

# Estimation of parameters

## Fitting probability laws to data

Many families of probability laws depend on a small number of parameters:

- Poisson family:  $\lambda$ , Gaussian family:  $\mu, \sigma \dots$
- Parameters are estimated from the data
- Need measures and tests of goodness of fit (later)

*Classical example.* Fitting Poisson Distribution to Emissions of  $\alpha$  Particles

*If the underlying rate of emission is constant over the period of observation (half-life  $\gg$  time period of observation), if the particles come from a very large # of independent sources (atoms), the Poisson distribution is frequently used as a model for radioactive decay.*

- (1) the underlying rate at which the events occur is constant in space or time,
- (2) events in disjoint intervals of space or time occur independently,
- (3) there are no multiple events.

## Fitting probability laws to data

Berkson (1966) & National Bureau of Standards.

Source of  $\alpha$  particles: americium (镅/95号) 241.

The experimenters recorded 10,220 times between successive emissions.

Observed mean emission rate: .8392 emissions per sec.

observed in 1207 intervals, each of length 10 sec.

$n$	Observed	$\frac{1207 p_k}{1207}$ Expected
0-2	18	12.2
3	28	27.0
4	56	56.5
5	105	94.9
6	126	132.7
7	146	159.1
8	164	166.9
9	161	155.6
10	123	130.6
11	101	99.7
12	74	69.7
13	53	45.0
14	23	27.0
15	15	15.1
16	9	7.9
17+	5	7.1
	1207	1207

View the 1207 counts as 1207 independent realizations of Poisson random variables, probability mass function

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

Average count in a 10s interval = 8.392  $\rightarrow \hat{\lambda}$

1. Estimate of  $\lambda$  is a random variable w. sampling distribution!
2. To what decimal place is 8.392 accurate? Need knowledge on sampling distribution

Qualitatively good fit!

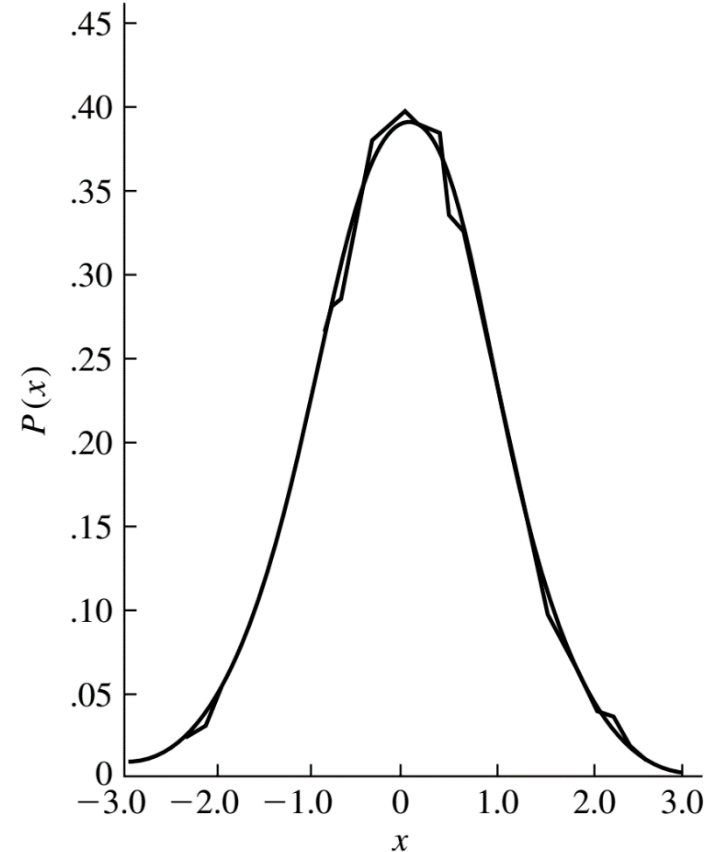
Quantitative measures later.

## Parameter estimation

*Example A: Gaussian.* (Usually justified using some version of CLT)

Bevan, Kullberg, and Rice (1979): 肌肉细胞膜中流体的随机波动性。细胞膜包含大量通道，随机打开和关闭，假定彼此间均独立运行。净流量为打开的通道通过的离子之和，即近似独立流之和。

A smoothed histogram of values obtained from 49,152 observations of the net current and an approximating Gaussian curve. Estimated parameters  $\mu$  and  $\sigma^2$  are used to extract useful microscopic information.



$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty$$

## Parameter estimation

### Example B: Gamma.

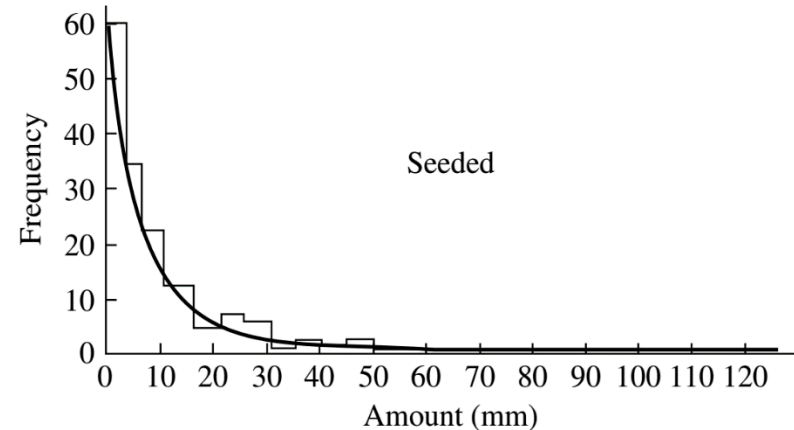
The family of gamma distributions provides a flexible set of densities for nonnegative random variables. Two parameters:

$$f(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}, \quad 0 \leq x \leq \infty$$

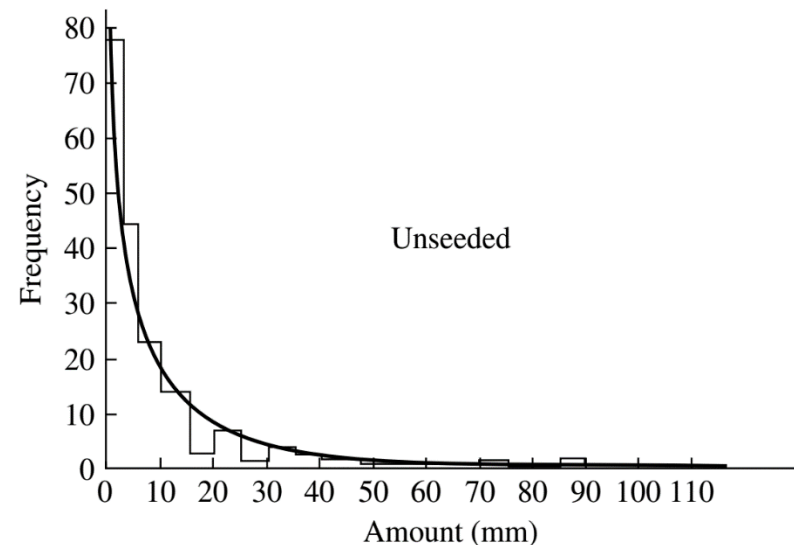
Different reasons of fitting to data:

- Scientific theory suggests a distribution of direct interest (e.g.  $\alpha$  particle emissions).
- Descriptive purposes, a method of data summary/compression (e.g. seeded storms)
- More complex modeling...(e.g. 水利专家规划水资源调度和使用方式时，采用随机模型模拟降水量)

Le Cam & Neyman (1967):  
人工降雨的效果



(a)



(b)

## Parameter estimation

**Basic approach:** The observed data are regarded as realizations of random variables  $X_1, X_2, \dots, X_n$ , whose joint distribution depends on unknown  $\theta$  (can be a vector).

**i.i.d.: independent & identically distributed (独立同分布)**

Usually the  $X_i$  is modeled as independent random variables all having the same distribution  $f(x|\theta)$ , their joint distribution is  $f(x_1|\theta)f(x_2|\theta) \cdots f(x_n|\theta)$ .

An estimate of  $\theta$  is a function of  $X_1, X_2, \dots, X_n$ , a random variable with a *sampling distribution*. We use approximations to the sampling distribution to assess the variability of the estimate (generally through its *standard error*).

## Parameter estimation

**Basic approach:** The observed data are regarded as realizations of random variables  $X_1, X_2, \dots, X_n$ , whose joint distribution depends on unknown  $\theta$  (can be a vector).

### **i.i.d.: independent & identically distributed (独立同分布)**

Usually the  $X_i$  is modeled as independent random variables all having the same distribution  $f(x|\theta)$ , their joint distribution is  $f(x_1|\theta)f(x_2|\theta) \cdots f(x_n|\theta)$ .

An estimate of  $\theta$  is a function of  $X_1, X_2, \dots, X_n$ , a random variable with a *sampling distribution*. We use approximations to the sampling distribution to assess the variability of the estimate (generally through its *standard error*).

General procedures for forming estimates so that each new problem does not have to be approached ***ab initio*** (从第一原理出发) .

- The method of moments
- The method of maximum likelihood (more useful in general)

## The method of moments

$$\mu_k = E(X^k)$$

If  $X_1, X_2, \dots, X_n$ , are i.i.d. random variables from a distribution, the **sample moment** can be viewed as an estimate of  $\mu_k$ .

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

Two steps:

1. Find expressions for them in terms of the lowest possible order moments,
2. Substitute sample moments into the expressions.

Suppose we wish to estimate  $\theta_1$  &  $\theta_2$ , if they can be expressed in terms of first 2 moments:

$$\theta_1 = f_1(\mu_1, \mu_2)$$

$$\theta_2 = f_2(\mu_1, \mu_2)$$

The method of moment estimates are

$$\hat{\theta}_1 = f_1(\hat{\mu}_1, \hat{\mu}_2)$$

$$\hat{\theta}_2 = f_2(\hat{\mu}_1, \hat{\mu}_2)$$



## The method of moments

The construction of a method of moments estimate involves three basic steps:

1. Calculate low order moments, finding expressions for the moments in terms of the parameters. Typically, the number of low order moments needed will be the same as the number of parameters.
2. Invert the expressions found in the preceding step, finding new expressions for the parameters in terms of the moments.
3. Insert the sample moments into the expressions obtained in the second step, thus obtaining estimates of the parameters in terms of the sample moments.

## The method of moments

### Example A. Gaussian.

The 1<sup>st</sup> and 2<sup>nd</sup> moments are

$$\mu_1 = E(X) = \mu$$

$$\mu_2 = E(X^2) = \mu^2 + \sigma^2$$

Therefore,

$$\mu = \mu_1$$

$$\sigma^2 = \mu_2 - \mu_1^2$$

The corresponding estimates of  $\mu$  and  $\sigma^2$  from the sample moments are

$$\hat{\mu} = \bar{X}$$

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

The sampling distribution of  $\bar{X}$  is  $N(\mu, \sigma^2/n)$ ,  $n\hat{\sigma}^2/\sigma^2 \sim \chi_{n-1}^2$ .



独立正态随机变量的组合还是正态



述而不证

## The method of moments

### Example B. Gamma.

The 1<sup>st</sup> and 2<sup>nd</sup> moments are

$$\mu_1 = \frac{\alpha}{\lambda}$$
$$\mu_2 = \frac{\alpha(\alpha + 1)}{\lambda^2}$$

Recall: we used its mgf

Express  $\alpha$ ,  $\lambda$ ,

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2}$$
$$\alpha = \lambda\mu_1 = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$$

Since  $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$ ,

$$\hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2} \quad \hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2}$$

In the previous precipitation (降水) problem (Le Cam & Neyman 1967),

$$\bar{X} = .224 \quad \hat{\sigma}^2 = .1338$$

and therefore  $\hat{\alpha} = .375$  and  $\hat{\lambda} = 1.674$ .

## Monte Carlo simulation (bootstrap)

In general, it is difficult to derive the exact forms of the sampling distributions of  $\hat{\alpha}$  and  $\hat{\lambda}$ , because they are each rather complicated functions of  $X_1, X_2, \dots, X_n$ .  
But *this problem can be approached by simulation!*

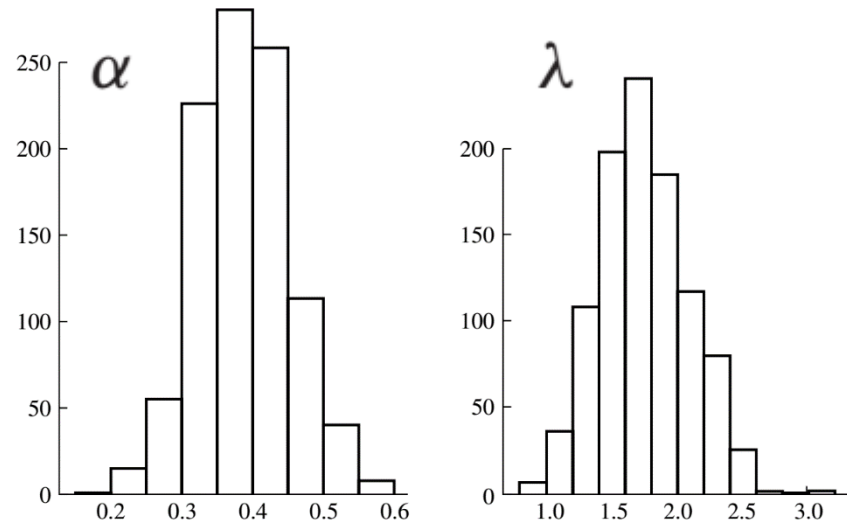
Imagine that we know the true values  $\lambda_0$  and  $\alpha_0$ . We can generate many, many samples of size  $n = 227$  from the gamma distribution with these parameter values, from each of which we can calculate estimates of  $\lambda$  and  $\alpha$ .

A histogram of the values of the estimates of  $\lambda$ , for example, should then give us a good idea of the sampling distribution of  $\hat{\lambda}$ . The only problem is we don't know the true values. Substitute our estimates of  $\lambda$  and  $\alpha$  for the true values.

Calculate the standard deviations of 1000 estimates to obtain estimated Standard errors of  $\hat{\alpha}$  and  $\hat{\lambda}$ .

$$s_{\hat{\alpha}} = \sqrt{\frac{1}{1000} \sum_{i=1}^{1000} (\alpha_i^* - \bar{\alpha})^2}$$

$$s_{\hat{\alpha}} = .06 \text{ and } s_{\hat{\lambda}} = .34.$$



## The method of moments: consistency (相合估计)

An estimate,  $\hat{\theta}$ , is said to be a **consistent** estimate of a parameter  $\theta$ :  
-- if  $\hat{\theta}$  approaches  $\theta$  as the sample size approaches infinity.

### DEFINITION

Let  $\hat{\theta}_n$  be an estimate of a parameter  $\theta$  based on a sample of size  $n$ . Then  $\hat{\theta}_n$  is said to be consistent in probability if  $\hat{\theta}_n$  converges in probability to  $\theta$  as  $n$  approaches infinity; that is, for any  $\epsilon > 0$ ,

$$P(|\hat{\theta}_n - \theta| > \epsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \blacksquare$$

Recall: *the weak law of large numbers implies that the sample moments converge in probability to the population moments.*

If the functions relating the estimates to the sample moments are continuous, the estimates will converge to the parameters as the sample moments converge to the population moments.

**The method of moments estimates is consistent!**

## A brief summary

1. We have shown how the method of moments can provide estimates of the parameters of a probability distribution based on a “sample” (an i.i.d. collection) of random variables from that distribution.
2. Variability or reliability of the estimates? We observe that if the sample is random, the parameter estimates are random variables having their sampling distributions. The standard deviation of the sampling distribution is called the *standard error of the estimate*.
3. How to assess the variability of an estimate from the sample itself? Sometimes the sampling distribution is of an explicit form depending on the unknown parameters, we can substitute our estimates for the unknown parameters in order to approximate the sampling distribution. In other cases the form of the sampling distribution is not obvious, but we can simulate it.
4. By using the **bootstrap** we avoid doing difficult analytic calculations by instructing a computer to generate random numbers.

## The method of moments

The construction of a method of moments estimate involves three basic steps:

1. Calculate low order moments, finding expressions for the moments in terms of the parameters. Typically, the number of low order moments needed will be the same as the number of parameters.
2. Invert the expressions found in the preceding step, finding new expressions for the parameters in terms of the moments.
3. Insert the sample moments into the expressions obtained in the second step, thus obtaining estimates of the parameters in terms of the sample moments.

## The method of moments

### Example A. Gaussian.

The 1<sup>st</sup> and 2<sup>nd</sup> moments are

$$\mu_1 = E(X) = \mu$$

$$\mu_2 = E(X^2) = \mu^2 + \sigma^2$$

Therefore,

$$\mu = \mu_1$$

$$\sigma^2 = \mu_2 - \mu_1^2$$

The corresponding estimates of  $\mu$  and  $\sigma^2$  from the sample moments are

$$\hat{\mu} = \bar{X}$$

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

The sampling distribution of  $\bar{X}$  is  $N(\mu, \sigma^2/n)$ ,  $n\hat{\sigma}^2/\sigma^2 \sim \chi_{n-1}^2$ .



独立正态随机变量的组合还是正态



述而不证



## The method of moments

### Example B. Gamma.

The 1<sup>st</sup> and 2<sup>nd</sup> moments are

$$\mu_1 = \frac{\alpha}{\lambda}$$
$$\mu_2 = \frac{\alpha(\alpha + 1)}{\lambda^2}$$

Recall: we used its mgf

Express  $\alpha$ ,  $\lambda$ ,

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2}$$

$$\alpha = \lambda\mu_1 = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$$

Since  $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$ ,

$$\hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2} \quad \hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2}$$

In the previous precipitation (降水) problem (Le Cam & Neyman 1967),

$$\bar{X} = .224 \quad \hat{\sigma}^2 = .1338$$

and therefore  $\hat{\alpha} = .375$  and  $\hat{\lambda} = 1.674$ .

## Monte Carlo simulation (bootstrap)

In general, it is difficult to derive the exact forms of the sampling distributions of  $\hat{\alpha}$  and  $\hat{\lambda}$ , because they are each rather complicated functions of  $X_1, X_2, \dots, X_n$ .  
But *this problem can be approached by simulation!*

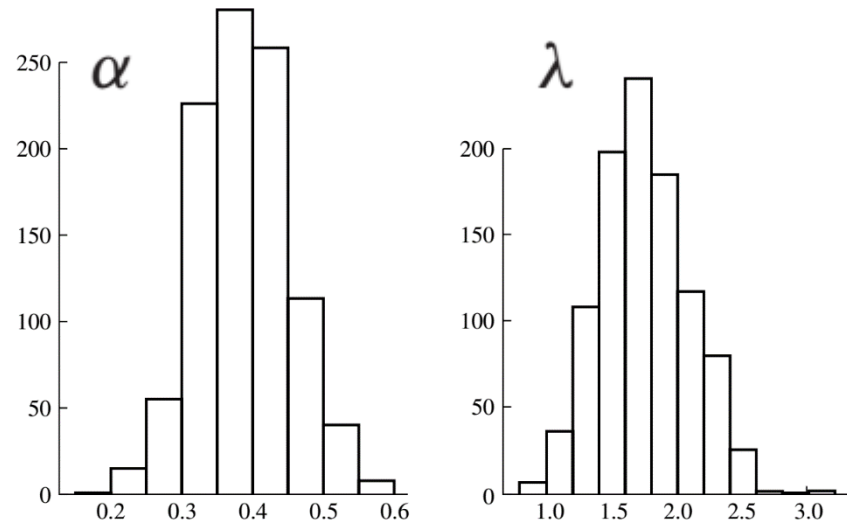
Imagine that we know the true values  $\lambda_0$  and  $\alpha_0$ . We can generate many, many samples of size  $n = 227$  from the gamma distribution with these parameter values, from each of which we can calculate estimates of  $\lambda$  and  $\alpha$ .

A histogram of the values of the estimates of  $\lambda$ , for example, should then give us a good idea of the sampling distribution of  $\hat{\lambda}$ . The only problem is we don't know the true values. Substitute our estimates of  $\lambda$  and  $\alpha$  for the true values.

Calculate the standard deviations of 1000 estimates to obtain estimated Standard errors of  $\hat{\alpha}$  and  $\hat{\lambda}$ .

$$s_{\hat{\alpha}} = \sqrt{\frac{1}{1000} \sum_{i=1}^{1000} (\alpha_i^* - \bar{\alpha})^2}$$

$$s_{\hat{\alpha}} = .06 \text{ and } s_{\hat{\lambda}} = .34.$$



## The method of moments: consistency (相合估计)

An estimate,  $\hat{\theta}$ , is said to be a **consistent** estimate of a parameter  $\theta$ :  
-- if  $\hat{\theta}$  approaches  $\theta$  as the sample size approaches infinity.

### DEFINITION

Let  $\hat{\theta}_n$  be an estimate of a parameter  $\theta$  based on a sample of size  $n$ . Then  $\hat{\theta}_n$  is said to be consistent in probability if  $\hat{\theta}_n$  converges in probability to  $\theta$  as  $n$  approaches infinity; that is, for any  $\epsilon > 0$ ,

$$P(|\hat{\theta}_n - \theta| > \epsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \blacksquare$$

Recall: *the weak law of large numbers implies that the sample moments converge in probability to the population moments.*

If the functions relating the estimates to the sample moments are continuous, the estimates will converge to the parameters as the sample moments converge to the population moments.

**The method of moments estimates is consistent!**

## A brief summary

1. We have shown how the method of moments can provide estimates of the parameters of a probability distribution based on a “sample” (an i.i.d. collection) of random variables from that distribution.
2. Variability or reliability of the estimates? We observe that if the sample is random, the parameter estimates are random variables having their sampling distributions. The standard deviation of the sampling distribution is called the *standard error of the estimate*.
3. How to assess the variability of an estimate from the sample itself? Sometimes the sampling distribution is of an explicit form depending on the unknown parameters, we can substitute our estimates for the unknown parameters in order to approximate the sampling distribution. In other cases the form of the sampling distribution is not obvious, but we can simulate it.
4. By using the **bootstrap** we avoid doing difficult analytic calculations by instructing a computer to generate random numbers.

# Maximum likelihood estimation

# The method of maximum likelihood

...is applied to a great variety of statistical problems (e.g. curving fitting, CMB analysis, galaxy formation modeling...)

- Mathematicians like it because of its nice theoretical properties

Suppose that random variables  $X_1, \dots, X_n$  have a joint density or frequency function  $f(x_1, x_2, \dots, x_n|\theta)$ . Given observed values  $X_i = x_i$ , where  $i = 1, \dots, n$ , the likelihood of  $\theta$  as a function of  $x_1, x_2, \dots, x_n$  is defined as

$$\text{lik}(\theta) = f(x_1, x_2, \dots, x_n|\theta)$$

Note: We consider the joint density as a function of  $\theta$  rather than a function of  $x_i$ !!!

The likelihood function gives the probability of observing the given data as a function of the parameter  $\theta$ .

The **maximum likelihood estimate (mle)** of  $\theta$  is the  $\theta$  value that maximizes the likelihood, making the observed data “most probable” or “most likely.”

## *What you may have learned before...*

If the  $X_i$  are i.i.d., their joint density = product of marginal densities,

$$\text{lik}(\theta) = \prod_{i=1}^n f(X_i|\theta)$$

Usually maximizing its logarithm is easier (equivalent, as log is monotonic).

If i.i.d., the **log likelihood** is

$$l(\theta) = \sum_{i=1}^n \log[f(X_i|\theta)]$$

Here “log” means “ln”

We simply need to find the maximum of log likelihood

– You all have learned techniques in calculus on single variable or multiple variables. It’s time to fully utilize them!

## MLE

*Example:* Poisson.

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

If  $X_1, \dots, X_n$  are i.i.d. and Poisson, joint frequency function = product of marginals

The log likelihood is

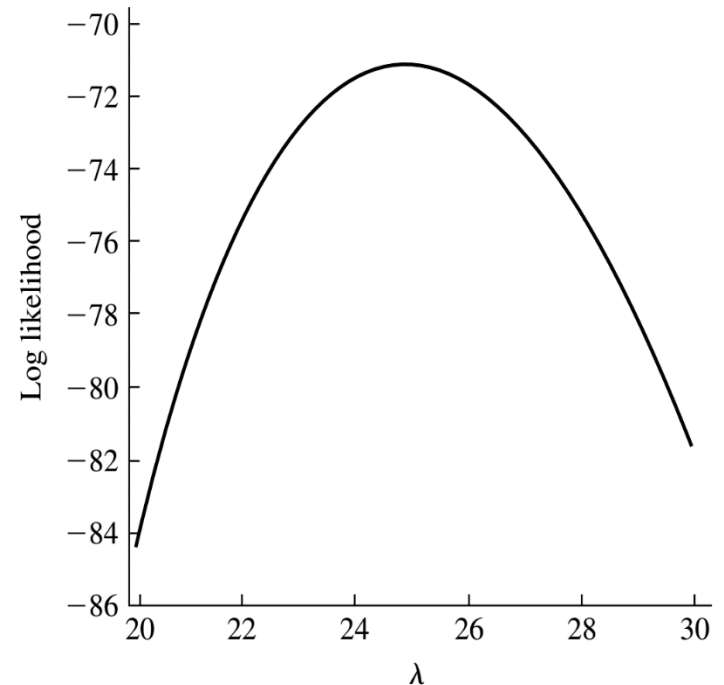
$$\begin{aligned} l(\lambda) &= \sum_{i=1}^n (X_i \log \lambda - \lambda - \log X_i!) \\ &= \log \lambda \sum_{i=1}^n X_i - n\lambda - \sum_{i=1}^n \log X_i! \end{aligned}$$

Setting 1<sup>st</sup> derivative to 0,

$$l'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^n X_i - n = 0$$

$$\hat{\lambda} = \bar{X}$$

-- Agree with the method of moments.





## MLE

*Example:* Gaussian.

If  $X_1, \dots, X_n$  are i.i.d. and Gaussian, joint frequency function = product of marginals.

$$f(x_1, x_2, \dots, x_n | \mu, \sigma) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left[\frac{x_i - \mu}{\sigma}\right]^2\right)$$

Regarded as a function of  $\mu, \sigma$ , this is the likelihood function.

$$l(\mu, \sigma) = -n \log \sigma - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

The partials w.r.t.  $\mu$  and  $\sigma$  are

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^n (X_i - \mu)^2$$

Setting them to 0,

$$\hat{\mu} = \bar{X} \quad \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

-- Same as the method of moments.

**Example: Gamma.**  $f(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}, \quad 0 \leq x < \infty$

The likelihood of an i.i.d. sample,  $X_1, \dots, X_n$ , is

$$l(\alpha, \lambda) = \sum_{i=1}^n [\alpha \log \lambda + (\alpha - 1) \log X_i - \lambda X_i - \log \Gamma(\alpha)]$$

$$= n\alpha \log \lambda + (\alpha - 1) \sum_{i=1}^n \log X_i - \lambda \sum_{i=1}^n X_i - n \log \Gamma(\alpha)$$

$$\frac{\partial l}{\partial \alpha} = n \log \lambda + \sum_{i=1}^n \log X_i - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0 \quad \frac{\partial l}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^n X_i = 0$$

$$n \log \hat{\alpha} - n \log \bar{X} + \sum_{i=1}^n \log X_i - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0 \quad \hat{\lambda} = \frac{n\hat{\alpha}}{\sum_{i=1}^n X_i} = \frac{\hat{\alpha}}{\bar{X}}$$

Solving it numerically, one finds it to be

-- *Different from the method of moments:*

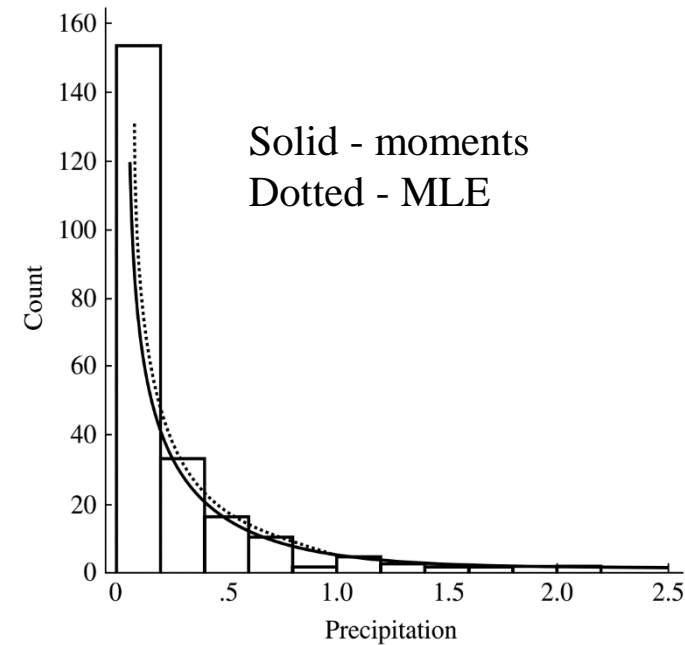
*Example: Gamma.*

Solving it numerically, one finds  $\hat{\alpha} = .441$   $\hat{\lambda} = 1.96$

The method of moments gives  $\hat{\alpha} = .375$   $\hat{\lambda} = 1.674$

Not in closed form, but we can do **bootstrap!**

Generate 1000 samples  $n = 227$  of gamma distributed random variables with  $\alpha = .441$  and  $\lambda = 1.96$ . For each of sample, the MLEs of  $\alpha$  and  $\lambda$  are calculated.



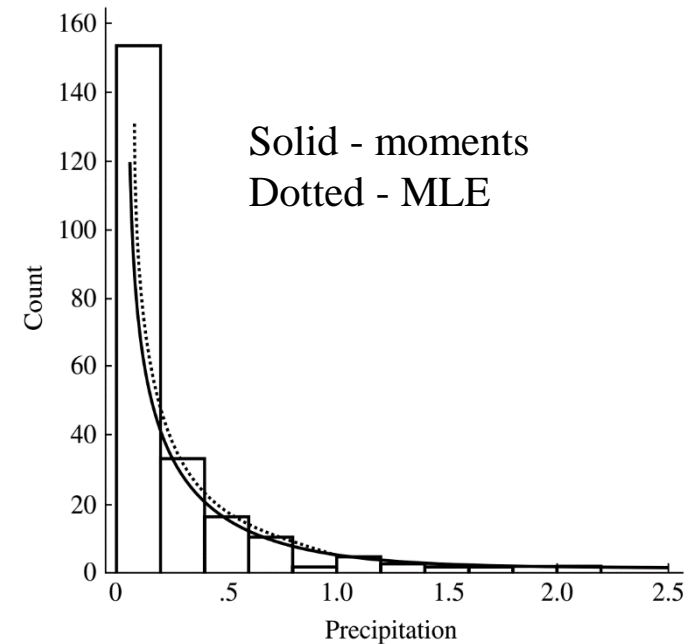
## Example: Gamma.

Solving it numerically, one finds  $\hat{\alpha} = .441$   $\hat{\lambda} = 1.96$

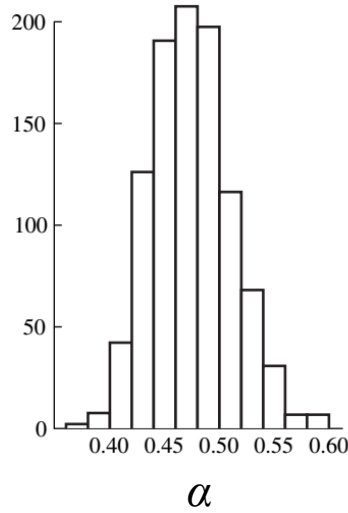
The method of moments gives  $\hat{\alpha} = .375$   $\hat{\lambda} = 1.674$

Not in closed form, but we can do **bootstrap**!

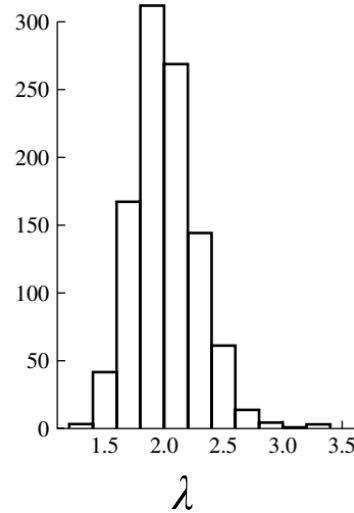
Generate 1000 samples  $n = 227$  of gamma distributed random variables with  $\alpha = .441$  and  $\lambda = 1.96$ . For each of sample, the MLEs of  $\alpha$  and  $\lambda$  are calculated.



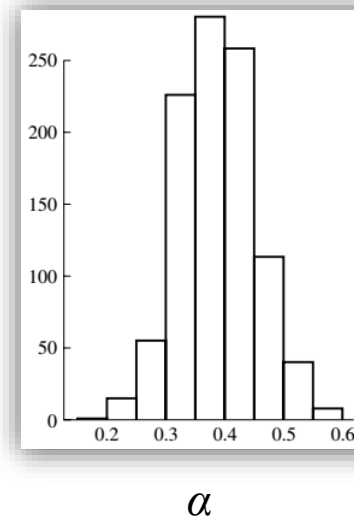
Standard error = 0.03



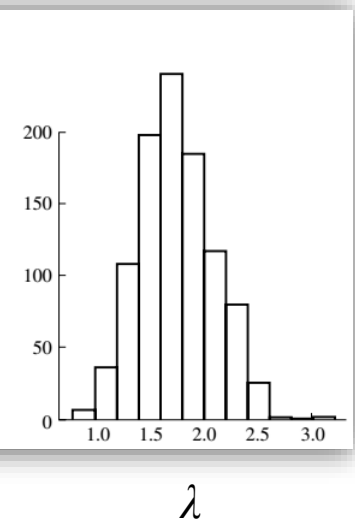
0.26



0.06



0.34



Why is MLE significantly more precise than using moments?

## Multinomial distribution -- Revisited!

Each of  $n$  independent trials can result in one of  $r$  types of outcomes, on each trial the probabilities of the  $r$  outcomes are  $p_1, p_2, \dots, p_r$ .

$N_i$  = total # of outcomes of type  $i$  in the  $n$  trials,  $i=1, \dots, r$ . (e.g. God is playing a dice...) Any particular sequence of trials giving rise to  $N_1=n_1, \dots, N_r=n_r$  occurs with the probability

$$p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$$

How many such sequences are there?

**Equivalent question:** What is the # of ways that  $n$  objects are grouped into  $r$  classes (types of outcomes) with  $n_i$  in the  $i$ th class,  $i=1, \dots, r$ ?

Joint frequency function:  $p(n_1, \dots, n_r) = \binom{n}{n_1 \cdots n_r} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$

$$\binom{n}{n_1 n_2 \cdots n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}$$

**Marginal distribution of  $N_i$ ?** -- Direct summation is daunting!

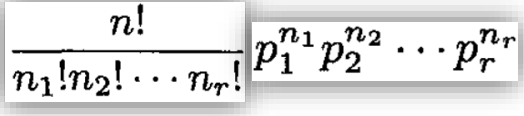
$N_i$  can be interpreted as # of success in  $n$  trials, each of which has  $p_i$  of success.

Binomial random variable  $N_i$  renders  $p_{N_i}(n_i) = \binom{n}{n_i} p_i^{n_i} (1 - p_i)^{n-n_i}$

## MLEs of Multinomial Cell Probabilities (多项单元概率)

Suppose that  $X_1, \dots, X_m$ , the counts in cells  $1, \dots, m$ , follow a multinomial distribution with a total count of  $n$  and cell probabilities  $p_1, \dots, p_m$ .

We wish to estimate the  $p$ 's from the  $x$ 's. The joint distribution of  $X_1, \dots, X_m$  is

$$f(x_1, \dots, x_m | p_1, \dots, p_m) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}$$


$X_i$  are not independent (constrained to sum to  $n$ , meaning  $p_i$  sum to 1). Log likelihood:

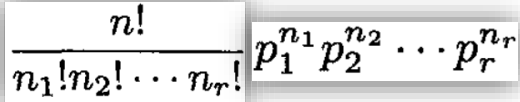
$$l(p_1, \dots, p_m) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i$$

To maximize it subject to the constraint, we should use a...

## MLEs of Multinomial Cell Probabilities (多项单元概率)

Suppose that  $X_1, \dots, X_m$ , the counts in cells  $1, \dots, m$ , follow a multinomial distribution with a total count of  $n$  and cell probabilities  $p_1, \dots, p_m$ .

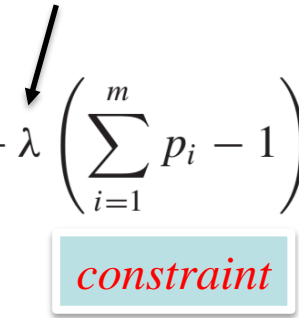
We wish to estimate the  $p$ 's from the  $x$ 's. The joint distribution of  $X_1, \dots, X_m$  is

$$f(x_1, \dots, x_m | p_1, \dots, p_m) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}$$


$X_i$  are not independent (constrained to sum to  $n$ , meaning  $p_i$  sum to 1). Log likelihood:

$$l(p_1, \dots, p_m) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i$$

To maximize it subject to the constraint, we should use a... **Lagrange multiplier**, and maximize, instead,

$$L(p_1, \dots, p_m, \lambda) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i + \lambda \left( \sum_{i=1}^m p_i - 1 \right)$$


*constraint*

## MLEs of Multinomial Cell Probabilities (多项单元概率)

$$L(p_1, \dots, p_m, \lambda) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i + \lambda \left( \sum_{i=1}^m p_i - 1 \right)$$

Setting the partial derivatives to 0, we find a system of equations:

$$\hat{p}_j = -\frac{x_j}{\lambda}, \quad j = 1, \dots, m$$

Summing both sides,

$$1 = \frac{-n}{\lambda} \quad \lambda = -n$$

Finally... Not surprising at all!

$$\hat{p}_j = \frac{x_j}{n}$$

In some situations (e.g. frequently occur in genetics), the multinomial cell probabilities are functions of other unknown parameters  $\theta$ ; that is,  $p_i = p_i(\theta)$ . The log likelihood of  $\theta$  is

$$l(\theta) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i(\theta)$$



## MLEs of Multinomial Cell Probabilities (多项单元概率)

Example: Hardy-Weinberg Equilibrium (Genetics)

例 8.5.1.1 (哈代-温伯格平衡) 如果基因频率是平衡的, 那么根据哈代-温伯格定律, 基因型  $AA$ ,  $Aa$  和  $aa$  在总体中出现的频率分别是  $(1 - \theta)^2$ ,  $2\theta(1 - \theta)$  和  $\theta^2$ . 在 1937 年中国香港人口总体的抽样中, 血型发生频率如下, 其中  $M$  和  $N$  是红细胞抗原:

频 率	血 型			总 计
	$M$	$MN$	$N$	
	342	500	187	1029

最简单的估计: 令  $\theta^2 = 187/1029$ ,  $\theta = 0.4263$ , 但显然没有充分利用信息; 如果令  $X_1, X_2, X_3$  表示三个单元的观测数,  $n = 1029$ ,

$$f(x_1, \dots, x_m | p_1, \dots, p_m) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}$$

$$\begin{aligned} l(\theta) &= \log n! - \sum_{i=1}^3 \log X_i! + X_1 \log(1 - \theta)^2 + X_2 \log 2\theta(1 - \theta) + X_3 \log \theta^2 \\ &= \log n! - \sum_{i=1}^3 \log X_i! + (2X_1 + X_2) \log(1 - \theta) + (2X_3 + X_2) \log \theta + X_2 \log 2 \end{aligned}$$

显然已有  $\sum_{i=1}^3 p_i(\theta) = 1$ , 不必另引入限制条件。

## MLEs of Multinomial Cell Probabilities (多项单元概率)

*Example:* Hardy-Weinberg Equilibrium (Genetics)

偏导数设为0，得到

$$-\frac{2X_1 + X_2}{1 - \theta} + \frac{2X_3 + X_2}{\theta} = 0$$

$$\begin{aligned}\hat{\theta} &= \frac{2X_3 + X_2}{2X_1 + 2X_2 + 2X_3} \\ &= \frac{2X_3 + X_2}{2n} \\ &= \frac{2 \times 187 + 500}{2 \times 1029} = .4247\end{aligned}$$

令 $\theta^2=187/1029$ , $\theta=0.4263$
--

## MLEs of Multinomial Cell Probabilities (多项单元概率)

*Example:* Hardy-Weinberg Equilibrium (Genetics)

偏导数设为0, 得到

$$-\frac{2X_1 + X_2}{1 - \theta} + \frac{2X_3 + X_2}{\theta} = 0$$
$$\hat{\theta} = \frac{2X_3 + X_2}{2X_1 + 2X_2 + 2X_3}$$
$$= \frac{2X_3 + X_2}{2n}$$
$$= \frac{2 \times 187 + 500}{2 \times 1029} = .4247$$

$$\text{令 } \theta^2 = 187/1029, \\ \theta = 0.4263$$

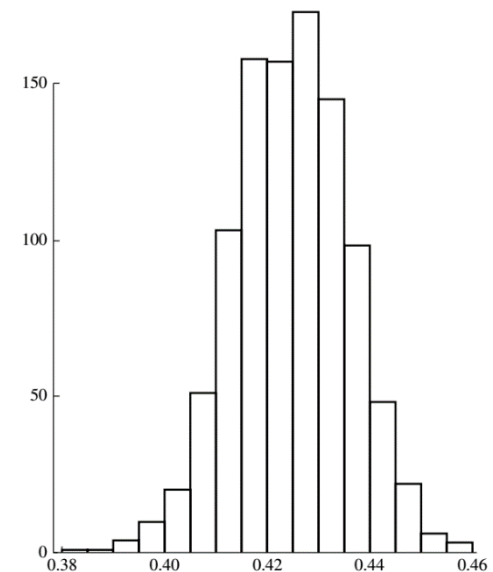
How precise is the estimate? How many digits should be kept? **Bootstrap!**

Simulate many multinomial random variables with these probabilities and  $n=1029$ , and for each we form an estimate of  $\theta$ . Histograms are  $\sim$  sampling distribution.

$\theta$  is unknown, we use  $\hat{\theta} = .4247$  in its place, cell probabilities are .331, .489, .180.

1000 computer experiments, each has a  $\theta^*$ . Finally,

$$s_{\hat{\theta}} = .011$$



## Large sample theory for MLEs

*We will skip all the proofs of the theorems in this section.*

Heuristically, we consider the case of an **i.i.d. sample** and a 1-d parameter.

$$l(\theta) = \sum_{i=1}^n \log f(x_i|\theta)$$

$\theta$  – true value =  $\theta_0$ , estimate =  $\hat{\theta}$

### THEOREM A 相合估计

Under appropriate smoothness conditions on  $f$ , the mle from an i.i.d. sample is consistent.

### LEMMA A

Define  $I(\theta)$  by

$$I(\theta) = E \left[ \frac{\partial}{\partial \theta} \log f(X|\theta) \right]^2$$

$$\begin{aligned} I(\theta) &= E [l'(\theta)]^2 \\ &= -E [l''(\theta)] \end{aligned}$$

Under appropriate smoothness conditions on  $f$ ,  $I(\theta)$  may also be expressed as

$$I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right]$$

## Large sample theory for MLEs

The large sample distribution of a MLE is approximately Gaussian  
-- mean =  $\theta_0$ , variance =  $1/[nI(\theta_0)]$ .

This is a limiting result that holds as the sample size  $\rightarrow \infty$ , we say that

- MLE is **asymptotically unbiased**
- variance of the limiting normal distribution = **asymptotic variance of the mle**

### THEOREM B

Under smoothness conditions on  $f$ , the probability distribution of  $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$  tends to a standard normal distribution.

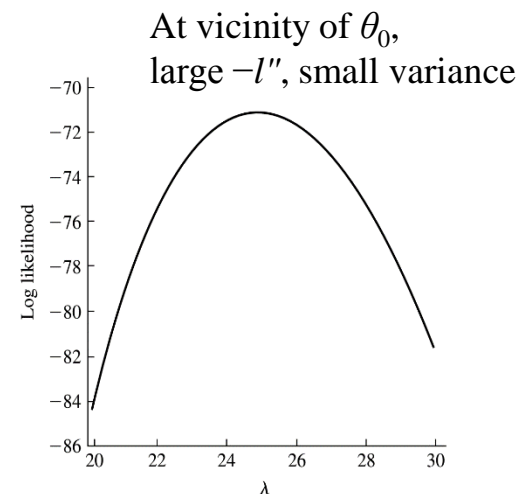
The MLE is the maximizer of the log likelihood function

$$l(\theta) = \sum_{i=1}^n \log f(X_i|\theta)$$

The asymptotic variance is

$$\frac{1}{nI(\theta_0)} = -\frac{1}{E l''(\theta_0)}$$

$$\begin{aligned} I(\theta) &= E [l'(\theta)]^2 \\ &= -E [l''(\theta)] \end{aligned}$$



## Large sample theory for MLEs: multi-d

Multidimensional case is similar: now  $\theta$  is a vector

Mean of asymptotic distribution = vector of true parameters,  $\theta_0$ .

Covariance of  $\hat{\theta}_i$  and  $\hat{\theta}_j$  is given by the  $ij$  entry of the matrix  $n^{-1}I^{-1}(\theta_0)$ , where the matrix  $I(\theta)$  has an  $ij$  component

$$E \left[ \frac{\partial}{\partial \theta_i} \log f(X|\theta) \frac{\partial}{\partial \theta_j} \log f(X|\theta) \right] = -E \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X|\theta) \right]$$

Conditions for the above to hold:

- True parameter value,  $\theta_0$ , is required to be an interior point of the set of all parameter values.
  - (e.g.  $|\alpha| \leq 1$ ,  $\alpha_0=1$ , inapplicable)
- The support of density/frequency function  $f(x|\theta)$  (*the set of values for which  $f(x|\theta) > 0$* ) does not depend on  $\theta$ .
  - (e.g. uniform distribution on  $[0, \theta]$ , inapplicable)

## Confidence intervals from MLEs

Recall that

- a confidence interval for  $\theta$  is an (random) interval based on the sample values used to estimate  $\theta$ .
- The probability of containing  $\theta$  is called the coverage probability of the interval.

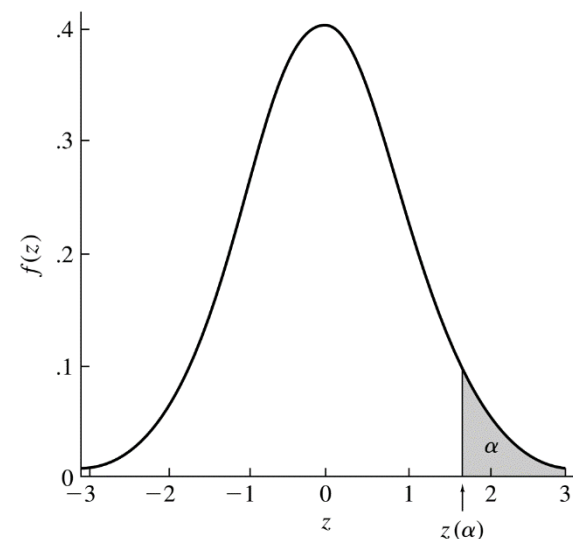
Three methods for forming confidence intervals for MLEs:

1. Exact methods
2. Approximations based on large sample properties of MLEs
3. Bootstrap confidence intervals

$$P\left(-z(\alpha/2) \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq z(\alpha/2)\right) \approx 1 - \alpha$$

$$P(\bar{X} - z(\alpha/2)\sigma_{\bar{X}} \leq \mu \leq \bar{X} + z(\alpha/2)\sigma_{\bar{X}}) \approx 1 - \alpha$$

区间为平均值±数倍 $\sigma$ 形式







## Confidence intervals from MLEs: Exact methods

*Example: Gaussian.* First of all, let's prove two important conclusions:

- (1) The distribution of  $(n-1)S^2/\sigma^2$  is the chi-square distribution with  $n-1$  degrees of freedom.

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 + \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 / \sigma^2 \sim \chi^2_{n-1}$$

In fact,  $W=U+V$ ,  $S^2$  and  $\bar{X}$  are independent (a theorem), mgf  $M_W = M_U M_V$ ,

$$M_U(t) = \frac{M_W(t)}{M_V(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2}$$

莫忘：卡方只是伽马分布的特例， $(n/2, 1/2)$

- (2)  $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)}{\sqrt{S^2/\sigma^2}} \begin{matrix} \nearrow \sim N(0, 1) \\ \searrow \sim \chi^2_{n-1} \end{matrix}$$

定义： $Z \sim N(0, 1)$ ,  $U \sim \chi^2_n$ ,  $Z$ 与 $U$ 独立，则

$$Z/\sqrt{U/n} \sim t_n$$

## Confidence intervals from MLEs: Exact methods

Example: Gaussian.

### Confidence interval for $\mu$ .

Earlier in this lecture we found that MLEs from an i.i.d. normal sample are

$$\hat{\mu} = \bar{X} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

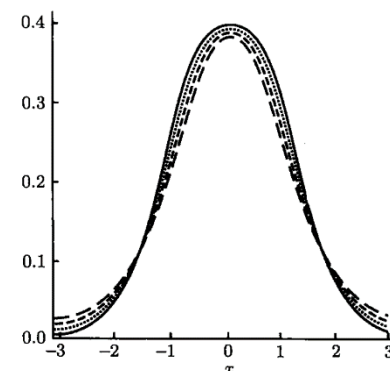
Since  $\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$

Let  $t_{n-1}(\alpha/2)$  denote the point beyond which  $t_{n-1}$  has probability  $\alpha/2$ .

$$P \left( -t_{n-1}(\alpha/2) \leq \frac{\sqrt{n}(\bar{X} - \mu)}{S} \leq t_{n-1}(\alpha/2) \right) = 1 - \alpha$$

$$P \left( \underline{\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2)} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2) \right) = 1 - \alpha$$

Symmetric  
about 0



## Confidence intervals from MLEs: Exact methods

*Example:* Gaussian. **Confidence interval for  $\sigma^2$ .**

Now that 
$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-1}^2$$

Let  $\chi_{n-1}^2(\alpha/2)$  denote the point  $>$  which  $\chi_{n-1}^2$  has prob.  $\alpha/2$ .

$$P\left(\chi_{n-1}^2(1 - \alpha/2) \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq \chi_{n-1}^2(\alpha/2)\right) = 1 - \alpha$$

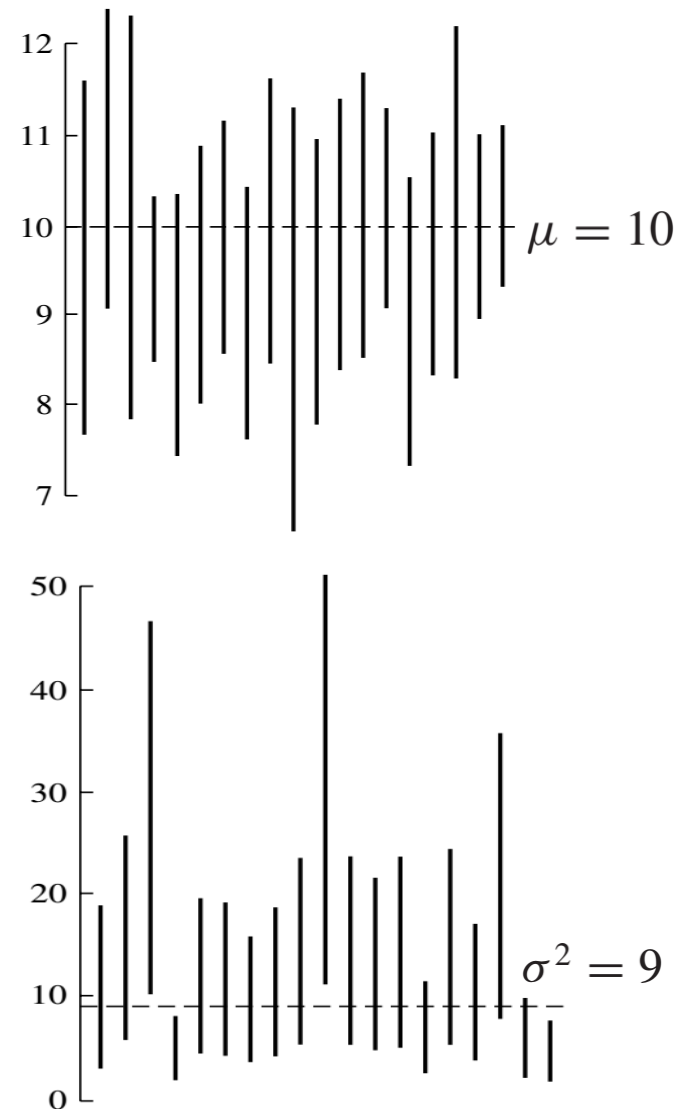
$$P\left(\frac{n\hat{\sigma}^2}{\chi_{n-1}^2(\alpha/2)} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{\chi_{n-1}^2(1 - \alpha/2)}\right) = 1 - \alpha$$

不对称!!

Simulation: Generate a random sample of size  $n=11$  from Gaussian with  $\mu=10$ ,  $\sigma^2=9$ . Do 20 experiments. Calculate 90% confidence intervals.

- 10% of the time true value falls outside?

simulation



## Confidence intervals from MLEs: Large sample theory

*Exact methods are the exception rather than the rule in practice.*

If i.i.d, we know the distribution of  $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$  is approximately standard normal.

$\theta_0$  is unknown, use  $I(\hat{\theta})$  in place of  $I(\theta_0)$ .

In fact, the distribution of  $\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0)$  is also approximately standard normal.

$$P\left(-z(\alpha/2) \leq \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \leq z(\alpha/2)\right) \approx 1 - \alpha$$

Confidence interval is

$$\hat{\theta} \pm z(\alpha/2) \frac{1}{\sqrt{nI(\hat{\theta})}}$$

## Confidence intervals from MLEs: Large sample theory

*Example:* Poisson

MLE of  $\lambda$  from a sample of size  $n$  from Poisson is  $\hat{\lambda} = \bar{X}$

For large samples, let's calculate  $I(\lambda)$  first.

$f(x|\lambda)$  -- probability mass function of a Poisson variable with  $\lambda$ .

$$\log f(x|\lambda) = x \log \lambda - \lambda - \log x!$$

$$\frac{\partial^2}{\partial \lambda^2} \log f(X|\lambda) = -\frac{X}{\lambda^2}$$

$$I(\lambda) = -E \left[ \frac{\partial^2}{\partial \lambda^2} \log f(X|\lambda) \right] = \frac{E(X)}{\lambda^2} = \frac{1}{\lambda}$$

Thus, an approximate  $100(1 - \alpha)\%$  confidence interval for  $\lambda$  is

$$\bar{X} \pm z(\alpha/2) \sqrt{\frac{\bar{X}}{n}}$$

## Confidence intervals from MLEs: Large sample theory

*Random multinomial counts:* no longer i.i.d., variance of parameter estimate is not  $1/[n I(\theta)]$ . But it can be shown that

$$\text{Var}(\hat{\theta}) \approx \frac{1}{E[l'(\theta_0)^2]} = -\frac{1}{E[l''(\theta_0)]}$$

*Example:* Hardy-Weinberg Equilibrium again.

$$l''(\theta) = -\frac{2X_1 + X_2}{(1 - \theta)^2} - \frac{2X_3 + X_2}{\theta^2}$$

Since the  $X_i$  are binomially distributed, we have

$$E(X_1) = n(1 - \theta)^2$$

$$E(X_2) = 2n\theta(1 - \theta)$$

$$E(X_3) = n\theta^2$$

$$E[l''(\theta)] = -\frac{2n}{\theta(1 - \theta)}$$

Substitute  $\theta$  with  $\hat{\theta}$ , the standard error of  $\hat{\theta}$ :

$$s_{\hat{\theta}} = \frac{1}{\sqrt{-I''(\hat{\theta})}} = \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{2n}} = .011$$

与自助法结果相同:

$$s_{\hat{\theta}} = .011$$

An ~95% confidence interval is

$$\hat{\theta} \pm 1.96s_{\hat{\theta}}, \text{ or } (.403, .447)$$

## Confidence intervals from MLEs: Bootstrap

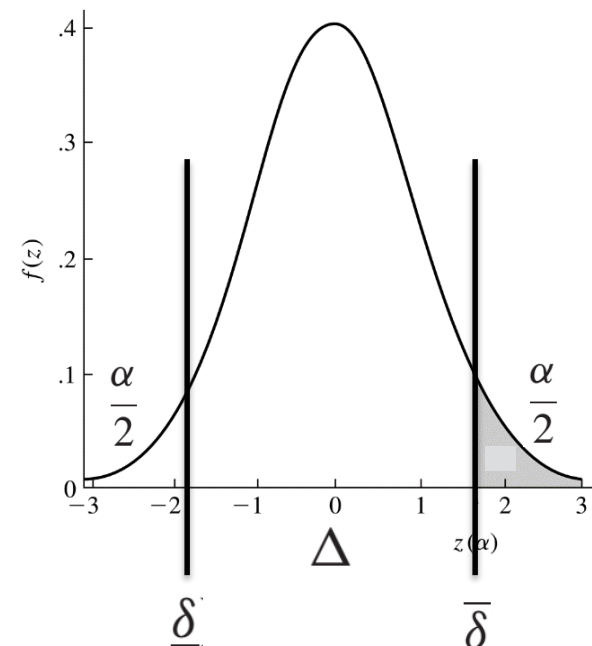
Suppose we know the distribution of  $\Delta = \hat{\theta} - \theta_0$

Denote  $\alpha/2$  and  $1 - \alpha/2$  quantiles by  $\underline{\delta}$  and  $\bar{\delta}$

$$P(\hat{\theta} - \theta_0 \leq \underline{\delta}) = \frac{\alpha}{2} \quad P(\hat{\theta} - \theta_0 \leq \bar{\delta}) = 1 - \frac{\alpha}{2}$$

$$P(\underline{\delta} \leq \hat{\theta} - \theta_0 \leq \bar{\delta}) = 1 - \alpha$$

$$P(\hat{\theta} - \bar{\delta} \leq \theta_0 \leq \hat{\theta} - \underline{\delta}) = 1 - \alpha$$



If  $\theta_0$  is known, the distribution of  $\hat{\theta} - \theta_0$  can be approximated arbitrarily well by simulation: randomly generate many, many samples of observations with  $\theta_0$ , for each sample, record  $\hat{\theta} - \theta_0$ . Once we have this distribution, determine  $\underline{\delta}$  and  $\bar{\delta}$  as accurately as desired.

-- But  $\theta_0$  is unknown! No worries, use  $\hat{\theta}$ !

Generate many, many samples (say,  $B$  in all) from a distribution with value  $\hat{\theta}$ ; for each sample construct an estimate of  $\theta$ , say  $\theta_j^*$ ,  $j=1, 2, \dots, B$ . Characterize the distribution of  $\theta^* - \hat{\theta}$ , the quantiles of which then form an approximate confidence interval.

## Confidence intervals from MLEs: Bootstrap

*Example.* Hardy-Weinberg Equilibrium again.

Use bootstrap to find an ~95% confidence interval, then compare to results from large-sample theory for MLEs.

We already did 1000 bootstrap estimates, ready to use.

The 25<sup>th</sup> largest is .403, the 975<sup>th</sup> largest is .446: estimates of .025 and .975 quantiles.

$\theta^* - \hat{\theta}$  distribution? Subtract  $\hat{\theta} = .425$  from each  $\theta_i^*$

.025 and .975 quantiles are estimated as

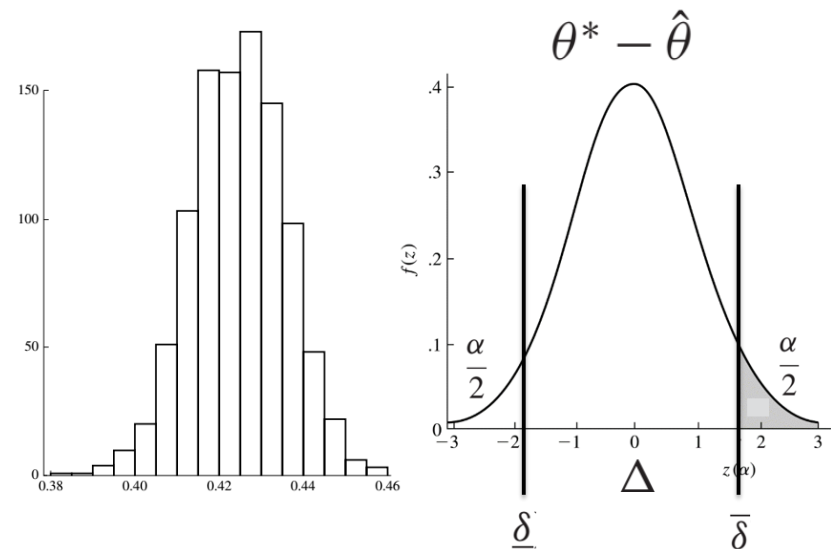
$$\underline{\delta} = .403 - .425 = -.022$$

$$\bar{\delta} = .446 - .425 = .021$$

Our approximate 95% confidence interval is

$$(\hat{\theta} - \bar{\delta}, \hat{\theta} - \underline{\delta}) = (.404, .447)$$

大样本理论给出结果: (.403, .447)





# Problem set #4

1. Consider simple random sampling *with* replacement.

a. Show that

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

is an unbiased estimate of  $\sigma^2$ .

b. Is  $s$  an unbiased estimate of  $\sigma$ ?

c. Show that  $n^{-1}s^2$  is an unbiased estimate of  $\sigma_{\bar{X}}^2$ .

d. Show that  $n^{-1}N^2s^2$  is an unbiased estimate of  $\sigma_T^2$ .

e. Show that  $\hat{p}(1 - \hat{p})/(n - 1)$  is an unbiased estimate of  $\sigma_{\hat{p}}^2$ .

2. Consider our old friend the **regression line**, for which we have values of  $Y_i$  measured at given values of the independent variable  $X_i$ . Our model is  $y(a, b) = ax + b$  and assuming that the  $Y_i$  have a Gaussian scatter, each term in the likelihood product is

$$\mathcal{L}_i(y|(a, b)) = \exp \left[ -\frac{(Y_i - (aX_i + b))^2}{2\sigma^2} \right]$$

i.e. the residuals are  $(Y_i - \text{model})$ , and our model has the free parameters  $(a, b)$ .

Maximizing the log likelihood and show that the estimates should take the form of the **ordinary least square** fit:

$$a = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\overline{X^2} - (\bar{X})^2}, \quad b = \bar{Y} - a\bar{X}.$$

# The Bayesian approach to parameter estimation