

# STAT400 Notes

## Table of contents

1. Basic Terms .....	1
2. Probability functions .....	2
2.1. Sigma algebra .....	2
2.2. Probability functions .....	2
3. Independent events .....	2
4. Random variables .....	2
5. Discrete distributions .....	2
5.1. Geometric distribution .....	2
6. Continuous distributions .....	2
6.1. Normal distribution .....	2
6.2. Gamma distribution .....	2
6.3. Exponential distribution .....	3
6.4. Chi-squared distribution .....	3
7. Relationships between variables .....	3
7.1. Covariance .....	3
7.2. Correlation coefficient .....	4
8. Linear combinations of random variables .....	4
9. Random samples and statistics .....	4
9.1. Central Limit Theorem .....	4
10. Estimators .....	5
10.1. Parametric estimation .....	5
10.1.1. Point estimators .....	5
10.2. Method of Moments .....	5

## 1. Basic Terms

**Parameters** Quantitative features of a population

**Statistics** Quantitative features calculated using a sample

**Deduction** A way of going from the population to the sample, e.g., if you get the class average (population) and then compare it to your personal score on an exam (sample), you can deduce that you did better than the average.

**Inference** A way of going from the sample to the population (opposite direction from deduction). For example, if you get a 100 on an exam (sample) and you infer that you did better than the class average (population).

Inference and deduction goes in a circle. When developing a model, you estimate some parameters, then you make deductions about the population, then you infer population parameters, then you draw deductions about the population again.

**Experiment** Repeatable task with well-defined outcomes

**Sample space** The sample space for an experiment is the set of all possible outcomes of that experiment, denoted by  $S$

**Event** Any subset  $E$  of the sample space  $S$  attached to an experiment will be called an event associated with the experiment. To say that an event  $E$  has happened means that the outcome of the experiment was in the set  $E$ .

**Simple event** An event with only one event in it

## 2. Probability functions

### 2.1. Sigma algebra

A sigma algebra is a collection of interesting events in some sample space.

A collection  $\mathcal{B}$  of subsets of a sample space  $S$  is a sigma algebra if:

1.  $\emptyset \in \mathcal{B}$
2.  $\forall A \in \mathcal{B}, A^c \in \mathcal{B}$
3. If  $\{A_i : i \in \mathbb{N}\}$  is a countable collection such that  $A_i \in \mathcal{B}$  for all  $i$ , then  $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{B}$

### 2.2. Probability functions

Consider a sample space  $S$  with a sigma algebra  $\mathcal{B}$ .

A probability function is a function from events to probabilities ( $\mathcal{B} \rightarrow \mathbb{R}$ ). It must satisfy the following axioms:

1. (finite measure)  $P(S) = 1$
2. (positivity)  $\forall A \in \mathcal{B}, P(A) \geq 0$
3. (countable additivity) For  $A_1, A_2, A_3, \dots$ , the collection of pairwise disjoint subsets of  $S$  in  $\mathcal{B}$ , we must have

$$P\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

## 3. Independent events

Two events  $A$  and  $B$  are independent if any of the following are true (all are equivalent):

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

## 4. Random variables

A random variable  $X$  maps outcomes in some sample space to real numbers, i.e.,  $X : \mathcal{S} \rightarrow \mathbb{R}$ . A random variable measures a specific quantitative feature of the sample space outcome.

The **range** of  $X$ , the set of all possible values that  $X$  can take, is denoted  $\mathcal{X}$ .

With the new sample space,  $\mathcal{X}$ , you can use the order relationship in real numbers and you can add, multiply, etc.

## 5. Discrete distributions

### 5.1. Geometric distribution

## 6. Continuous distributions

### 6.1. Normal distribution

todo

### 6.2. Gamma distribution

The gamma distribution is based on the gamma function, which extends the factorial function to complex numbers:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

Some properties of  $\Gamma$ :

- $\Gamma(x) = x!$  if  $x$  is a non-negative integer
- $\Gamma(a+1) = a\Gamma(a)$

If you have a gamma distribution  $\text{Gamma}(\alpha, \lambda)$ , then its probability density function is

$$f_{X(x)} = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Note: there are two different parameterizations for Gamma:

- $\text{Gamma}(\alpha, \beta/\lambda)$  (with a rate parameter)
- $\text{Gamma}(\alpha, \theta)$  (with a scale parameter)

The  $\alpha$  is the shape parameter.  $\beta = \frac{1}{\theta}$

### 6.3. Exponential distribution

This is a special case of the gamma distribution:  $\text{Exponential}(\lambda) = \text{Gamma}(1, \lambda)$ . If you sum  $n$  independent  $\text{Exponential}(\lambda)$  random variables, you get a  $\text{Gamma}(n, \lambda)$  random variable.

Application: Variables that model the amount of time you have to wait before something happens follow an exponential distribution. E.g., time between clicks of a Geiger counter.

This is the continuous analog of the geometric distribution. Only the geometric and exponential distributions are **memoryless**.

### 6.4. Chi-squared distribution

Also a special case of the gamma distribution:  $\chi_k^2 \sim \text{Gamma}(\alpha = \frac{k}{2}, \theta = 2)$ .

Stuff in real life isn't distributed this way. Chi-squared is mostly just used for hypothesis tests. It's closely related to the standard normal distribution ( $Z$ ):  $\chi_1^2 = Z^2$ . In general:

$$\chi_k^2 = (Z_1)^2 + (Z_2)^2 + \dots + (Z_k)^2$$

## 7. Relationships between variables

**Large value** A value of some random variable is large if it's greater than the mean

**Small value** A value of some random variable is small if it's less than the mean

**Positive relationship**  $X$  and  $Y$  have a positive relationship if large values of  $X$  are associated with large values of  $Y$  and small values of  $X$  are associated with small values of  $Y$

**Negative relationship**  $X$  and  $Y$  have a negative relationship if large values of  $X$  are associated with small values of  $Y$  and small values of  $X$  are associated with large values of  $Y$

### 7.1. Covariance

$$\text{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

When  $X$  and  $Y$  have a positive relationship, the covariance should be positive. When they have a negative relationship, the covariance should be negative.

Units of  $\text{Cov}(X, Y)$  are (units of  $X$ )  $\cdot$  (units of  $Y$ ). This is a problem, since the magnitude of the covariance depends on the units.

Properties:

- $\text{Cov}(X, X) = V(X)$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(aX + bY, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$

## 7.2. Correlation coefficient

$$\text{Corr}(X, Y) = \rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

This one is unitless, unlike covariance. It always lies in  $[-1, 1]$ .

$\rho_{X,Y} = \pm 1$  iff there is a perfect linear relationship between  $X$  and  $Y$ .

The correlation coefficient measures the extent of the linear relationship between  $X$  and  $Y$ .

## 8. Linear combinations of random variables

If you have  $Y = a_1X_1 + a_2X_2 + \dots + a_nX_n$ , then  $E(Y) = a_1E(X_1) + \dots + a_nE(X_n)$

Variance is more complicated:

$$V(Y) = \text{Cov}(Y, Y) = \sum_{i=1}^n a_i^2 V(X_i) + 2 \sum_{1 \leq i < j \leq n} a_i a_j \text{Cov}(X_i, X_j)$$

If all the random variables  $X_i$  are independent, then the covariance terms cancel out, meaning that:

$$V(Y) = \sum_{i=1}^n a_i^2 V(X_i)$$

## 9. Random samples and statistics

**Random sample** We say the collection of random variables  $\{X_1, X_2, \dots, X_n\}$  is a **random sample** of size  $n$  from the population distribution  $X$  if:

- The  $X_i$ s are identically distributed to the distribution of  $X$ , i.e.,  $X_i \sim X$
- The  $X_i$ s are mutually independent, i.e., joint pdf/pmf splits into its marginals

**Statistic** A quantity calculated using a random sample.

**Joint sample space** Suppose  $\{X_1, \dots, X_n\}$  is a random sample from population  $X$  and  $X$  takes values in  $\mathcal{X}$ . The joint sample space for the random sample is  $\mathcal{X}^n = \{(x_1, \dots, x_n) \mid x_i \in \mathcal{X}\}$ .

The joint sample space is the set of all possible sample data for the random sample.

If  $T$  is a statistic calculated using the random sample  $\{X_1, \dots, X_n\}$  with joint sample space  $\mathcal{X}^n$ , then we can think of  $T$  as the function

$$T : \mathcal{X}^n \rightarrow \mathbb{R}$$

Therefore,  $T$  is a random variable.

**Sampling distribution** The associated probability distribution of  $T$  above is called the sampling distribution of the statistic  $T$ .

### 9.1. Central Limit Theorem

If you have a random sample of size  $n$  from a distribution  $X$  with  $E(X) = \mu$  and  $V(X) = \sigma^2$ , then, if  $n$  is large enough, the sample mean  $\bar{X}$  will approximately have the distribution  $N\left(\mu, \frac{\sigma^2}{n}\right)$

## 10. Estimators

### 10.1. Parametric estimation

Suppose the actual distribution has parameters  $\theta_1, \theta_2, \dots, \theta_k$ .

We want to estimate these parameters. The estimates are  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$

#### 10.1.1. Point estimators

Suppose a population is fixed and has distribution  $X$  with parameters  $\theta_1, \theta_2, \dots, \theta_k$

**Point estimator** A point estimator for the parameter  $\theta_i$  is a statistic  $\hat{\theta}_i$  calculated using a random sample of size  $n$  coming from the population distribution whose values are used as estimates for  $\theta_i$

**Bias** The bias of a point estimator is the expected deviation of values from  $\hat{\theta}$  from  $\theta$

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta = E(\hat{\theta} - \theta)$$

**Unbiased** A point estimator  $\hat{\theta}$  is unbiased if  $\text{Bias}(\hat{\theta}) = 0$  for all  $\theta$

**Consistent** A point estimator  $\hat{\theta}$  is consistent if  $\lim_{n \rightarrow \infty} \text{MSE}(\hat{\theta}) = 0$

**Very important theorem:**

$$\text{MSE}(\hat{\theta}) = V(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$$

### 10.2. Method of Moments

Given:  $k$  parameters attached to a population distribution.

Goal: given a random sample of size  $n$ , find  $k$  estimators for these parameters.

Intuition: the estimators will show up as solutions to a system of  $k$  equations.

**Population moment** The  $k$ th population moment is defined as  $\mu_k := E(X^k)$

**Sample moment** Given a random sample  $\{X_1, X_2, \dots, X_n\}$  coming from the population distribution  $X$ , the  $k$ th sample moment is defined as  $\frac{1}{n} \sum_{i=1}^n X_i^k$

Intuition: For large enough  $n$ , the  $k$ th sample moments estimate the  $k$ th population moments.

MoM Algorithm:

1. Calculate the first  $k$  population moments
2. Derive a system of  $k$  equations by equating each  $i$ th population moment to the  $i$ th sample moment.
3. Find solutions to the system, if they exist. These are the MoM estimators.