Lecture 3

Sampling, sampling distributions, and parameter estimation

Sampling

Definition

- Population is defined as the collection of all the possible observations of interest.
- The collection of observations we take from the population is called a sample.
- The number of observations in the sample is called the sample size.

Sampling

- When we are interested in a population, we typically study a sample of that population rather than attempt to study the entire population.
- The sample should ideally be a representation of the population with similar characteristics.

Principles of sampling

- 1. Same distribution. All variables in the sample X_1, …, X_n have the same distribution as in the entire population.
- 2. Independence. X_1, …, X_n are independent. In other words, each observation has no relationship with others.

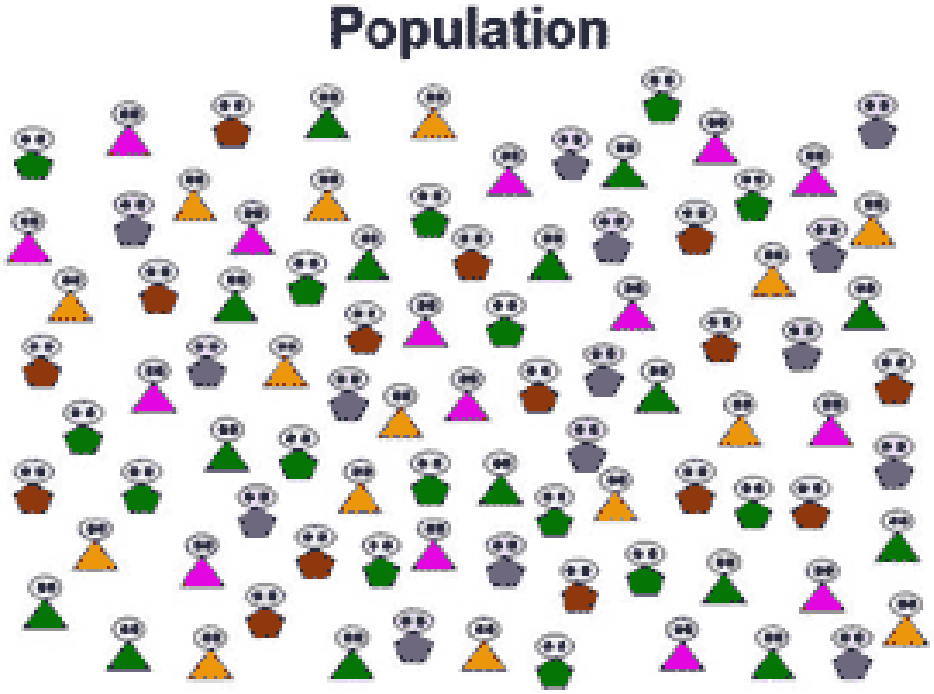
Simple random sampling

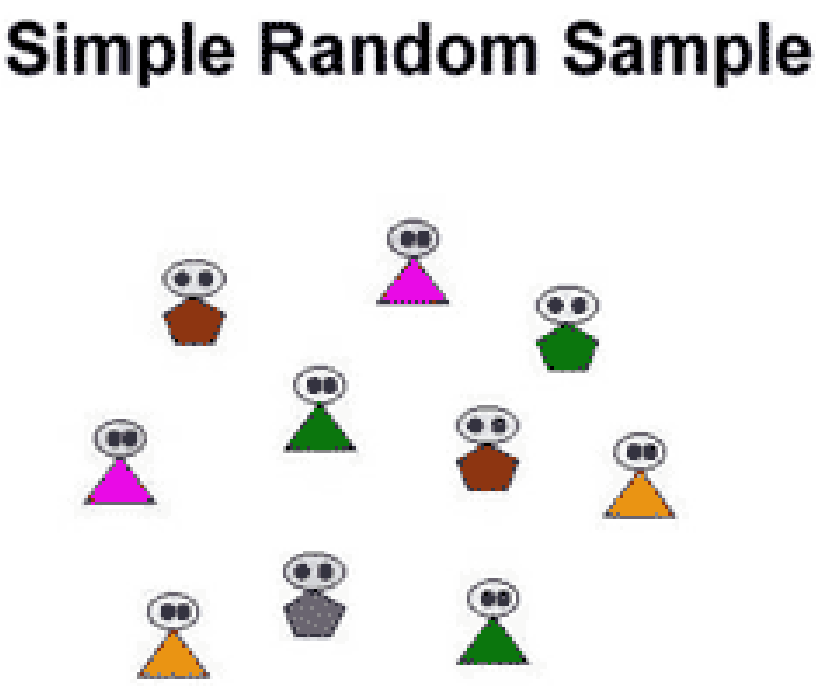
- Simple random sampling is the most straightforward of the random sampling strategies. We use this strategy when we believe that the population is relatively homogeneous for the characteristic of interest. i.e. no population structure

Simple random sampling

- For example, let's say you were surveying first-time parents about their attitudes toward mandatory seat belt laws. You might expect that their status as new parents might lead to similar concerns about safety.
- On campus, those who share a major might also have similar interests and values; we might expect psychology majors to share concerns about access to mental health services on campus.

Simple Random Sampling





Other sampling methods

- Systematic sampling
- Stratified sampling
- Proportionate sampling
- Cluster sampling
- Multistage sampling
- And so on

Sample statistic and distribution

Sample mean and sample variance

- Let $X_1, \ldots, X_n$ be a random sample
- Sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2\right)$$

- Sample standard error (deviation)

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

For frequency or grouping data

- Let X_1, …, X_k be group values, f_1, …, f_k be the frequency of each group,
  $f_1 + \ldots + f_k = 1$
- Sample mean $\overline{X} = \sum_{i=1}^{k} f_i X_i$

- Sample variance

$$S^2 = \sum_{i=1}^{k} f_i \left(X_i - \overline{X} \right)^2 = \sum_{i=1}^{k} f_i X_i^2 - \overline{X}^2$$

Properties of sample mean and variance

- Let $X_1, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$, and let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

- Then we have
- (1) $\overline{X}$ and S^2 are independent random variables.
- (2) $\overline{X}$ has a normal distribution, i.e. $N(\mu, \sigma^2/n)$
- (3) $(n-1)S^2/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom.

$$(n-1)S^2 = SS = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

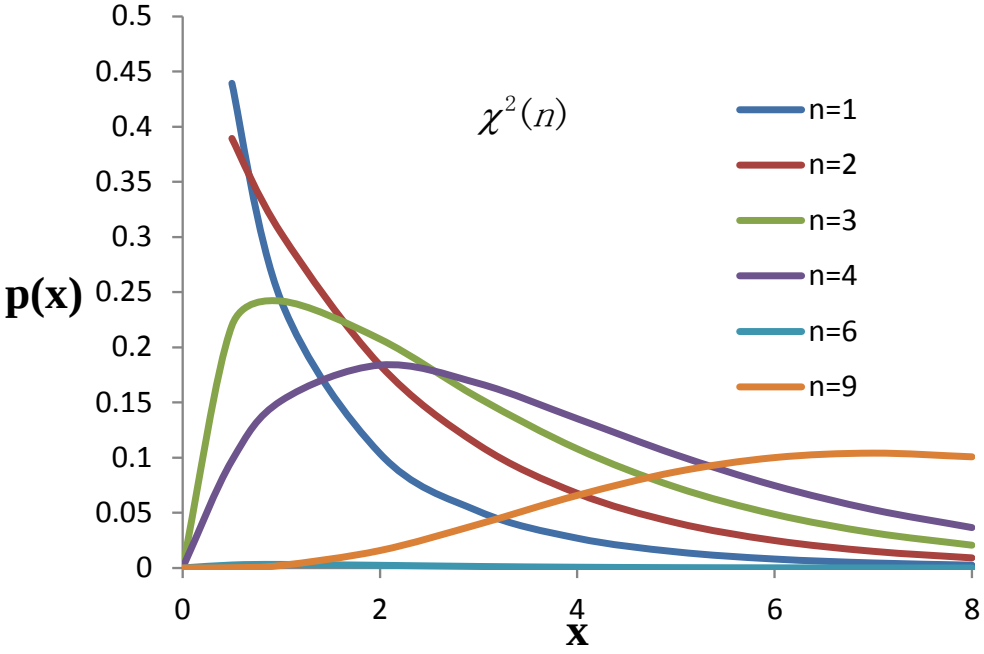
χ^2 distribution

- If $X_i \sim N(0,1)$, i=1,…,n, and X_i are independent

Define $\chi^2 = \sum_{i=1}^{n} X_i^2$

The χ^2 obeys χ^2 distribution with n degrees of freedom, denoted by

$$\chi^2 \sim \chi^2(n)$$



Student's *t* distribution

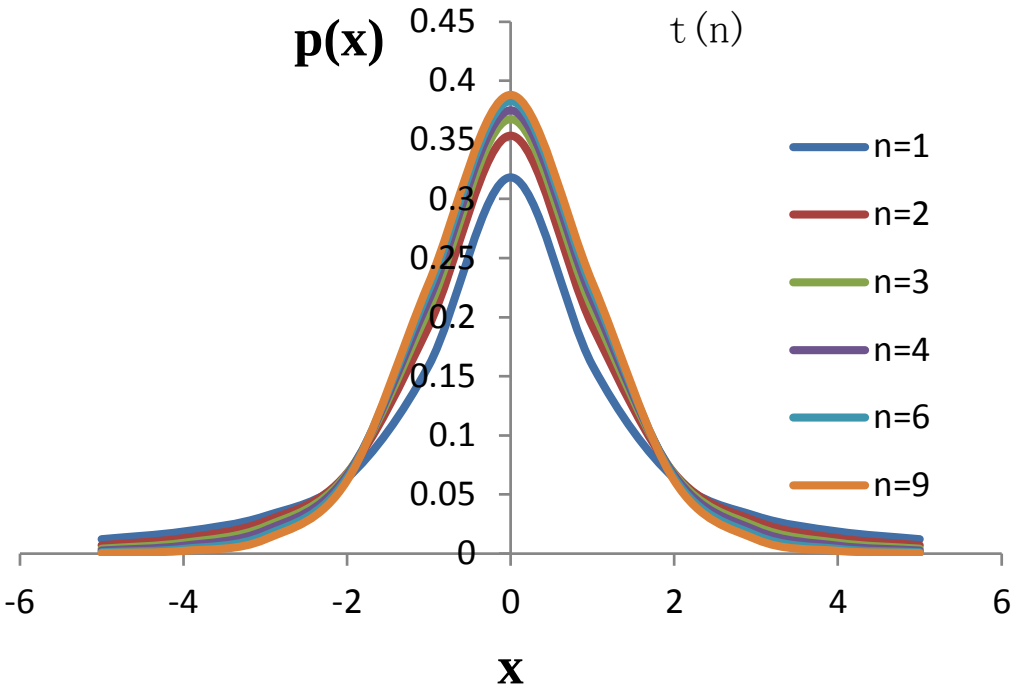
- If X~N(0,1), Y~ $\chi^2(n)$, X and Y are independent

Define

$$t = \frac{X}{\sqrt{Y/n}}$$

Then *t* obeys *t* distribution with n degrees of freedom, denoted by

$$t \sim t(n)$$

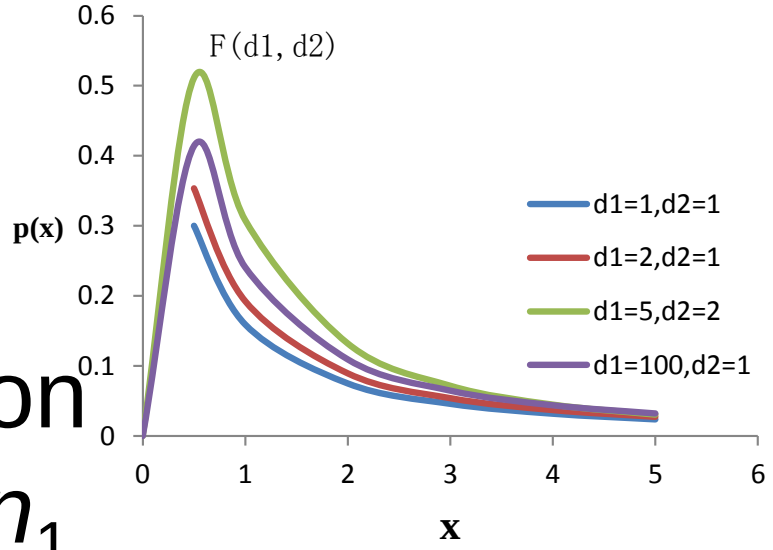


F distribution

- X and Y are independent,

$$X \sim \chi^2(n_1) \quad Y \sim \chi^2(n_2)$$

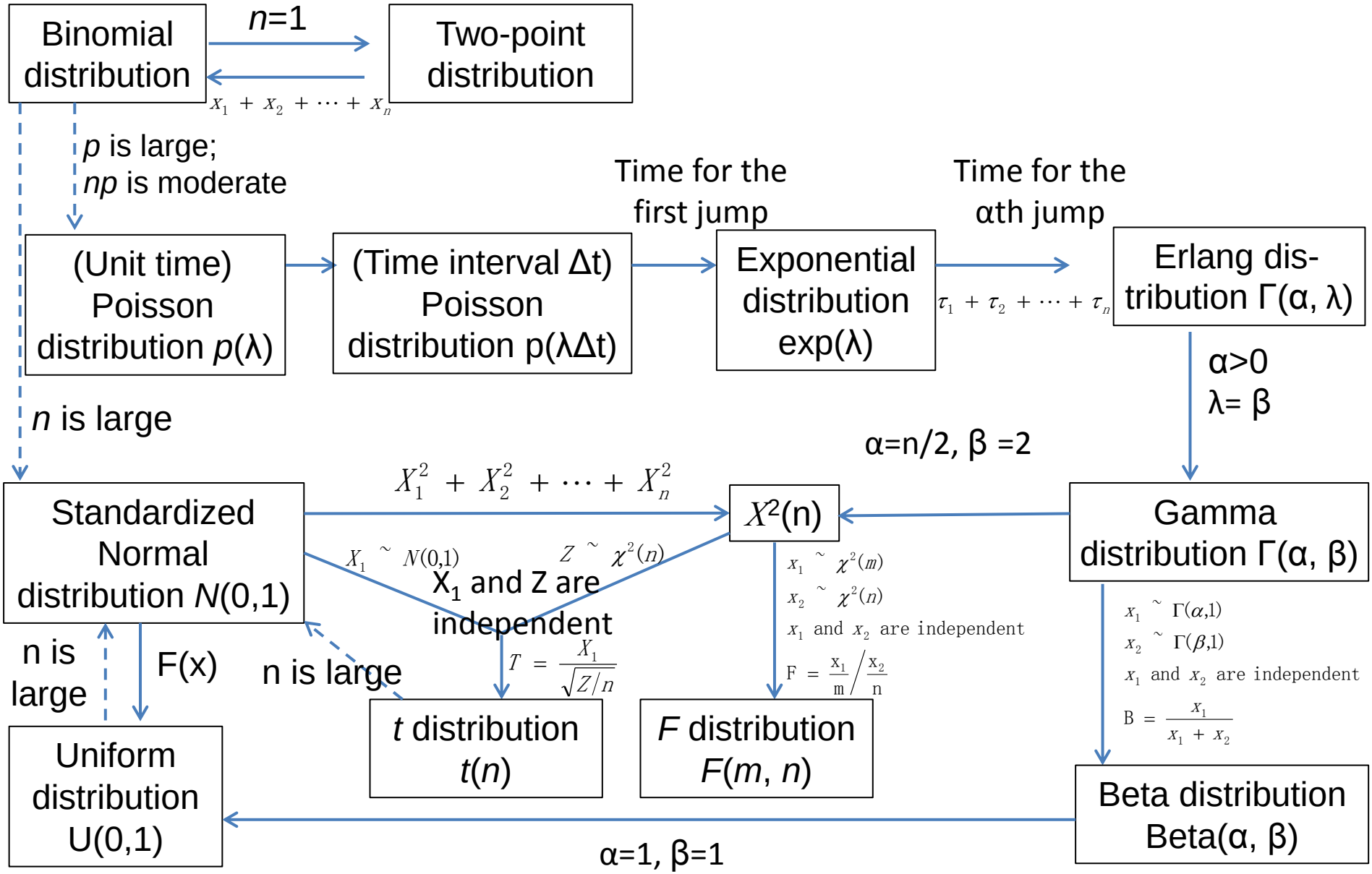
Define $$F = \frac{X/n_1}{Y/n_2}$$



Then *F* obeys *F* distribution with degrees of freedom n_1 and n_2, denoted by

$$F \sim F(n_1, n_2)$$

Relationship between different distributions



Statistical inference

- Statistical inference: Drawing conclusions about the whole population on the basis of a sample

- Precondition for statistical inference: A sample is randomly selected from the population (=probability sample)

Parameter estimation

Parameter estimation

- Parameter estimation is an important problem in statistics. It can divided into two types:
  - 1. Point estimation: it involves the use of sample data to calculate a single value (known as a statistic) which is to serve as a "best guess" or "best estimate" of an unknown (fixed or random) population parameter.
  - 2. Interval estimation: it is the use of sample data to calculate an interval of possible (or probable) values of an unknown population parameter.

Point estimation

- X~F(x, θ), θ is unknown. The target of point estimation is to give a statistic and there is a group of observations X_1, X_2, …, X_n. The estimator of θ denotes as

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$$

- For example, when $\theta = E(X)$, we can use mean of samples as the estimator of θ, i.e.

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Two commonly used point estimation methods

- Maximum likelihood method
- Moment method

Maximum likelihood estimate (MLE)

- This method is to maximize the likelihood function for getting the estimator of parameters.
- The probability density function of X is $p(x; \theta)$, and θ is unknown. Suppose there is a sample observations $X_1, X_2, \ldots, X_n$ for X.

Maximum likelihood method

- Then the combined probability function is

$$L(\theta) = L(\mathrm{x}_1, x_2, \cdots, x_n; \theta) = p(\mathrm{x}_1; \theta)\, p(\mathrm{x}_2; \theta) \cdots p(\mathrm{x}_n; \theta) = \prod_{i=1}^{n} p(\mathrm{x}_i; \theta)$$

- We call the above function the likelihood function. Define the logarithm of likelihood as

$$\ln L(\theta) = \ln L(\mathrm{x}_1, x_2, \cdots, x_n; \theta) = \sum_{i=1}^{n} \ln\, p(x_i; \theta)$$

- Let $\frac{d \ln L(\theta)}{d\theta} = 0$, then we can calculate the maximum likelihood estimator (MLE) of θ.

Maximum likelihood method

- When the likelihood function contains *k* parameters $\theta_1, \theta_2, \cdots, \theta_k$, then

$$L(\theta_1, \theta_2, \cdots, \theta_k) = \prod_{i=1}^{n} \mathrm{p}(\mathrm{x}_i; \theta_1, \theta_2, \cdots, \theta_k)$$

- The maximum likelihood estimator of $\theta_1, \theta_2, \cdots, \theta_k$: $\hat{\theta}_i = \hat{\theta}_i\left(X_1, X_2, \cdots, X_n\right)$, i=1,…,k are the solution of *k* equations

$$\frac{\partial \ln L(\theta_1, \theta_2, \cdots, \theta_k)}{\partial \theta_i} = 0, i = 1, 2, \cdots, k$$

Example

- Assume X_1 , X_2 , … , X_n are random samples from a normal distribution $N(\mu,\ \sigma^2)$, how to get the maximum likelihood estimator of parameters μ and σ^2 .
- Solution: The likelihood function is

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{(x_i-\mu)^2}{2\sigma^2}]$$

$$=\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

Example

- Then

$$\ln L(\mu,\ \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

- Then the derivative equations are

$$\begin{cases} \dfrac{\partial}{\partial\mu}\ln L(\mu,\ \sigma^2) = \dfrac{1}{\sigma^2}\displaystyle\sum_{1}^{n}(x_i-\mu) = 0 \\ \dfrac{\partial}{\partial\sigma^2}\ln L(\mu,\ \sigma^2) = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{1}^{n}(x_i-\mu)^2 = 0 \end{cases}$$

Example

- So the solutions are

$$\begin{cases} \mu = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \\ \sigma^2 = \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2 \end{cases}$$

- The maximum likelihood estimator of μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Example

- When μ is known, MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

- If μ is unknown, MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- MLE of σ^2 is not equal to the sample variance! Which one is better?

Moment method

- Basic idea: equating sample moments with unobservable population moments and then solving those equations for the quantities to be estimated.
- Suppose the probability density function of X is $p(x;\ \theta_1, \theta_2, \cdots, \theta_k)$, then the r th moment of X is

$$\upsilon_r = E(X^r) = \int_{-\infty}^{+\infty} x^r\, p(x;\theta_1, \theta_2, \cdots, \theta_k)dx$$

Moment method

- Suppose there is a sample observations $X_1, X_2, \ldots, X_n$ for X. Then the *r* th moment of samples are

$$a_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

- Equate the *j* th (*j*=1, …, *k*) sample moments with unobservable population moments

$$\begin{cases} \upsilon_1(\theta_1, \theta_2, \cdots, \theta_k) = a_1 \\ \upsilon_2(\theta_1, \theta_2, \cdots, \theta_k) = a_2 \\ \cdots \\ \upsilon_k(\theta_1, \theta_2, \cdots, \theta_k) = a_k \end{cases}$$

Moment method

- Solve the equations, then we can get the estimator of θ: $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k$
- We call them the moment estimators of θ.

Example

- $X_1, X_2, \ldots, X_n$ are samples from Uniform distribution

$$p(\mathrm{x}, \theta) = \begin{cases} \frac{1}{\theta}, 0 < x \le \theta \\ 0, \text{otherwise} \end{cases}$$

- Then $$\upsilon_1 = \mu = \int_{-\infty}^{+\infty} x p(x, \theta) dx = \frac{1}{\theta} \int_0^{\theta} x dx = \frac{\theta}{2}$$

$$a_1 = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- So $$\frac{\theta}{2} = \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$
- The moment estimator of θ is $\hat{\theta} = 2\bar{x}$

Desirable properties of estimator

- Since estimator gives an estimate that depends on sample points (X_1, X_2,…, X_n), estimate is a function of sample points. Sample points are random variable, therefore estimate is random variable and has probability distribution.
- We want that estimator to have several desirable properties like
- 1. Unbiasedness
- 2. Effectiveness
- 3. Minimum mean square error

Unbiasedness

- An estimator is said to be unbiased if the expected value of the estimator is equal to true value of the parameter being estimated, or

$$E(\hat{\theta}) = \theta$$

- Example: sample proportion is the unbiased estimator of population proportion

Effectiveness

- The most efficient estimator among a group of unbiased estimators is the one with the smallest variance.
- Generally speaking, assuming $\hat{\theta}_1(\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_n)$ and $\hat{\theta}_2(\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_n)$ are two unbiased estimators of θ, and $V(\hat{\theta}_1) \le V(\hat{\theta}_2)$, then $\hat{\theta}_1$ is said to be more effective than $\hat{\theta}_2$.

Minimum mean square error (MSE)

- Basic idea: minimize the average deviation between the estimation and true value.
- We call the estimator which minimize

$$E\{[\hat{\theta}(\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_n) - \theta]^2\}$$

as the minimum mean square error estimator of θ.

Interval estimation

- Estimation of the parameter is not sufficient. It is necessary to analyze and see how confident we can be about this particular estimation.
- One way of doing it is defining confidence intervals. If we have estimated θ we want to know if the “true” parameter is close to our estimate. In other words we want to find an interval that satisfies following relation:

$$P(G_L < \theta < G_U) \geq 1 - \alpha$$

Interval estimation

- i.e. probability that “true” parameter θ is in the interval (G_L, G_U) is greater than 1-α.
- Actual realization of this interval - (g_L, g_U) is called a 100(1- α)% of confidence interval, limits of the interval are called lower and upper confidence limits. 1- α is called confidence level.

Example

- If population variance is known (σ^2) and we estimate population mean then

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

- We can find from the table that probability of Z is more than 1 is equal to 0.1587. Probability of Z is less than -1 is again 0.1587. These values comes from the table of the standard normal distribution.

Example

- Now we can find confidence interval for the sample mean. Since:

$$P(-1 < Z < 1) = P(Z < 1) - P(Z < -1) = 1 - P(Z > 1) - P(Z < -1)$$

$$= 1 - 2 * 0.1587 = 0.6826$$

- Then for μ we can write

$$P(-1 < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < 1) = P(\bar{x} - \sigma / \sqrt{n} < \mu < \bar{x} + \sigma / \sqrt{n}) = 0.6826$$

- Confidence level that “true” value is within 1 standard error (standard deviation of sampling distribution) from the sample mean is 0.6826. Probability that “true” value is within 2 standard error from the sample mean is 0.9545.

Interval estimation

- Above we considered the case when population variance is known in advance. It is rarely the case in real life. When both population mean and variance are unknown we can still find confidence intervals. In this case we calculate population mean and variance and then consider distribution of the statistic:

$$Z = \frac{\bar{x} - \mu}{S / \sqrt{n}}$$

- Here S^2 is the sample variance.

Interval estimation

- Since it is the ratio of the standard normal random variable to square root of χ^2 random variable with n-1 degrees of freedom, Z has Student's t distribution with n-1 degrees of freedom. In this case we can use table of t distribution to find confidence levels.
- It is not surprising that when we do not know sample variance confidence intervals for the same confidence levels becomes larger. That is price we pay for what we do not know.

Interval estimation

- If number of degrees of freedom becomes large, then *t* distribution is approximated well with normal distribution. For $n>100$ we can use normal distribution to find confidence levels, intervals.

The Law of Large Numbers and Central Limit Theorem

The Law of Large Numbers

- Assume X_1 , X_2 , … , X_n are random samples of X. $E(X)=\mu$ and *V*(X) exist. Let $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$, then for any given $\varepsilon>0$,

$$\lim_{n\to\infty} P\left\{\left|\overline{X}-\mu\right|<\varepsilon\right\}=1$$

The Central Limit Theorem

Let $\overline{X}$ be the mean of a random sample $X_1, X_2, \ldots, X_n$, of size *n* from a distribution with a finite mean μ and a finite positive variance σ^2. Then

$$Y = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \rightarrow N(0,1)$$

Small probability event

- A **small probability event** is an event that has a low probability of occurring.
- The small probability event will hardly happen in one experiment. This principle is used for hypothesis and tests.
- An event is a small probability event, so it will hardly happen in theory. But if it happens actually, then we reject H_0.

Experiments on the distribution of sample mean and sample variance

- Use RAND() in EXCEL to generate pseudo-random numbers X_1 and X_2 of U(0,1): uniform distribution on the interval [0, 1]
- Use transformation to generate random numbers Y_1 and Y_2 of N(0, 1)

$$Y_1 = \sqrt{-2\ln(X_1)}\sin(2\pi X_2), Y_2 = \sqrt{-2\ln(X_1)}\cos(2\pi X_2)$$

- Use transformation to generate random numbers Z_1 and Z_2 of N(μ, σ^2)

$$Z = \sigma Y + \mu$$

Let's do some exercises together

- Draw 100 random samples from U(0, 1)
- Draw the frequency distribution of the 100 samples
- Draw 100 sets of 5 samples from N(10, 10)
- Draw the frequency distribution of the 500 samples, and compare it with N(10, 10)
- Draw the frequency distribution of the 100 sample means and 100 sample variances
- Compare the distribution of $\bar{X}$ with N(10, 2)
- Compare the distribution of $4S^2/10$ with $\chi^2(4)$