COMPUTATIONAL FINANCE & RISK MANAGEMENT
UNIVERSITY *of* WASHINGTON
Department of Applied Mathematics

CFRM 410: Probability and Statistics for Computational Finance

Week 8 Random Samples

Jake Price
Instructor, Computational Finance and Risk Management
University of Washington
Slides originally produced by Kjell Konis

Random Samples

Outline

Motivation

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent Bernoulli(p) trials

Let $S_n = \sum_{i=1}^{n} X_i$

Then $S_n \sim \text{Binomial}(n, p)$

Suppose we estimate p using the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left[\frac{1}{n}S_n\right] = \frac{1}{n}\mathrm{E}(S_n) = \frac{1}{n}np = p$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left[\frac{1}{n}S_n\right] = \frac{1}{n^2}\mathrm{Var}(S_n) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

$$\lim_{n\to\infty}\left[\frac{p(1-p)}{n}\right] = 0$$

Random Sample

Random variables $X_1, X_2, \ldots, X_n$ are called a *random sample* (of size n) if

- X_i and X_j are independent when $i \neq j$
- the marginal pdf (pmf) of each X_i is the same function $f_X(x)$

Terminology:

- *mutually independent*: X_i, X_j independent when $i \neq j$
- *identically distributed*: two random variables have the same pdf (pmf)
- A random sample is also called an *iid* sample with pdf (pmf) $f_X(x)$

The joint density (mass) function of a random sample is given by

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) = \prod_{i=1}^{n} f_X(x_i)$$

Random Sample (continued)

If the population pdf (pmf) is a member of a parametric family, the joint pdf (pmf) is given by

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

For example, if f is a normal pdf then $\theta = (\mu, \sigma^2)$

Modeling assumption: the population distribution is a member of a known parametric family

- but the *true* value of θ is unknown

Idea: study how a random sample behaves for different populations by considering different values of θ

Outline

Definition of a Statistic

Let $X_1, \ldots, X_n$ be a random sample from a population

Let $T(x_1, \ldots, x_n)$ be a real-valued (or vector-valued) function

The random variable $Y = T(X_1, \ldots, X_n)$ is called a *statistic*

Terminology:

- The distribution of a statistic Y is called the *sampling distribution* of Y

Three Statistics

The *sample mean*:

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

The *sample variance*:

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

The *sample standard deviation*:

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

Unbiased Statistics

A statistic Y is an *unbiased estimator* of θ if $\mathsf{E}(Y) = \theta$

Example: Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ

$$\begin{aligned}
\mathsf{E}(\bar{X}) &= \mathsf{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] \\
&= \frac{1}{n}\mathsf{E}(X_1 + \cdots + X_n) \\
&= \frac{1}{n}\big(\mathsf{E}(X_1) + \cdots + \mathsf{E}(X_n)\big) \\
&= \frac{1}{n}(\overbrace{\mu + \cdots + \mu}^{n}) \\
&= \mu
\end{aligned}$$

Variance of the Sample Mean

Example: Let $X_1, \ldots, X_n$ be a random sample from a population with variance σ^2

$$
\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left[\frac{X_1 + \cdots + X_n}{n}\right] \\
&= \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) \\
&= \frac{1}{n^2}\big(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\big) \\
&= \frac{1}{n^2}(\overbrace{\sigma^2 + \cdots + \sigma^2}^{n}) \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

Why $n-1$?

Example: Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2

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

Why $n-1$? (continued)

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

Recall that: $E(U^2) = \mathrm{Var}(U) + [E(U)]^2$

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

Why $n-1$? (continued)

$$
\begin{aligned}
\mathrm{E}(S^2) &= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \\
&= \frac{1}{n-1}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) \\
&= \frac{1}{n-1}(n-1)\sigma^2 \\
&= \sigma^2
\end{aligned}
$$

Outline

Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of iid random variables

Implies that $\mathrm{E}(X_i) = \mu$ for all i

Let

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

Given $\epsilon > 0$,

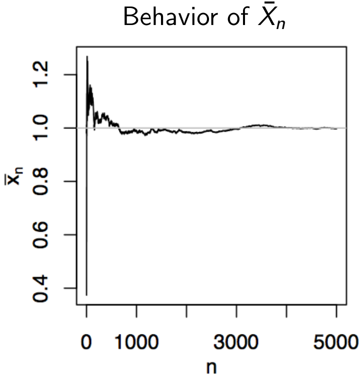
$$\mathrm{P}\left(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon\right) = 1$$

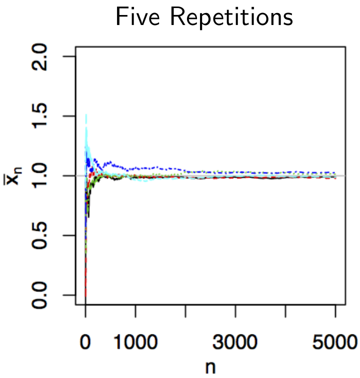
In other words, it is certain that $\bar{X}$ will be close to μ for large n

Further if $\mathrm{Var}(X_i)$ is finite and bounded, then $\mathrm{Var}(\bar{X}) \to 0$

Illustration of the Law of Large Numbers

Example for $X_i \sim \text{Normal}(1, 1^2)$





Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of *iid* random variables with

- $\mathrm{E}(X_i) = \mu$
- $\mathrm{Var}(X_i) = \sigma^2 < \infty$

Let $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$

Define $G_n(x)$ to be the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$

Then, for any $x \in (-\infty, \infty)$

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

In other words, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting standard normal distribution

Central Limit Theorem

Implications:

- The distribution of the sample mean is approximately normal for large values of n
- Normality comes from sums of *small* (i.e., finite variance), independent disturbances

Caveats:

- How large is large n? The Central Limit theorem does not tell us how good the approximation is
- The *goodness* of the approximation depends on the distribution of the population, hence it must be calculated on a case-by-case basis

Example

Let $X_1, \ldots, X_{100}$ be 100 independent Bernoulli(0.5) trials

What is the probability that $S_{100} = \sum_{i=1}^{100} X_i \geq 60$?

$$
\begin{aligned}
P(S_{100} \geq 60) &= \sum_{i=60}^{100} P(S_{100} = i) \\
&= \sum_{i=60}^{100} \binom{100}{i} 0.5^i (1 - 0.5)^{(100-i)} \\
&= \sum_{i=60}^{100} \binom{100}{i} 0.5^{100} \\
&\approx 0.0284
\end{aligned}
$$

Example

And again, using the Central Limit Theorem

$$P(S_n \geq 60) = P\left(\frac{S_n}{100} \geq \frac{60}{100}\right) = \mathrm{P}(\bar{X} \geq 0.6)$$

Since $X_i \sim \text{Bernoulli}(0.5)$

- $\mathrm{E}(X_i) = 0.5$
- $\mathrm{Var}(X_i) = 0.5(1 - 0.5) = 0.25$

$$\begin{aligned}
\mathrm{P}(\bar{X} \geq 0.6) &= \mathrm{P}\left(\frac{\bar{X} - 0.5}{\sqrt{0.25/100}} \geq \frac{0.6 - 0.5}{\sqrt{0.25/100}}\right) \\
&= \mathrm{P}\left(Z \geq 2\right) \\
&= 1 - \Phi(2) \\
&\approx 0.0228
\end{aligned}$$

Outline

Sampling from a Normal Population

Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population

The sample mean $\bar{X}$ and the sample variance S^2 are independent random variables

$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

$$\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2_{n-1}$$

If $U \sim \mathcal{N}(\mu, \sigma^2)$ and $V \sim \mathcal{N}(\gamma, \tau^2)$ are independent, then

$$Y = U + V \sim \mathcal{N}(\mu + \gamma, \sigma^2 + \tau^2)$$

Sampling from a Normal Population

- Let $X_1, \ldots, X_n$ be a random sample from a Normal(μ, σ^2) population
- The quantity
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
has a standard normal distribution
- But what do we do if σ^2 is unknown?
- Replacing σ^2 with the sample variance S^2 gives the statistic
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$
- What is the distribution of T?

Sampling from a Normal Population

- We can rewrite the expression for T as follows:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu)}{\frac{\sqrt{n}}{\sigma}\frac{S}{\sqrt{n}}} = \frac{\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}}{\sqrt{\frac{n-1}{n-1}\frac{S^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{U}{n-1}}}$$

- The numerator is distributed standard normal
- Since

$$\frac{(n-1)S^2}{\sigma^2} = U \sim \chi^2_{n-1}$$

  the denominator is the square root of a χ^2_{n-1} random variable divided by its degrees of freedom

- Z is a function of $\bar{X}$ and U is a function of S^2
  - Since $\bar{X}$ and S^2 are independent, so are Z and U

- $T \sim t_{n-1}$, that is, Student's t with $n-1$ degrees of freedom

COMPUTATIONAL FINANCE & RISK MANAGEMENT

UNIVERSITY *of* WASHINGTON

Department of Applied Mathematics

http://computational-finance.uw.edu