STA261: Lecture 1

Introduction and Review

Alex Stringer

July 4th, 2018

Welcome

- Previous course (STA257): Introduction to probability
- This course (STA261): Introduction to statistics, mainly estimation theory and hypothesis testing
- This week: Brief review of STA257; detailed review of limit theorems and convergence of random variables

Disclaimer

The materials in these slides are intended to be a companion to the course textbook, *Mathematical Statistics and Data Analysis, Third Edition*, by John A Rice. Material in the slides may or may not be taken directly from this source. These slides were organized and typeset by Alex Stringer.

A big thanks to Jerry Brunner as well for providing inspiration for assignment questions.

License

Course Information

- Instructor (me!): Alex Stringer
- Course Website: q.utoronto.ca
- Piazza: piazza.com/configure-classes/summer2018/sta261

The syllabus is posted on Quercus.

Review: Probability

Recall: a **random variable** X is a function from a **sample space** Ω to (possibly a subset of) the real numbers. The subset of the reals to which X maps is referred to as the **support** of X.

The probability distribution of X represents the assignment of probability measure to events in the sample space.

- Discrete:
  - $p(x) = P(X = x)$
  - $F(x) = P(X \leq x) = \sum_{a=1}^{x} p(a)$
- Continuous:
  - $F(x) = P(X \leq x)$
  - $f(x) = \partial F / \partial x$
- General:
  - $F(A) = P(X \in A) = \int_{x \in A} dF(x)$

Review: Expectation

The **expected value**, **expectation**, or **mean** of a random variable X is defined as:

$$E(X) = \int_x x dF(x)$$

where the integral is taken across the entire support of X. Specifically,

- Discrete: $E(X) = \sum_x xP(X = x)$
- Continuous: $E(X) = \int_x xf(x)dx$

Expectation is a linear operator, satisfying $E(aX + b) = aE(X) + b$. The expectation of a function $g(X)$ is obtained by plugging $g(X)$ in for X in the above definition. $E(g(X)) \neq g(E(X))$ unless g is linear.

Review: Standard Deviation and Variance

The **standard deviation** of a random variable is the Euclidean distance from the random variable to its mean:

$$SD(X) = \sqrt{E\Big[(X - E(X))^2\Big]}$$

Numerical values of the standard deviation, computed for actual data, will have the same metric units as X, which is convenient for interpretation and communication.

Often, for mathematical tractability, we work with the variance, which is the squared standard deviation:

$$Var(X) = SD(X)^2 = E\Big[(X - E(X))^2\Big]$$

Review: Moment-Generating Functions

The **moment-generating function** of X is defined as

$$M_X(t) = E(e^{tX})$$

This has two major uses in mathematical statistics:

- Computing moments: $E(X^k) = M_X^{(k)}(0)$
- The fact that $X \overset{d}{=} Y \iff M_X(t) = M_Y(t) \forall t$ gives us a really convenient way of asserting that two random variables have the same distribution

Recall: two random variables are equal in distribution, $X \overset{d}{=} Y$, if and only if their distribution functions are equal at all points in their support, $F_X(x) = F_Y(x) \forall x$.

Review: Inequalities

If X is a random variable with $E(X) < \infty$ and $Var(X) < \infty$, then we have:

Chebyshev: $P(|X - E(X)| > t) \leq Var(X)/t^2$ for any $t > 0$

Markov: if X is nonnegative with probability 1 in addition to the above criteria, then $P(X \geq t) \leq E(X)/t$ for any $t > 0$

Sequence of Random Variables

Recall: a *sequence* of random variables is a set of random variables indexed by some natural number n. We write $\{X_n\}_{n=1}^{N} = (X_1, X_2, \ldots, X_N)$.

In general, the sequence may be infinite, but in this course it will always be *countable*.

The order may or may not matter.

n is often the sample size of an experiment.

Example

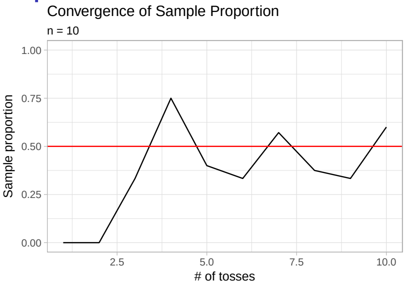
Example: let $X_i = 1$ if the flip of a fair coin yields heads, and we flip the coin n times. Then the random variables $(X_1, X_2, \ldots, X_n)$ form an unordered binary sequence- but a *random* one. This sequence itself has a probability distribution, equal to the joint distribution of the X_i.

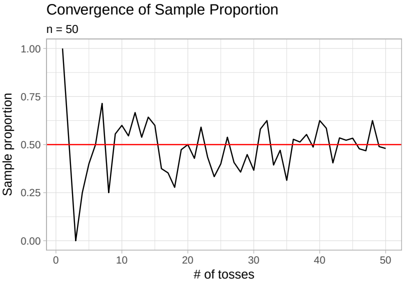
Example

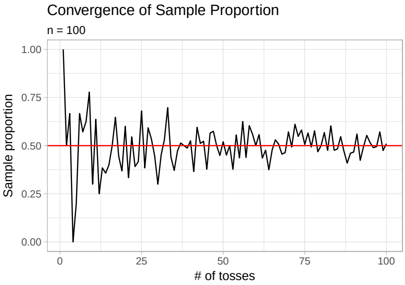
Consider the sequence $S_n = (1/1 \times X_1, 1/2 \times (X_1 + X_2), \ldots, 1/n \times \sum_{i=1}^{n} X_n)$. We can write this more succinctly as $S_n = \left\{ \frac{1}{i} \sum_{j=1}^{i} X_j \right\}_{i=1}^{n}$. This is the sequence of sample proportions of heads obtained by flipping the coin $1, 2, ..., n$ times.

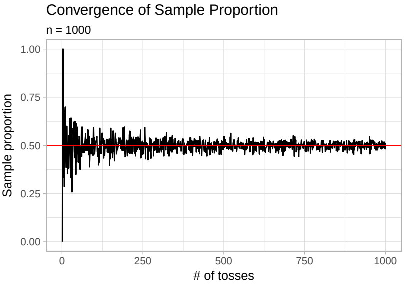
As $n \to \infty$, our intuition tells us that the tail of S_n should get closer and closer to $1/2$, the true population proportion of heads. But for any finite n, each element of S_n is still a random variable. So, we would expect the tail to fluctuate about $1/2$, but less and less as $n \to \infty$.

Example


Convergence of Sample Proportion
n = 10
Sample proportion
0.00
0.25
0.50
0.75
1.00
2.5
5.0
7.5
10.0
# of tosses


Convergence of Sample Proportion
n = 50
Sample proportion
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50
# of tosses


Convergence of Sample Proportion
n = 100
Sample proportion
0.00
0.25
0.50
0.75
1.00
0
25
50
75
100
# of tosses


Convergence of Sample Proportion
n = 1000
Sample proportion
0.00
0.25
0.50
0.75
1.00
0
250
500
750
1000
# of tosses

Convergence in Probability (textbook, page 178)

We can formalize our intuition as follows. Let $\{Z_n\}$ be any sequence of random variables, and let $\mu \in \mathbb{R}$ be any scalar. Then we say that the sequence $\{Z_n\}$ **converges in probability** to μ, $Z_n \xrightarrow{p} \mu$, if for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|Z_n - \mu| > \epsilon) = 0$$

This means no matter how close we want the sequence to get to μ, we can always pick an n so that the probability of any further term being farther away is as small as we want.

Convergence in Probability (textbook, page 178)

This can be defined for convergence to a random variable: we say X_n converges in probability to X if $X_n - X \xrightarrow{p} 0$.

Example: Coin Toss

In the example above, we actually observe the whole sequence $\left\{\bar{X}_i\right\}_{i=1}^n$. We throw the coin once and observe $\bar{X}_1$, throw it again and observe $\bar{X}_2$, and so forth.

Let's say we want to be really sure that our $\bar{X}_n$ is within $\epsilon = 0.001$ of the true population proportion, $1/2$. That is, we want $P\left(\left|\bar{X}_n - 1/2\right| > 0.001\right)$ to be small. We might hope that we can pick n to make that probability as small as we want.

We might hope that we can do this for $\epsilon = 0.0001$, $\epsilon = 0.00001$, or any arbitrarily small ϵ.

Example: Sample Size

Example: consider an experiment in which we measure a single quantity X_i on n individuals, $i = 1 \ldots n$. Increasing n corresponds to increasing the size of the experiment. "As $n \to \infty$" means "as we make our sample bigger and bigger and bigger".

The sequence $S_n = \left\{ \bar{X}_i \right\}_{i=1}^{n}$ is an abstract *idea*- in practice we only actually pick one n and then compute that $\bar{X}_n$.

Thinking about this theoretical sequence of random variables that *might* have been observed at any n lets us study what happens as we make the sample size bigger.

Testing Convergence

In practice, that limit is often difficult to evaluate, so we have the following theorem:

Theorem: Suppose $\{Z_n\}$ is a sequence of random variables with $E(Z_n) = \mu$ and $\lim_{n\to\infty} Var(Z_n) = 0$. Then $Z_n \xrightarrow{p} \mu$.

Proof: $\ldots$

Law of Large Numbers (textbook, page 178)

We can now state and prove the weak law of large numbers.

Theorem: Suppose $\{X_n\}$ is a sequence of *independent* random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Let $\bar{X}_n = (1/n) \times \sum_{i=1}^{n} X_i$. Then $\bar{X}_n \xrightarrow{p} \mu$.

Proof: $E(\bar{X}_n) = \mu$, $Var(\bar{X}_n) = \sigma^2/n$ (this is where independence is used), so the result follows immediately from the theorem on the last slide.

Applications and Limitations

The major application of this was hinted in the previous example (with the coin tosses): if we increase the sample size enough, we can be sure that we get an estimate of the population mean that is "close enough". This has direct applications in, for example, monte carlo integration (textbook, page 179).

How do we tell whether we are close enough, in an actual experiment? We still haven't made any assumptions about the distribution of the X_i, or about $\bar{X}_n$, so we can't make any probability statements. In particular, we can't actually evaluate $P(|\bar{X}_n - \mu| > \epsilon)$ for any specific (n, ϵ).

Tosses of a Fair Coin

We saw previously that if the $\bar{X}_n$ represents the sample average number of heads in n tosses of a fair coin, then as $n \to \infty$, $\bar{X}_n$ gets close to $E(X)$ in probability. That is, by the LLN, $\bar{X}_n \xrightarrow{p} 1/2$. Can we say anything about the manner in which $\bar{X}_n$ fluctuates about its mean?

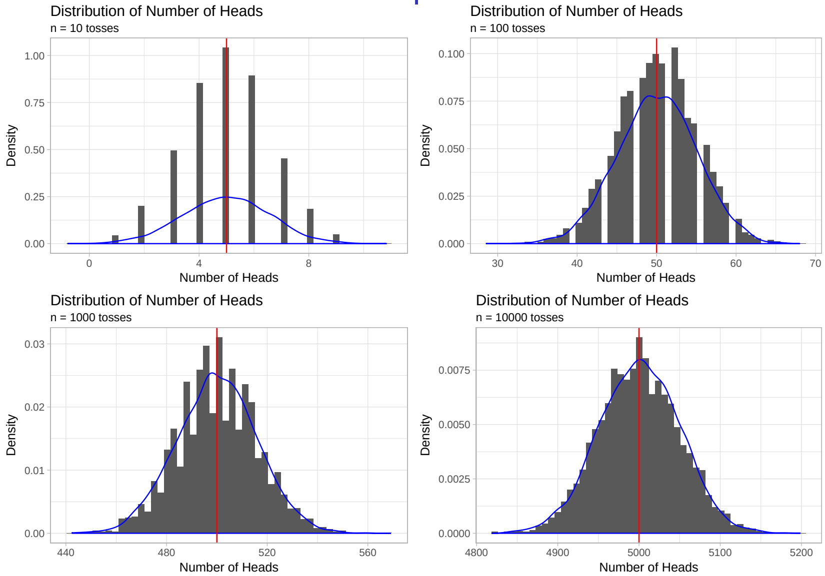
Can we evaluate probability statements, like $P(0.4 < \bar{X}_n < 0.6)$ or $P(0.49 < \bar{X}_n < 0.51)$, for any actual n?

Distribution of a Sum of Independent Random Variables

Let X_i represent a single toss of a fair coin, taking value 1 if heads and 0 else. Let $S_n = \sum_{i=1}^{n} X_i$, which is just the number of heads observed in the n tosses. S_n is a random variable for any finite n. Can we say anything about its probability distribution?

Let's look at some simulated experiments.

Distribution of a Sum of Independent Random Variables



Convergence in Distribution

Let $\{X_n\}$ be a sequence of random variables with corresponding distribution functions $\{F_n(x)\}$, and let X be a random variable with cdf $F_X(x)$. We say that the sequence $\{X_n\}$ **converges in distribution** to X, $X_n \xrightarrow{d} X$, if $\lim_{n\to\infty} F_n(x) = F_X(x)$ for all x at which these distribution functions are continuous.

This also works for moment-generating functions:
$\lim_{n\to\infty} M_n(t) = M_X(t) \forall t \implies X_n \xrightarrow{d} X$.

The Central Limit Theorem (textbook, pg 184)

Let $\{X_n\}$ be a sequence of **independent** random variables each with **mean 0** and common variance σ^2. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{S_n}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$$

The Central Limit Theorem (textbook, pg 184)

Proof: The proof relies on computing the moment-generating function of $Z_n = \frac{S_n}{\sigma\sqrt{n}}$ and showing that it goes to the mgf of the standard normal distribution. We have

$$M_{Z_n}(t) = \left[M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

Consider the taylor expansion of $M_X(t)$ about $t = 0$:

$$\begin{aligned} M(s) &= M(0) + sM'(0) + (1/2)s^2 M''(0) + \epsilon_s \\ &= 1 + (1/2)\sigma^2 s^2 + \epsilon_s \end{aligned}$$

where $\epsilon_s \to 0$ as $s \to 0$ and we have used the fact that $M(0) = 1, M'(0) = 0$ and $M''(0) = \sigma^2$.

The Central Limit Theorem (textbook, pg 184)

Hence,

$$M_{Z_n}(t) = \left(1 + \frac{t^2/2}{n} + \epsilon_n\right)^n \to e^{t^2/2}$$

as $n \to \infty$. This is the mgf of a standard normal random variable.

Application

The CLT can (and should!) be used to evaluate probability statements like the ones shown previously.

We know, in our previous example of coin tosses, that $S_n \sim Bin(n, 1/2)$ exactly. We can compare the actual probabilities defined by the binomial distribution to the approximations obtained from the CLT. We have for $n = 100$ (for example):

$$
\begin{aligned}
P(0.4 < \bar{X}_{100} < 0.6) &= P(40 < S_{100} < 60) \\
&= \sum_{i=40}^{60} P(S_{100} = i) \\
&= 0.9540
\end{aligned}
$$

Application

We can approximate this using the CLT.

Recall that $E(S_n) = np$, $Var(S_n) = np(1-p)$. For $n = 100, p = 1/2$, we have $E(S_{100}) = 50$ and $Var(S_{100}) = 25$, so

$$\begin{aligned}
P(0.4 < \bar{X}_{100} < 0.6) &= P(40 < S_{100} < 60) \\
&= P\left(\frac{40-50}{\sqrt{25}} < \frac{S_{100}-50}{\sqrt{25}} < \frac{60-50}{\sqrt{25}}\right) \\
&= P\left(-2 < Z_n < 2\right) \\
&\approx P(-2 < Z < 2), Z \sim N(0,1) \\
&= \Phi(2) - \Phi(-2) \\
&= 0.9545
\end{aligned}$$