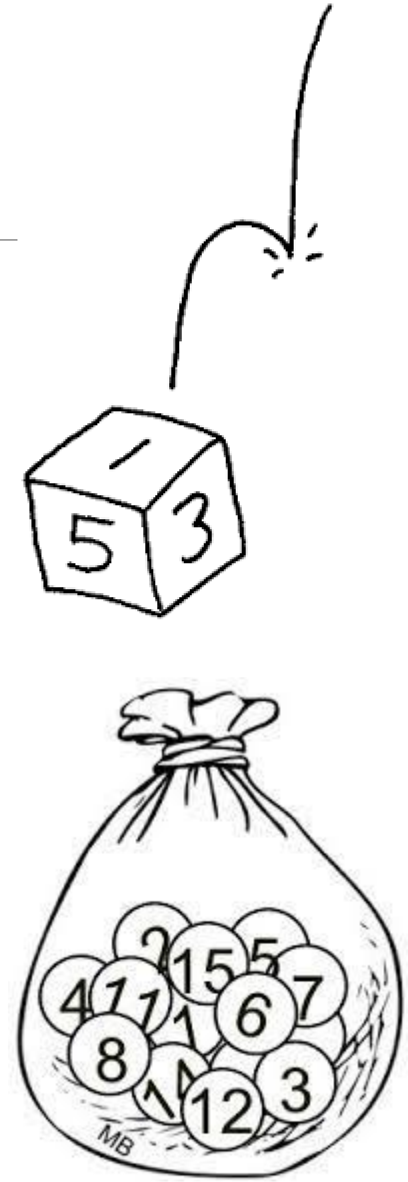
Lecture 15

- Sampling
- Statistics (统计量) from a sample



Statistics

数理统计

Probability:
From PMF/PDF/CDF of population to event probabilities.

Statistics:
From sample(s) to statistics/properties of population.

Statistics

- Sampling
- Parameter Estimation
- Hypothesis Testing (Optional)

Motivating example

Bhutan (不丹)

You want to know the true mean and variance of happiness in Bhutan.

- But you can't ask everyone.
- You poll 200 random people.
- Your data looks like this:

Happiness = {72, 85, 79, 91, 68, …, 71}

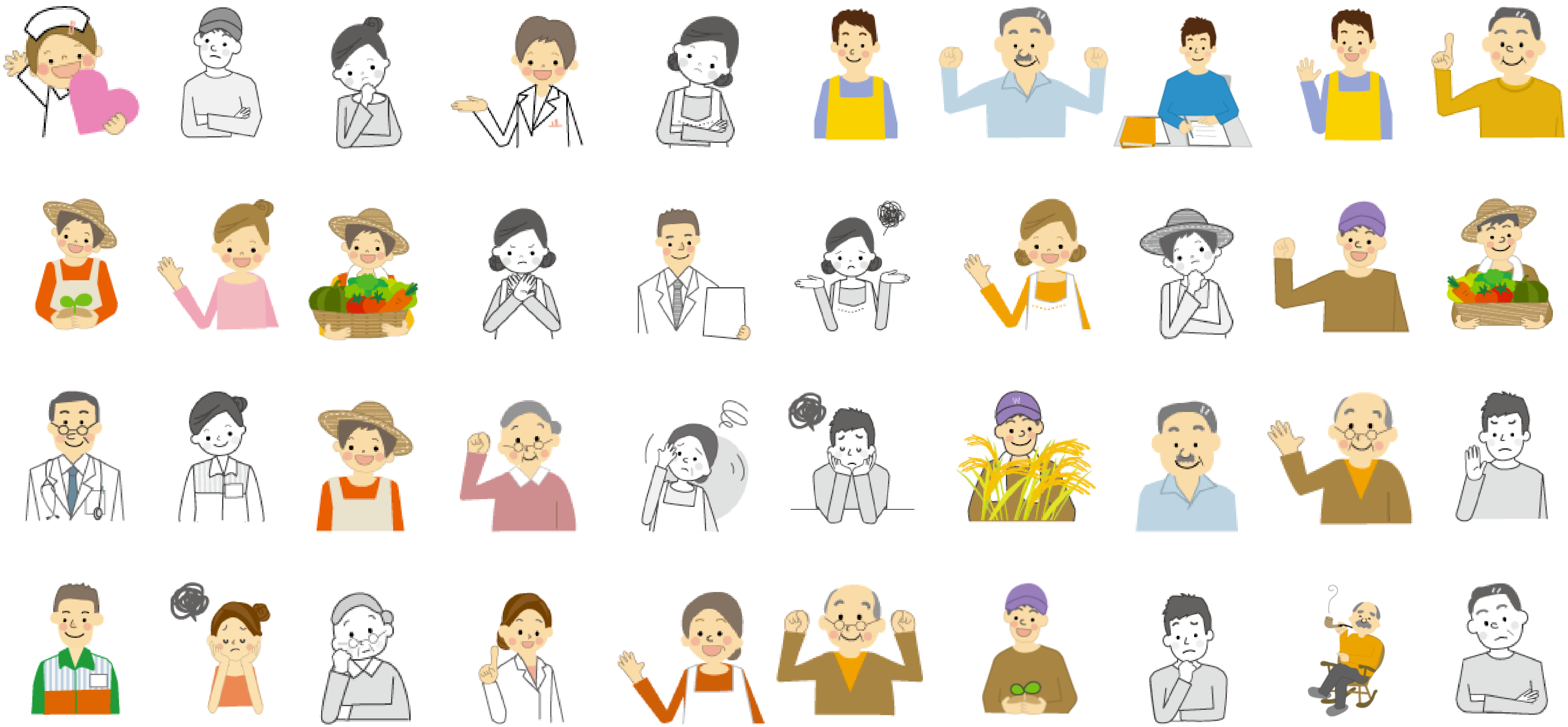
- The mean of all these numbers is 83.

Is this the true mean happiness of Bhutanese people?

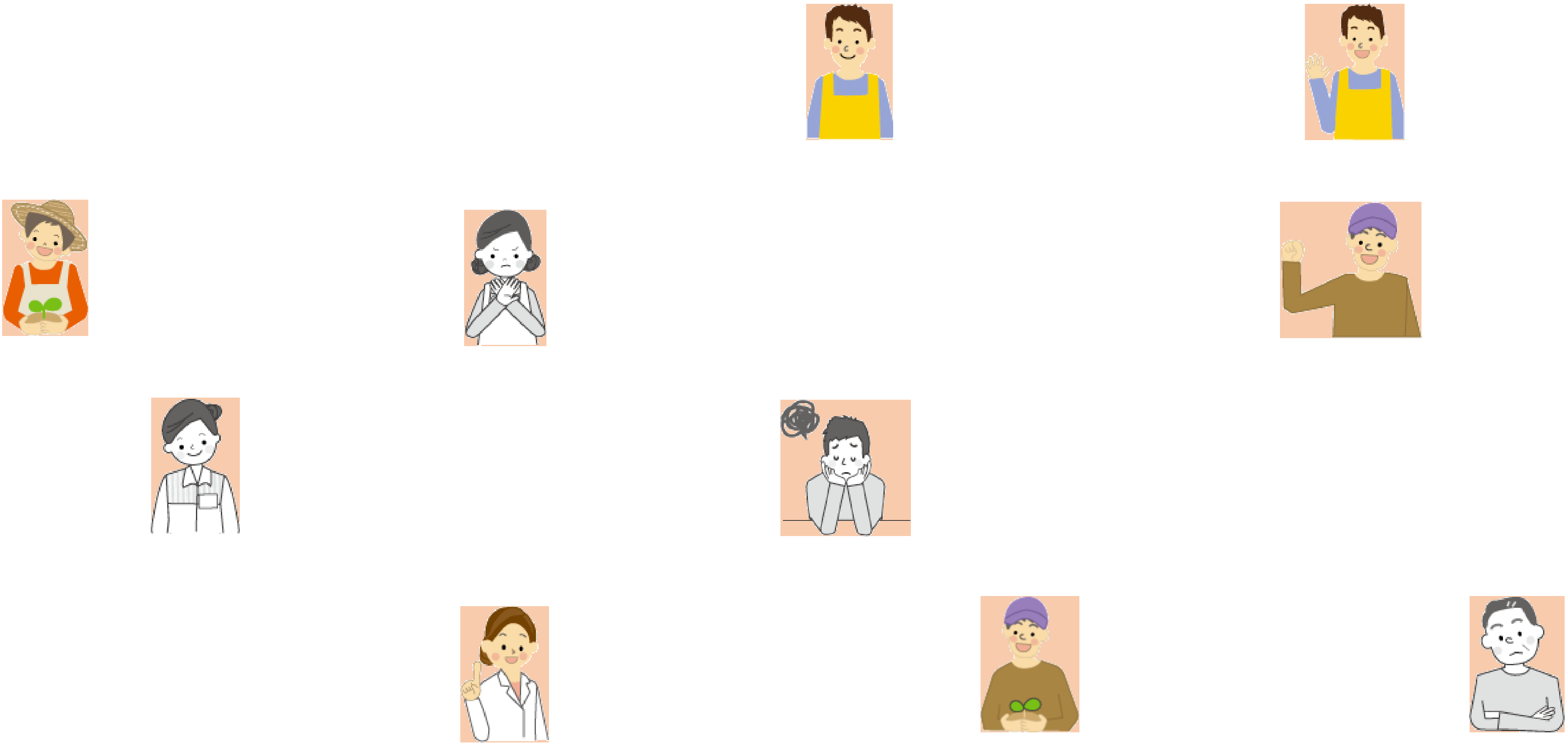
Of course NOT!

But what can we learn from these data.

Population (总体)



Sample (样本)



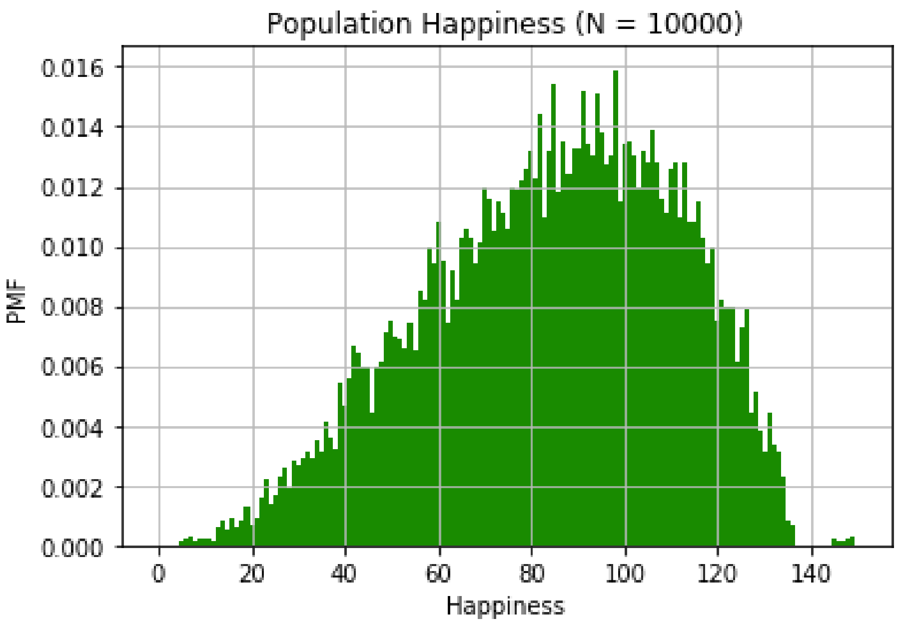
A sample is selected from a population.

A sample (一个样本), mathematically

Consider n random variables $X_1, X_2, \ldots, X_n$.

The sequence $X_1, X_2, \ldots, X_n$ is a sample from distribution F if:

- X_i are pairwise independent
- n is the size of sample (样本容量)
- All X_i have the same distribution function F (the underlying distribution), where $E[X_i] = \mu, D[X_i] = \sigma^2$



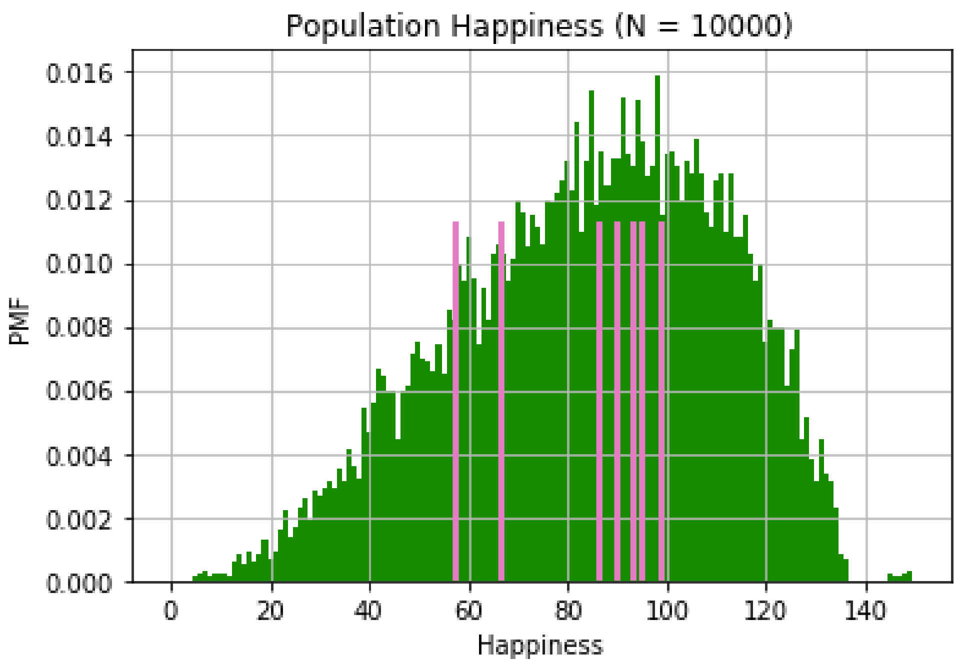
A sample (一个样本), mathematically

A sample of sample size (样本容量) 8:

$$(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

A realization (观察值) of a sample of size 8:

(59, 87 , 94 , 99, 87 , 78 , 69 , 91)



A sample (一个样本)

A happy Bhutanese person

If we had a distribution F of our entire population, we could compute exact statistics about happiness.

But we only have 200 people (a sample).

In this part: If we only have a single sample,

- How do we report *estimated* statistics?
- How do we report estimated error of these estimates?
- How do we perform hypothesis testing?

Statistics (统计量) from a sample

A happy Bhutanese person

If we had a distribution F of our entire population, we could compute exact statistics about happiness.

But we only have 200 people (a sample).

- Therefore, these population statistics are unknown:
  - μ, the population mean (总体均值)
  - σ^2, the population variance (总体方差)

Estimating the population mean

1. What is our best estimate of μ, the mean happiness of Bhutanese people?

If we only have a sample, $(X_1, X_2, \ldots, X_n)$:

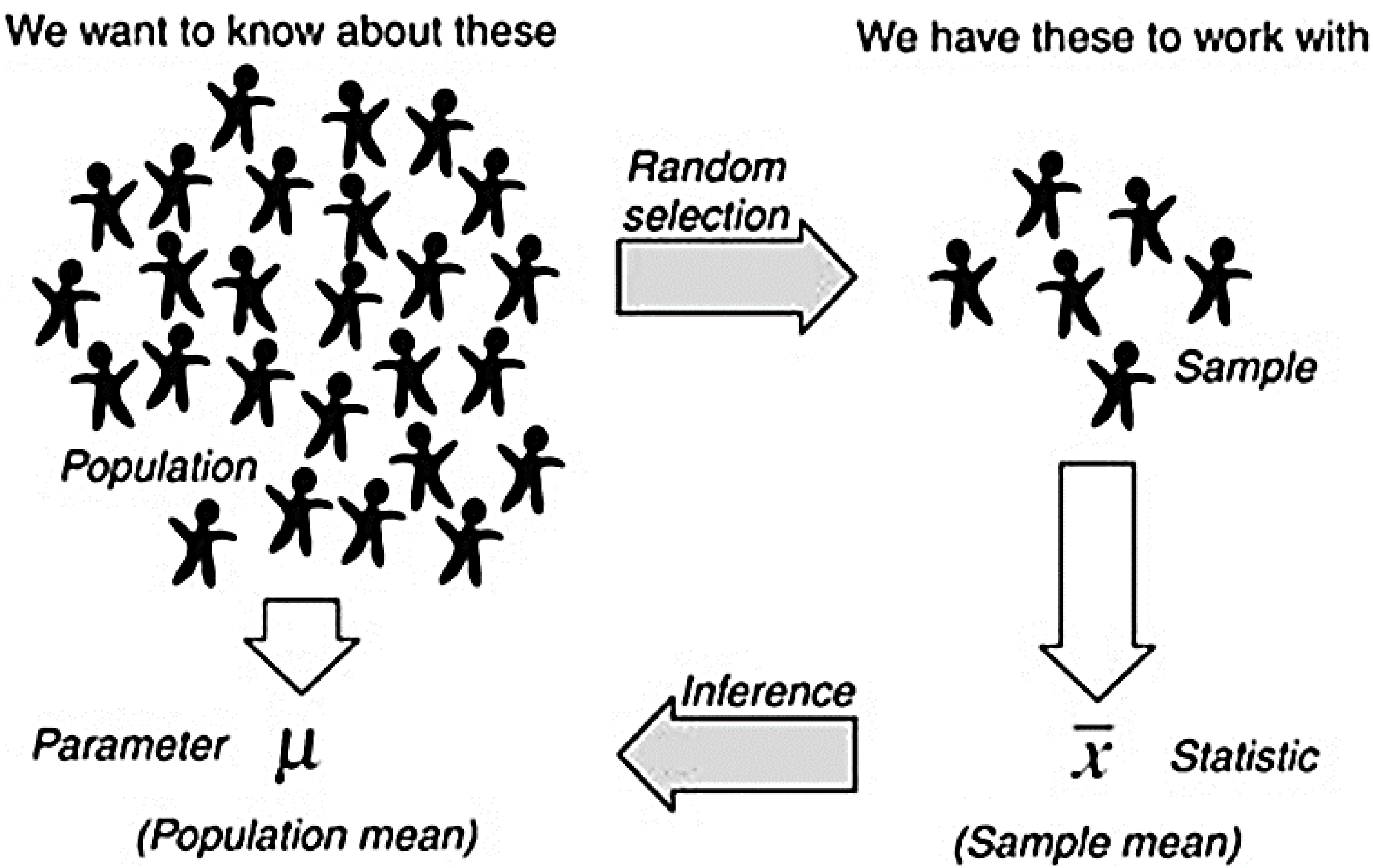
The best estimate of μ is the sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

(样本均值)

$\bar{X}$ is an **unbiased estimator** of the population mean.

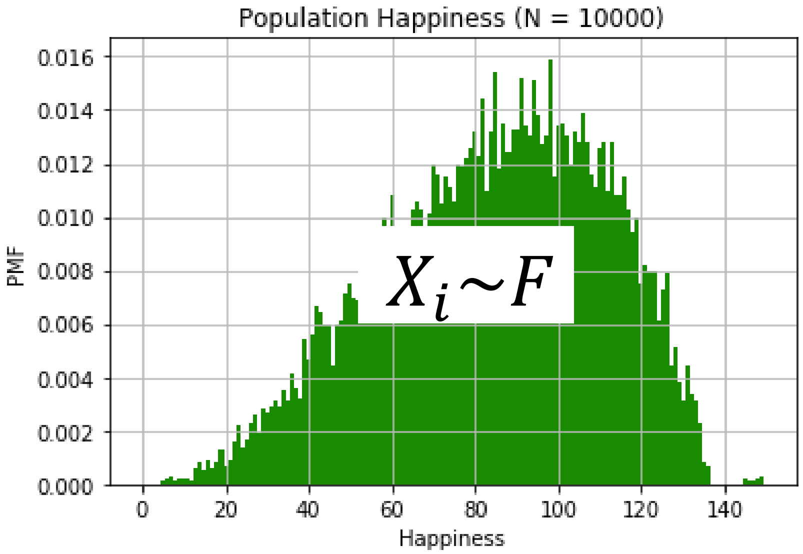
(无偏估计)

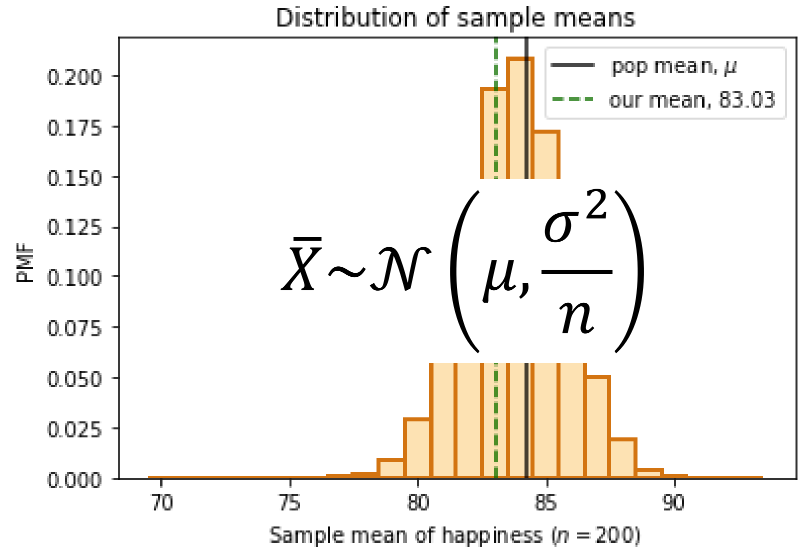
From C.L.T., $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \Longrightarrow E[\bar{X}] = \mu$.



[Statistical inference for data science]()

Sample mean





Even if we can't report μ, we can report our sample mean 83.03, which is an unbiased estimate of μ.

Estimating the population variance

2. What is σ^2, the variance of happiness of Bhutanese people?

If we knew the entire population $(X_1, X_2, \ldots, X_N)$:

Population variance: $\sigma^2 = E[(X-\mu)^2] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

总体方差

population mean

If we only have a sample, $(X_1, X_2, \ldots, X_n)$:

Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

样本方差

sample mean

Estimating the population variance

<u>Actual, σ^2</u>

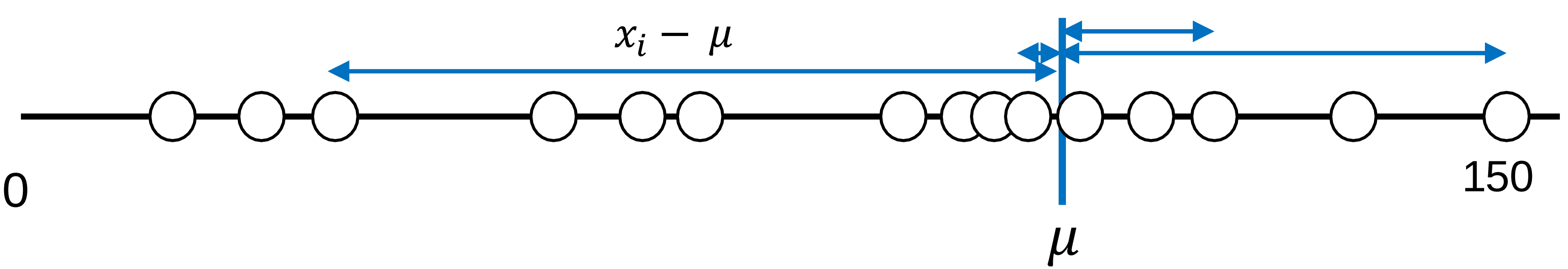
population variance

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

<u>Estimate, S</u>

sample variance

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



Happiness

Population size, N

Calculating population statistics **<u>exactly</u>** requires us knowing all N data points.

Estimating the population variance

<u>Actual, σ^2</u>

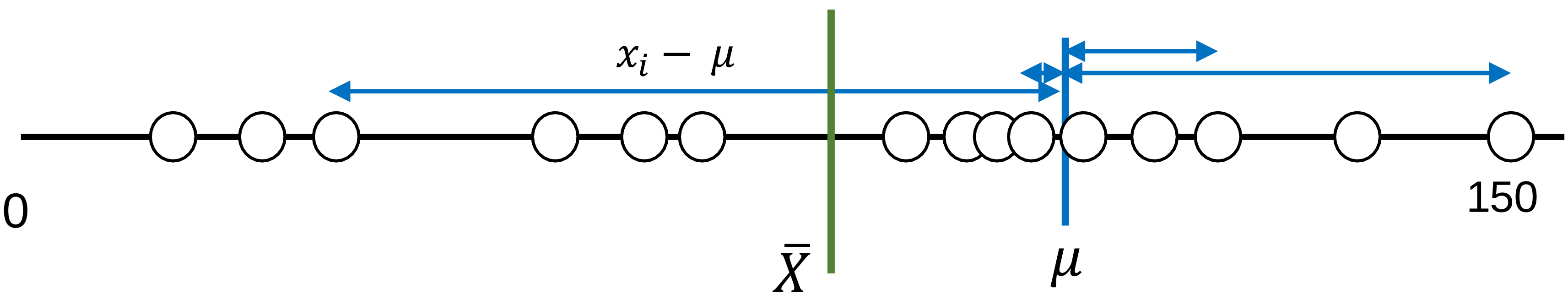
population variance

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

<u>Estimate, S</u>

sample variance

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



Happiness

Population size, N

Sample variance is "an estimate <u>using an estimate</u>", so it needs **additional scaling**.

Estimating the population variance

2. What is σ^2, the variance of happiness of Bhutanese people?

If we only have a sample, $(X_1, X_2, \dots, X_n)$:

The best estimate of σ^2 is the sample variance:

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

S^2 is an unbiased estimator of the population variance,

$$E[S^2] = \sigma^2$$

Proof in the next slide.

Proof: S^2 is an unbiased estimator of σ^2

$$E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{N}(X_i - \bar{X})^2\right] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right]$$

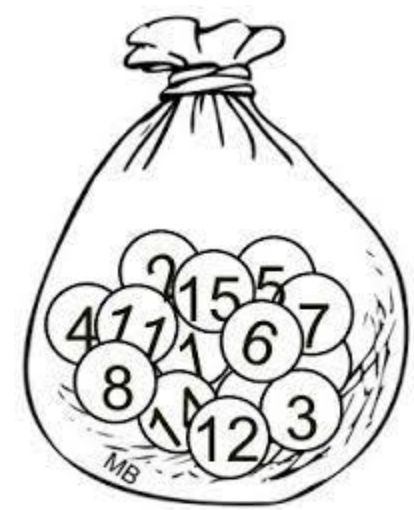
$$= E\left[\frac{1}{n-1}\left(\sum\nolimits_{i=1}^{n} X_i^2 - 2\bar{X}\sum\nolimits_{i=1}^{n} X_i + n\bar{X}^2\right)\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)\right]$$

Given $E\left(X_i^2\right) = D(X_i) + [E(X_i)]^2 = \sigma^2 + \mu^2$, $E(\bar{X}^2) = D(\bar{X}) + [E(\bar{X})]^2 = \frac{\sigma^2}{n} + \mu^2$

$$= \frac{1}{n-1}\left[n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] = \sigma^2$$

Lecture 15

- Sampling
- Statistics (统计量) from a sample



Statistics from a sample

Given a sample $(X_1, X_2, \dots, X_n)$ from the population, with value $(x_1, x_2, \dots, x_n)$

- Sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$;
- Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Sample standard deviation $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- Sample k-th order raw moment $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

$(k = 1,2, \dots)$

- Sample k-th order central moment $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$

$(k = 2,3, \dots)$

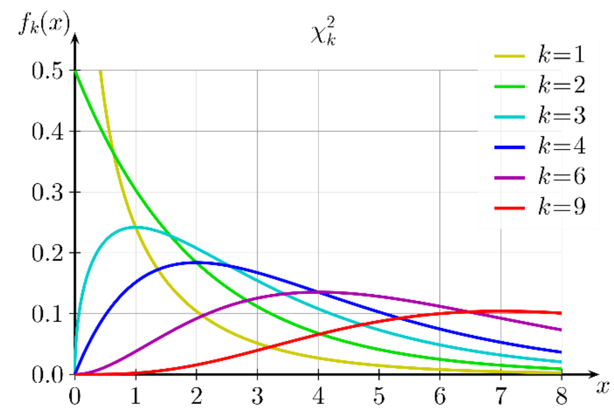
Distribution of sample statistics (统计量的分布)

χ^2 distribution

Def. Given $X_1, X_2, \dots, X_n$ is a sample from the population follows $\mathcal{N}(0,1)$, the statistics

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

follows χ^2 (Chi-square 卡方) distribution with n degree of freedom, $\chi^2 \sim \chi^2(n)$.



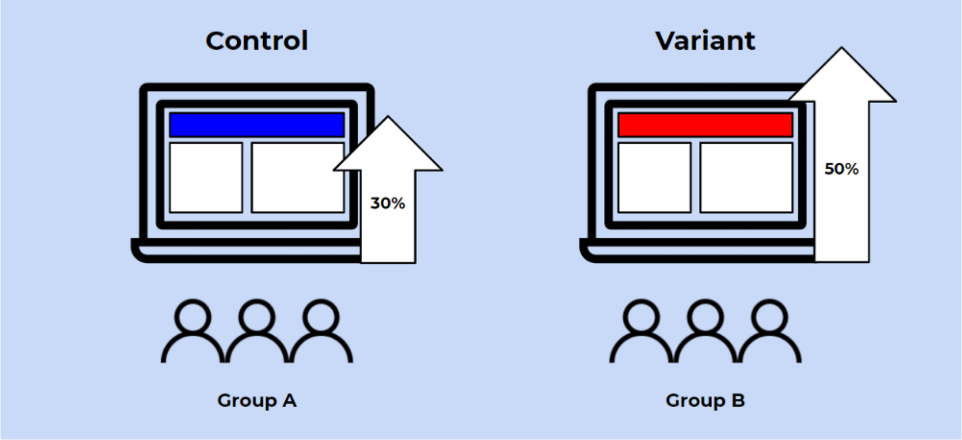
Note:

1) $X_i \sim \mathcal{N}(0,1)$. $X_1, X_2, \dots, X_n$ are i.i.d..

2) for $n = 1$, $X_1 \sim \mathcal{N}(0,1)$, then $X_1^2 \sim \chi^2(1)$.

Ex. $X \sim \mathcal{N}(0,2), Y \sim \mathcal{N}(0,4)$, then $\frac{1}{2}X^2 + \frac{1}{4}Y^2 = \underline{\boldsymbol{\chi}^2(2)}$.

A chi-square test is a statistical test used to compare observed results with expected results.



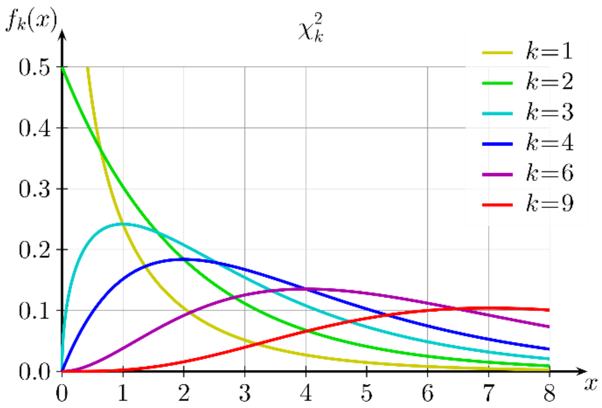
Chi-Squared Test Statistic

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

X_i observed data, $\bar{X}$ expected value

Proof (very complicated!)

	Click	No Click	Click + No Click
Advertisement A	360	140	500
Advertisement B	300	250	550
Ad A + Ad B	660	390	1050



Additive rule: Given $\chi_1^2 \sim \chi^2(n_1), \chi_2^2 \sim \chi^2(n_2)$, and χ_1^2, χ_2^2 are independent, then $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$.

Expected value and variance: Given $\chi^2 \sim \chi^2(n)$,

$$E[\chi^2] = n, \qquad D[\chi^2] = 2n$$

Proof: Given $X_i \sim \mathcal{N}(0,1)$,

$E\left[X_i^2\right] = D[X_i] = 1, \quad E\left[X_i^4\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^4 e^{-\frac{x^2}{2}} dx = 3,$ (integration by part)

$$D\left[X_i^2\right] = E\left[X_i^4\right] - \left[E\left[X_i^2\right]\right]^2 = 3 - 1 = 2$$

X_i are independent, thus

$$E[\chi^2] = E\left(\sum_{i=1}^n X_i^2\right) = \sum_{i=1}^n E\left[X_i^2\right] = n,$$

$$D[\chi^2] = D\left(\sum_{i=1}^n X_i^2\right) = \sum_{i=1}^n D\left[X_i^2\right] = 2n$$

Quick test

Expected value and variance: Given $\chi^2 \sim \chi^2(n)$,

$$E[\chi^2] = n, \qquad D[\chi^2] = 2n$$

Example: $X \sim \chi^2(5)$, $Y \sim U(0,4)$, X and Y are independent, thus $E(X - Y) =$ _____ $D(X - Y) =$ _______.

Sol.

$$E(X - Y) = E(X) - E(Y) = 5 - 2 = 3$$

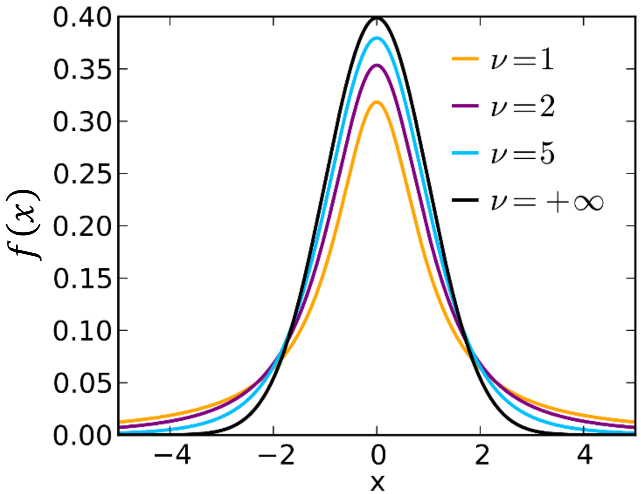
$$D(X - Y) = D(X) + D(Y) = 10 + \frac{4^2}{12} = 11\frac{1}{3}$$

Distribution of sample statistics

t distribution. Def. Given $X \sim \mathcal{N}(0,1), Y \sim \chi^2(n)$, and X, Y are independent,

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

follows t distribution with n degree of freedom.



$\nu \to \infty, \qquad f(x) \to \mathcal{N}(0,1)$

Ex. Given $X \sim \mathcal{N}(2,1)$, $Y_1, Y_2, \dots, Y_4$ follow $\mathcal{N}(0,4)$ and independent, how to form a t distribution with X and Y?

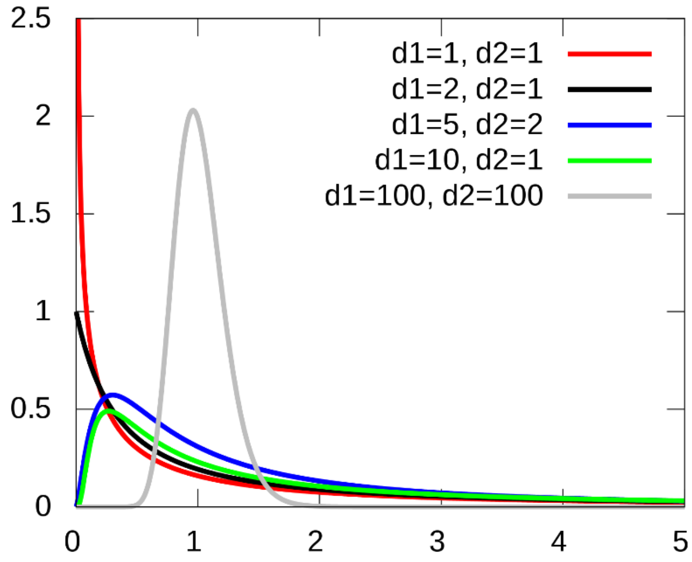
$$\frac{X-2}{\sqrt{\sum_{i=1}^{4}\left(\frac{Y_i}{2}\right)^2 \Big/ 4}} = \frac{4(X-2)}{\sqrt{\sum_{i=1}^{4} {Y_i}^2}} \sim t(4)$$

Distribution of sample statistics

F distribution. Def. Given $U \sim \chi^2(n_1), V \sim \chi^2(n_2)$, and U, V are independent,

$$F = \frac{U/n_1}{V/n_2}$$

follows F distribution with (n_1, n_2) degree of freedom.



Ex. Given $X_1, X_2, \ldots, X_n, X_{n+1}, \ldots, X_{n+m}$ follow $\mathcal{N}(0, \sigma^2)$, and

$$V = \frac{m\sum_{i=1}^{n} X_i^2}{n\sum_{i=n+1}^{n+m} X_i^2} \sim F(?, ?)$$

$F(n, m)$