A)$ and $P(B|A) = P(B)$, or in simple put:

$$P(A \cap B) = P(A)P(B)$$

- pairwise independent: each pair in events list is independent. pairwise independent cannot illustrate independent.
- conditional independent: $P(A_1 \cap \ldots \cap A_n | B) = P(A_1|B) \ldots P(A_n|B)$

Random variables

> P103

Given a sample space , a set of events F, a probability function P, and a countable set of real numbers D, a discrete random variable is a function with domain Ω and range D.

probability distribution function: $P(\{X = x\})$

cumulative distribution function: $P(\{X \leq x\})$

join probability function: $P(\{X = x\} \cap \{Y = y\}) = P(x, y)$

Bayes' Rule:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

independent random variable: $P(x, y) = p(x)p(y)$

probability density function

> P107

Let $p(x)$ be a probability density function (often called a pdf or density) for a continuous random variable X. We interpret this function by thinking in terms of small intervals. Assume that dx is an infinitesimally small interval. Then: $p(x)dx = P$

- no negative
- $\int_{-\infty}^{\infty} p(x)dx = 1$

normalizing constant: $\frac{1}{\int_{-\infty}^{\infty} g(x)dx}$

Expected Values

> P110

Given a discrete random variable X which takes values in the set D and which has probability distribution P, we define the expected value:

$$\mathbb{E}[X] = \sum_{x \in D} xP(X = x) = \mathbb{E}_p[X]$$

for the continuous random variable X which takes value in the set D, and which has probability distribution P, we define the expect value as:

$$\mathbb{E}[X] = \int_{x \in D} xp(x)dx = \mathbb{E}_p[X]$$

Assume we have a function f that maps a continuous random variable X into a set of numbers D_f. Then $f(X)$ is a continuous random variable, too, which we write F. The expected value of this random variable is:

$$\mathbb{E}[f] = \int_{x \in D} f(x)p(x)dx = \text{the expection of } f$$

- $\mathbb{E}[0] = 0$
- for any constant k, $\mathbb{E}[kf] = k\mathbb{E}[f]$
- $\mathbb{E}[f+g] = \mathbb{E}[f] + \mathbb{E}[g]$
- expectation are linear
- the mean/expect value of random variable X is $\mathbb{E}[X]$

variance of random variable:

$$var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

- for constant k, $var[k] = 0$
- $var[kX] = k^2 var[X]$
- if X, Y are independent, then $var[X+Y] = var[X] + var[Y]$

covariance for expected value:

$$cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

- if X, Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- if X, Y are independent, then $cov(X,Y) = 0$
- $var[X] = cov(X,X)$
- $corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$
- $var(X-Y) = var(X) + var(Y) - 2cov(X,Y)$
- $var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$

standard deviation of random variable:

$$std(\{X\}) = \sqrt{var[X]}$$

Markov's inequality:

> P116

the probability of a random variable taking a particular value must fall off rather fast as that value moves away from the mean

$$P(\{|X| \geq a\}) \leq \frac{\mathbb{E}[|X|]}{a}$$

Chebyshev's inequality:

- give us the weak law of large number

$$P(\{|X - \mathbb{E}[X]| \geq k\sigma\}) \leq \frac{1}{k^2}$$

indicator function:

An indicator function for an event is a function that takes the value zero for values of x where the event does not occur, and one where the event occurs. For the event E, we write:

$$\mathbb{I}_{[\{|x|\} \leq a]}(x) = \begin{cases} 1 \text{ if } & -a < x < a \\ 0 & \text{otherwise} \end{cases}$$

- $\mathbb{E}_P[\mathbb{I}_{[\varepsilon]}] = P(\varepsilon)$

Distribution

> P131

discrete uniform distribution:

> e.g. fair die, fair coin flip

A random variable has the discrete uniform distribution if it takes each of k values with the same probability $\frac{1}{k}$, and all other values with probability zero.

Bernoulli Random Variables:

> e.g. biased coin toss

Bernoulli random variable takes the value 1 with probability p and 0 with probability $1-p$. This is a model for a coin toss, among other things

- $mean = p$
- $variance = p(1-p)$

The Geometric Distribution:

> e.g. we flip this coin until the first head appears, the number of flip required to get one head

$$P(\{X = n\}) = (1-p)^{n-1}p$$

- $mean = \frac{1}{p}$
- $variance = \frac{1-p}{p^2}$

The Binomial Probability Distribution:

> e.g. toss a coin, the probability that it comes up head h times in N flips

$$P_b(h; N, p) = \binom{N}{h} p_h (1-p)^{N-h}$$

- as long as $0 \leq h \leq N$, in other case, the probability is equal to 0
- $mean = Np$
- $variance = Np(1-p)$
- different with Bernoulli: binomial represents the number of successes in n successive independent trials of a Bernoulli experiment

Multinomial Probabilities:

> e.g. toss a die with k sides, the probability that it comes up a outcome in N flips

Definition 5.5 (Multinomial Distribution) Perform N independent repetitions of an experiment with k possible outcomes. The i'th such outcome has probability p_i. The probability of observing outcome 1 n_1 times, outcome 2 n_2 times, etc. (where $n_1 + n_2 + n_3 + \ldots + n_k = N$) is

$$P_m(n_1, \ldots, n_k; N, p_1, \ldots, p_k) = \frac{N!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}.$$

The Poisson distribution:

> e.g. the marketing phone calls you receive during the day time

$$P(\{X = k\}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where $\lambda > 0$ is a parameter often known as the intensity of the distribution

- $mean = \lambda$
- $variance = \lambda$