INTRODUTION TO STATISTICAL INFERENCE

Kakhramon Yusupov

IFPRI/WIUT Summer School 2015

REVIEW OF PROBABILITY

A continuous-valued random variable takes on a range of real values, e.g. x ranges from 0 to +∞.

Examples of continuous(-valued) random variables: income of a randomly selected consumer, time when a crisis ends, etc.

Thus, for a continuous random variable *X*, we can define its probability density function (pdf)

$$f_x(x) = F'_X(x) = \frac{dF_X(x)}{dx}$$

REVIEW OF PROBABILITY

Note that since $F_X(x)$ is non-decreasing in x we have

$$f_X(x) \geq 0 \qquad \text{for all x.}$$

From the Fundamental Theorem of Calculus, we have

$$F_X(x) = \int_{-\infty}^{x} f_x(x)dx$$

In particular,

$$\int_{-\infty}^{\infty} fx(x)dx = F_X(\infty) = 1$$

REVIEW OF PROBABILITY

Note that since $F_X(x)$ is non-decreasing in x we have

$$f_X(x) \geq 0 \qquad \text{for all x.}$$

From the Fundamental Theorem of Calculus, we have

$$F_X(x) = \int_{-\infty}^{x} f_x(x)dx$$

In particular,

$$\int_{-\infty}^{\infty} fx(x)dx = F_X(\infty) = 1$$

REVIEW OF PROBABILITY

More generally,

$$\int_a^b f_X(x)dx = F_X(b) - F_X(a) = P(a < X \leq b)$$

The expectation (average) of a continuous random variable *X* is given by

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx$$

Note that this is just the continuous equivalent of the discrete expectation

$$E(X) = \sum_{x=-\infty}^{\infty} x P_X(x)$$

STATISTICAL INFERENCE

The process of making guesses about the truth from a sample

- Truth is population parameters (non-observable)

- Sample parameters are calculated and observable

- Make guesses about the population

STATISTICS VS. PARAMETERS

Sample Statistic – any summary measure calculated from data; e.g., could be a mean, a difference in means or proportions, an odds ratio, or a correlation coefficient

- E.g., the mean income level in a sample of 1000 men is 810 thousand soums
- E.g., the correlation coefficient between educational attainment and earning power in the sample of 1000 men is 0.65

Population parameter – the true value/true effect in the entire population of interest

- E.g., the mean income level of all male workers is 825 thousand soums
- E.g., the correlation coefficient between educational attainment and earning power in of all male workers is 0.66

DISTRIBUTION OF A STATISTIC…

Statistics follow distributions too…

But the distribution of a statistic is a theoretical question.

Economists ask themselves: how much would the value of the statistic fluctuate if one could repeat a particular study over and over again with different samples of the same size?

By answering this question, economists are able to pinpoint exactly how much uncertainty is associated with a given statistic.

MATHEMATICAL THEORY… THE CENTRAL LIMIT THEOREM

If all possible random samples, each of size *n*, are taken from any population with a mean μ and a standard deviation σ, the sampling distribution of the sample means (averages) will:

1. have mean: $\mu_{\bar{x}} = \mu$

2. have standard deviation: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

3. be approximately normally distributed regardless of the shape of the parent population (normality improves with larger *n*). It all comes back to Z

MATHEMATICAL PROOF

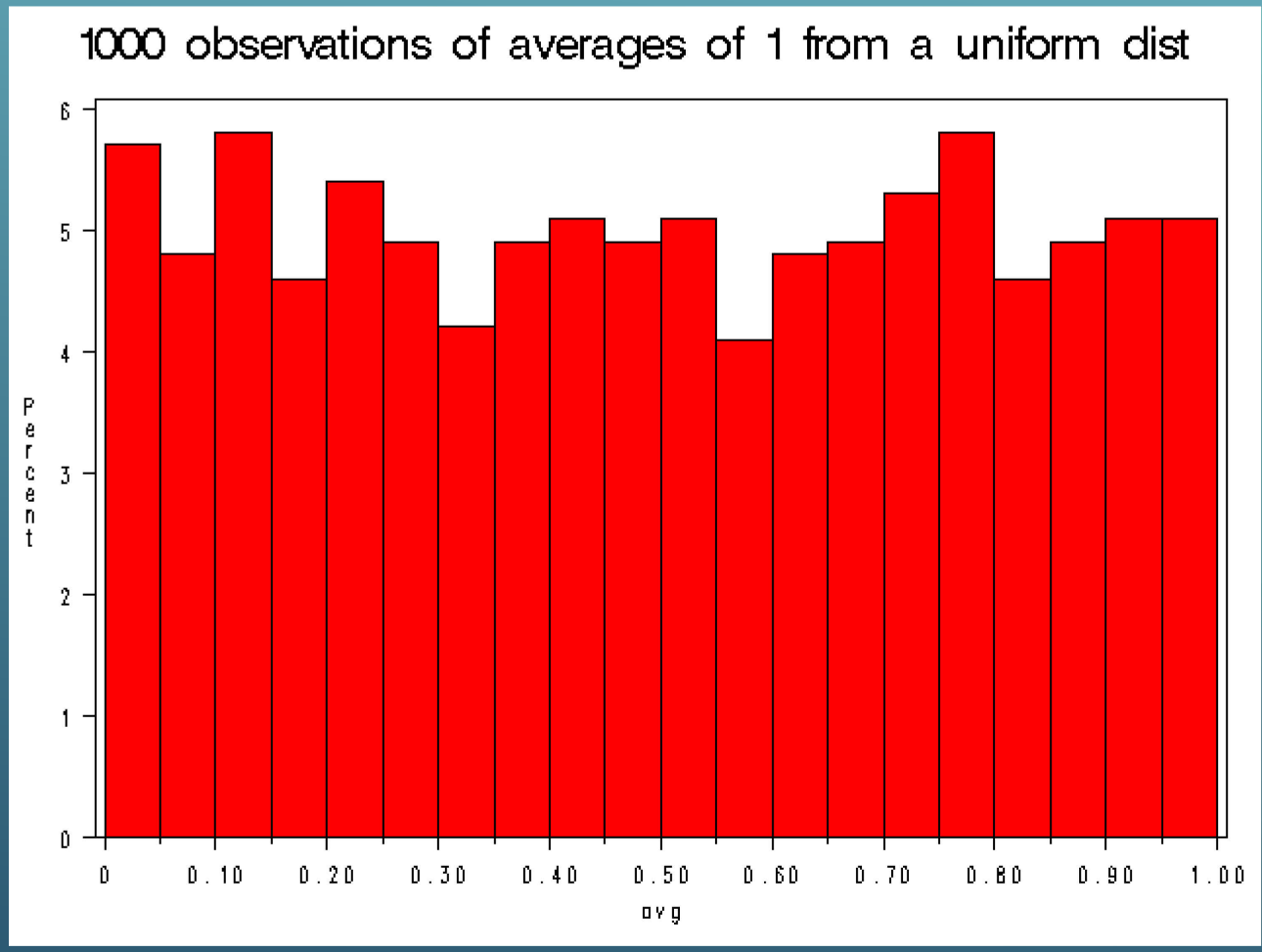
If *X* is a random variable from any distribution with known mean, E(x), and variance, Var(x), then the expected value and variance of the average of *n* observations of X is:

$$E(\overline{X}_n) = E(\frac{\sum_{i=1}^{n} x_i}{n}) = \frac{\sum_{i=1}^{n} E(x)}{n} = \frac{nE(x)}{n} = E(x)$$

$$Var(\overline{X}_n) = Var(\frac{\sum_{i=1}^{n} x_i}{n}) = \frac{\sum_{i=1}^{n} Var(x)}{n^2} = \frac{nVar(x)}{n^2} = \frac{Var(x)}{n}$$

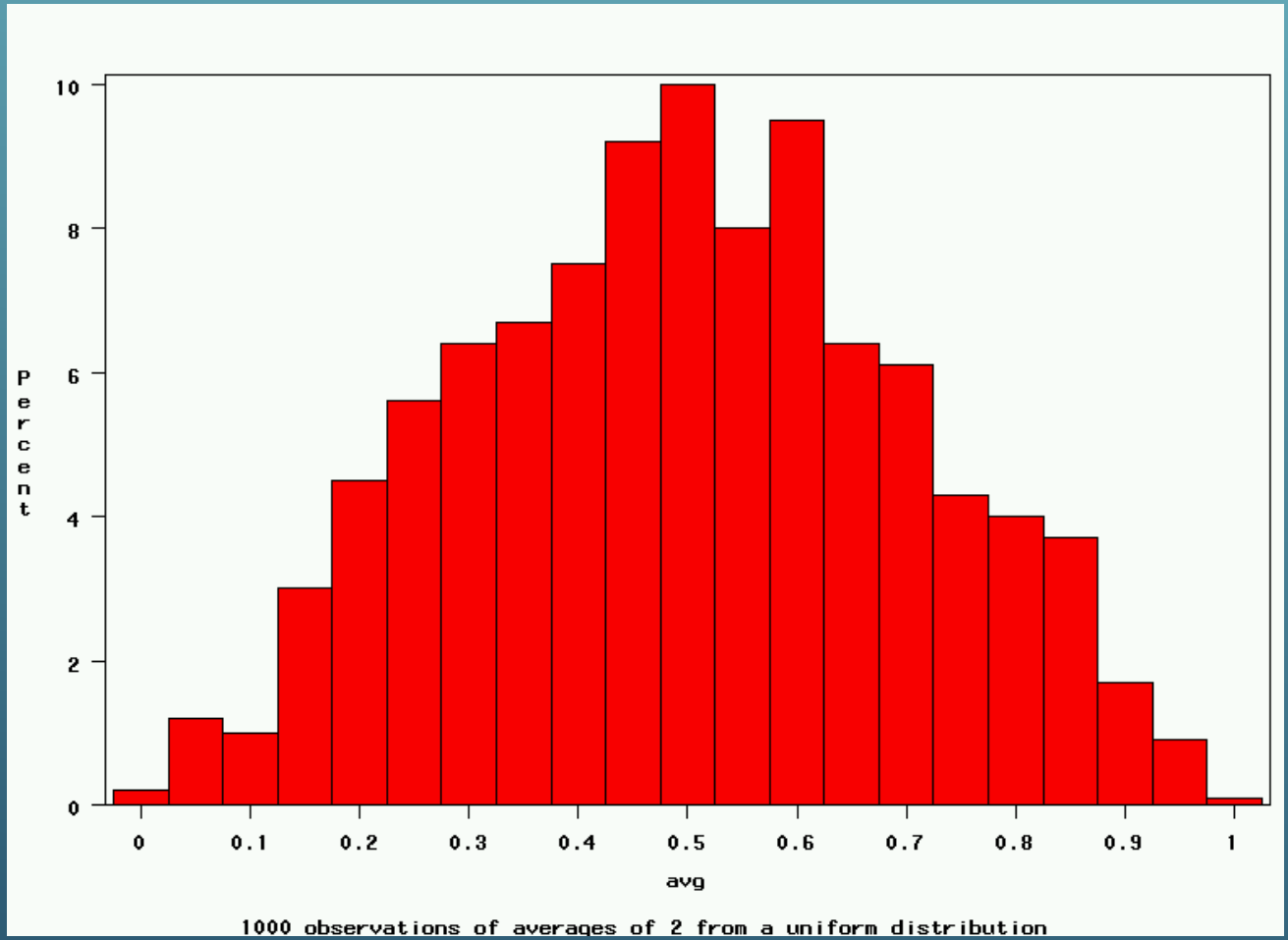
COMPUTER SIMULATION OF THE CLT

Uniform on [0,1]: average of 1 (original distribution)



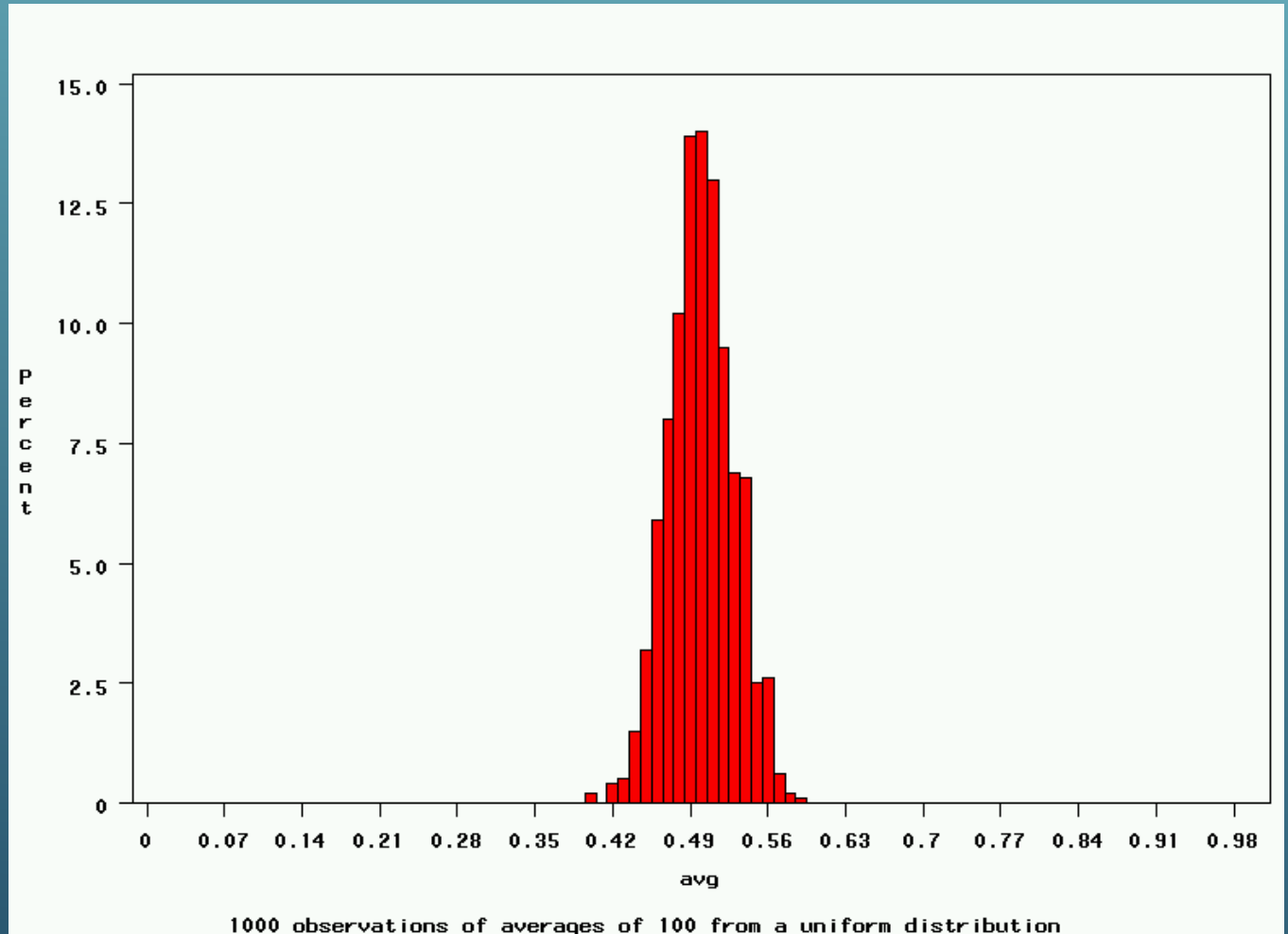
COMPUTER SIMULATION OF THE CLT

Uniform: 1000 averages of 2



COMPUTER SIMULATION OF THE CLT

Uniform: 1000 averages of 100



HYPOTHESIS TESTING (LARGE SAMPLE)

Hypothesis test:

$$Z = \frac{\text{observed mean} - \text{null mean}}{\frac{s}{\sqrt{n}}}$$

Confidence Interval

$$\text{confidence interval} = \text{observed mean} \pm Z_{\alpha/2} * (\frac{s}{\sqrt{n}})$$

HYPOTHESIS TESTING (SMALL SAMPLE)

Hypothesis test:

$$T_{n-1} = \frac{\text{observed mean} - \text{null mean}}{\frac{s}{\sqrt{n}}}$$

Confidence Interval

$$\text{confidence interval} = \text{observed mean} \pm \mathrm{T}_{n-1,\alpha/2} * (\frac{s}{\sqrt{n}})$$

95% CONFIDENCE INTERVAL

Goal: capture the true effect (e.g., the true mean) most of the time.

A 95% confidence interval should include the true effect about 95% of the time.

A 99% confidence interval should include the true effect about 99% of the time.

Confidence Intervals give:

- A plausible range of values for a population parameter.

- The precision of an estimate.(When sampling variability is high, the confidence interval will be wide to reflect the uncertainty of the observation.)

- Statistical significance (if the 95% CI does not cross the null value, it is significant at .05)

TESTING HYPOTHESES

The Steps:

1. Define your hypotheses (null, alternative)
2. Specify your null distribution
3. Collect sample data
4. Calculate the sample moments and test statistic
5. Reject or fail to reject (~accept) the null hypothesis

FORMAL HYPOTHESIS TEST

1. Null hypothesis: r=0

   Alternative: $r \neq 0$ (two-sided)

2. Determine the null distribution

   Normally distributed

   Standard error = 0.1

3. Collect Data, r=0.35

4. Calculate the p-value for the data:

$$Z = \frac{0.35 - 0}{.1} = 3.5$$

5. Reject or fail to reject the null (fail to reject)

P-VALUE

p-value is the probability of obtaining the observed sample results when the null hypothesis is actually true

Small p-values mean the null value is unlikely given our data.

Our data are so unlikely given the null hypothesis that we would rather reject the null hypothesis

P-VALUE

By convention, p-values of <.05 are often accepted as “statistically significant” in the economic literature; but this is an arbitrary cut-off.

A cut-off of p<.05 means that in about 5 of 100 experiments, a result would appear significant just by chance (“Type I error”).

TYPE I AND TYPE II ERRORS

	STATE OF NATURE	
Decision	*Ho is true*	*Ho is false*
Reject	Type I error	Correct
Do not reject	Correct	Type II error