STA261 Lecture 1 — 2017-07-05

Neil Montgomery

Last edited: 2017-07-05 18:58

admin

contact, notes

date format	YYYY-MM-DD – *All Hail ISO8601!!!*
instructor	Neil Montgomery
email	`neilmontg@gmail.com`
office	TBA
office hours	Monday and Wednesday 17:00 – 18:00
website	portal (announcements, grades, suggested exercises, etc.)
github	`https://github.com/sta261-summer-2017` (lecture material, code, etc.)

evaluation, book(s), tutorials

what	when	how much
midterm 1	Probably 2017-07-19 18:00 to 20:00 (short lecture after)	25%
midterm 2	Probably 2017-08-02 18:00 to 20:00 (short lecture after)	25%
exam	TBA	50%

Book: Mathematical Statistics and Data Analysis, 3rd ed. by John Rice

Tutorials start Monday.

propositions, and theorems

There is actually no difference, aside from the style of reserving the word "theorem" to results that are more important, for some reason.

propositions, and theorems

There is actually no difference, aside from the style of reserving the word "theorem" to results that are more important, for some reason.

In this course, every lecture will have exactly one result which I will call a **Theorem**.

propositions, and theorems

There is actually no difference, aside from the style of reserving the word "theorem" to results that are more important, for some reason.

In this course, every lecture will have exactly one result which I will call a **Theorem**.

The importance of a **Theorem** to you is that each test will ask you to prove one of the theorems from the previous four lectures. The final exam will ask you to prove one of the theorems from the entire course.

what we’re going to do, and why

probability versus statistics

The important objects from STA257:

- Random variable

probability versus statistics

The important objects from STA257:

- Random variable
- Distribution

probability versus statistics

The important objects from STA257:

- Random variable
- Distribution

probability versus statistics

The important objects from STA257:

- Random variable
- Distribution

You learned about several families of distributions, discrete and continuous. The specific family member was identified by one or more *parameters*.

probability versus statistics

The important objects from STA257:

- Random variable
- Distribution

You learned about several families of distributions, discrete and continuous. The specific family member was identified by one or more *parameters*.

In *statistics* we don't know the parameter values, so we'll (...imagine we can...) use a *dataset* to make *inferences* about the parameter values.

mathematical model for the idea of *sample*

In this course a *sample* is defined as a sequence of random variables $X_1, X_2, \ldots, X_n$ which are:

- **i**ndependent

mathematical model for the idea of *sample*

In this course a *sample* is defined as a sequence of random variables $X_1, X_2, \ldots, X_n$ which are:

- **i**ndependent
- come from the same distribution, also known as "**i**dentically **d**istributed".

mathematical model for the idea of *sample*

In this course a *sample* is defined as a sequence of random variables $X_1, X_2, \ldots, X_n$ which are:

- **i**ndependent
- come from the same distribution, also known as "**i**dentically **d**istributed".

mathematical model for the idea of *sample*

In this course a *sample* is defined as a sequence of random variables $X_1, X_2, \ldots, X_n$ which are:

- **i**ndependent
- come from the same distribution, also known as "**i**dentically **d**istributed".

Abbreviation: **i.i.d.**

mathematical model for the idea of *sample*

In this course a *sample* is defined as a sequence of random variables $X_1, X_2, \ldots, X_n$ which are:

- **i**ndependent
- come from the same distribution, also known as "**i**dentically **d**istributed".

Abbreviation: **i.i.d.**

We might refer to a "parent" or "population" random variable X, with some distribution we might call an "underlying distribution", and the sample is considered to be i.i.d. "replications" of X.

data, dataset

The most common dataset is in the form of a rectangle, made up of *variables* and *observations*.

data, dataset

The most common dataset is in the form of a rectangle, made up of *variables* and *observations*.

The columns are called the *variables*. Every element of a variable will be of the same “type” (numerical, categorical, free form text.)

data, dataset

The most common dataset is in the form of a rectangle, made up of *variables* and *observations*.

The columns are called the *variables*. Every element of a variable will be of the same "type" (numerical, categorical, free form text.)

The rows are the *observations*. The number of rows is the *sample size*.

data, dataset

The most common dataset is in the form of a rectangle, made up of *variables* and *observations*.

The columns are called the *variables*. Every element of a variable will be of the same "type" (numerical, categorical, free form text.)

The rows are the *observations*. The number of rows is the *sample size*.

The way the dataset was collected will dictate the method of analysis.

probabilistic model for the notion of “dataset” - I

One model for this prospective dataset can be to consider it as a mix of columns of length n where some (or all) of the columns are random.

probabilistic model for the notion of “dataset” - I

One model for this prospective dataset can be to consider it as a mix of columns of length n where some (or all) of the columns are random.

A random column is headed by random variable (with “the underlying distribution”), and the contents are random variable “copies” of that underlying distribution.

probabilistic model for the notion of "dataset" - I

One model for this prospective dataset can be to consider it as a mix of columns of length n where some (or all) of the columns are random.

A random column is headed by random variable (with "the underlying distribution"), and the contents are random variable "copies" of that underlying distribution.

There could be non-random columns with categorical or numerical information.

probabilistic model for the notion of “dataset” - II

probabilistic model for the notion of “dataset” - II

X
X_1
X_2
X_3
X_4
$\vdots$
X_n

probabilistic model for the notion of “dataset” - II

X	Y
X_1	Y_1
X_2	Y_2
X_3	Y_3
X_4	Y_4
$\vdots$	$\vdots$
X_n	Y_n

probabilistic model for the notion of “dataset” - II

"Subject ID"	X	Y
ID345	X_1	Y_1
ID952	X_2	Y_2
ID826	X_3	Y_3
ID118	X_4	Y_4
$\vdots$	$\vdots$	$\vdots$
ID503	X_n	Y_n

probabilistic model for the notion of “dataset” - II

"Subject ID"	X	Y	"Group ID"
ID345	X_1	Y_1	A
ID952	X_2	Y_2	A
ID826	X_3	Y_3	B
ID118	X_4	Y_4	B
$\vdots$	$\vdots$	$\vdots$	$\vdots$
ID503	X_n	Y_n	A

probabilistic model for the notion of “dataset” - II

"Subject ID"	X	Y	"Group ID"	"InputVariable"
ID345	X_1	Y_1	A	w_1
ID952	X_2	Y_2	A	w_2
ID826	X_3	Y_3	B	w_3
ID118	X_4	Y_4	B	w_4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ID503	X_n	Y_n	A	w_n

a snippet of real dataset from the wild

Ident	Date	WorkingAge	TakenBy	Fe	Al
448574	11579	3112	EMPL_9134	11	6
448574	11082	1562	EMPL_0592	15	6
448576	12435	5124	EMPL_0917	8	2
448579	11894	1208	EMPL_4925	10	4
448590	11005	210	EMPL_9375	11	6
448595	11039	170	EMPL_0917	12	3
448572	11434	2436	EMPL_0592	13	7
448594	11292	1230	EMPL_0917	24	5

inference on unknown parameter values

So the basic idea will be to have a population X from a distribution with one or more unknown parameter values.

inference on unknown parameter values

So the basic idea will be to have a population X from a distribution with one or more unknown parameter values.

We'll (imagine...) gathering a sample $X_1, \ldots, X_n$ from this distribution, and using one or more functions of this sample to "guess" values for the unknown parameters.

inference on unknown parameter values

So the basic idea will be to have a population X from a distribution with one or more unknown parameter values.

We'll (imagine...) gathering a sample $X_1, \ldots, X_n$ from this distribution, and using one or more functions of this sample to "guess" values for the unknown parameters.

A function of a sample is called a **statistic**.

inference on unknown parameter values

So the basic idea will be to have a population X from a distribution with one or more unknown parameter values.

We'll (imagine...) gathering a sample $X_1, \ldots, X_n$ from this distribution, and using one or more functions of this sample to "guess" values for the unknown parameters.

A function of a sample is called a **statistic**.

Many parameters have a close connection to $E(X)$ or $\mathrm{Var}(X)$. It will turn out that $\overline{X}$ and S^2 play a large role in statistical inference.

inference on unknown parameter values

So the basic idea will be to have a population X from a distribution with one or more unknown parameter values.

We'll (imagine...) gathering a sample $X_1, \ldots, X_n$ from this distribution, and using one or more functions of this sample to "guess" values for the unknown parameters.

A function of a sample is called a **statistic**.

Many parameters have a close connection to $E(X)$ or $\text{Var}(X)$. It will turn out that $\overline{X}$ and S^2 play a large role in statistical inference.

This explains the centrality of the normal distributions to statistics.

back to probability - distributions of functions of random variables

single variable case - a general formula

You might (or might not, which is fine!) recall from STA257, when you have a random variable X with density $f_X(x)$, and a monotone, differentiable function g, the density of $Y = g(X)$ is:

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|$$

single variable case - a general formula

You might (or might not, which is fine!) recall from STA257, when you have a random variable X with density $f_X(x)$, and a monotone, differentiable function g, the density of $Y = g(X)$ is:

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|$$

Note 1.0: This result is really not all that grand, or mysterious. It is just an application of the "change of variables" or "substitution" strategy from single variable calculus.

bivariate transformations - I

A joint density $f(x, y)$ for random variables X and Y likes to live inside a two dimensional integral.

bivariate transformations - I

A joint density $f(x, y)$ for random variables X and Y likes to live inside a two dimensional integral.

At some point in MAT237 you will (or will have) learned how to change both variables in such an integral. We'll use this technique in STA261 for a few specific cases.

bivariate transformations - I

A joint density $f(x, y)$ for random variables X and Y likes to live inside a two dimensional integral.

At some point in MAT237 you will (or will have) learned how to change both variables in such an integral. We'll use this technique in STA261 for a few specific cases.

We might want to define random variables U and V in terms of X and Y, in general as:

$$U = g_1(X, Y)$$
$$V = g_2(X, Y)$$

Question: what is the joint density for U and V?

bivariate transformations - II

In the simplest case the transformation is smooth and invertible, so that one can determine inverse functions:

$$X = h_1(U, V)$$
$$Y = h_2(U, V)$$

bivariate transformations - II

In the simplest case the transformation is smooth and invertible, so that one can determine inverse functions:

$$X = h_1(U, V)$$
$$Y = h_2(U, V)$$

The "differential" term $g^{-1}(y)$ in the single variable case is played by the *Jacobian*, which is the determinant of the matrix of partial derivatives.

Jacobian, and density formula

$$J = \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix}$$

Jacobian, and density formula

$$J = \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix}$$

The joint density of U and V is then given by:

$$f_{U,V}(u, v) = f_{X,Y}\left(h_1(u, v), h_2(u, v)\right)|J|$$

Jacobian, and density formula

$$J = \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix}$$

The joint density of U and V is then given by:

$$f_{U,V}(u,v) = f_{X,Y}\left(h_1(u,v), h_2(u,v)\right)|J|$$

Examples are easier than the formula itself!

bivariate transformation examples

Example 1.1: Suppose X and Y are independent with $N(0,1)$ distributions. What is the joint density of $U = X + Y$ and $V = X - Y$?

bivariate transformation examples

Example 1.1: Suppose X and Y are independent with $N(0,1)$ distributions. What is the joint density of $U = X + Y$ and $V = X - Y$?

Example 1.2: Suppose X and Y are independent with joint density $f_{X,Y}(x,y)$. What is the density of $U = X + Y$?

In this example we only have a $g_1(x,y)$. The technique is to add your own $g_2(x,y)$, and find the relevant marginal at the end.

multivariate transformations - we'll do this only once!

This technique works to transform any number of $X_1, \ldots, X_n$ into the same number $U_1, \ldots, U_n$ using functions $g_1, \ldots, g_n$.

multivariate transformations - we'll do this only once!

This technique works to transform any number of $X_1, \ldots, X_n$ into the same number $U_1, \ldots, U_n$ using functions $g_1, \ldots, g_n$.

Find the inverse transformations $h_1, \ldots, h_n$. The Jacobian is now:

$$J = \begin{vmatrix} \frac{\partial h_1}{\partial u_1} & \cdots & \frac{\partial h_1}{\partial u_n} \\ \vdots & & \vdots \\ \frac{\partial h_n}{\partial u_1} & \cdots & \frac{\partial h_n}{\partial u_n} \end{vmatrix}$$

and the new density is:

$$f_{U_1,\ldots,U_n}(u_1, \ldots, u_n) = f_{X_1,\ldots,X_n}(h_1(u_1, \ldots, u_n), \ldots, h_n(u_1, \ldots, u_n))|J|$$

other techniques for functions of random variables

Those general formulae are used as a last resort.

We will also make free use of moment generating functions:

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

and their most important properties, which are:

1. If the moment generating function exists, it uniquely describes the distribution.

other techniques for functions of random variables

Those general formulae are used as a last resort.

We will also make free use of moment generating functions:

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

and their most important properties, which are:

1. If the moment generating function exists, it uniquely describes the distribution.
2. If $X \perp Y$, then $M_{X+Y}(t) = M_X(t)M_Y(t)$

distributions of statistics from normal samples

motivation

We tend to be interested in the unknown mean of a distribution, with only a sample $X_1, \ldots, X_n$ to work with.

motivation

We tend to be interested in the unknown mean of a distribution, with only a sample $X_1, \ldots, X_n$ to work with.

Lots of things actually have normal distributions, so learning about distributions of functions of normal samples is a good idea.

motivation

We tend to be interested in the unknown mean of a distribution, with only a sample $X_1, \ldots, X_n$ to work with.

Lots of things actually have normal distributions, so learning about distributions of functions of normal samples is a good idea.

More crucially, even if we don't know what the underlying distribution is, things like $\overline{X}$ will still be approximately normal, due to the speed of convergence of the central limit theorem.

motivation

We tend to be interested in the unknown mean of a distribution, with only a sample $X_1, \ldots, X_n$ to work with.

Lots of things actually have normal distributions, so learning about distributions of functions of normal samples is a good idea.

More crucially, even if we don't know what the underlying distribution is, things like $\overline{X}$ will still be approximately normal, due to the speed of convergence of the central limit theorem.

It also turns out the central limit theorem has some friends that also converge quickly in practice.

sums of i.i.d. normals, and variations

Proposition 1.3: If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then:

1. $\sum X_i \sim N(n\mu, n\sigma^2)$

sums of i.i.d. normals, and variations

Proposition 1.3: If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then:

1. $\sum X_i \sim N(n\mu, n\sigma^2)$
2. $\overline{X} \sim N(\mu, \sigma^2/n)$

sums of i.i.d. normals, and variations

Proposition 1.3: If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then:

1. $\sum X_i \sim N(n\mu, n\sigma^2)$
2. $\overline{X} \sim N(\mu, \sigma^2/n)$
3. $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

squares of i.i.d. standard normals, and their sums

Straightforward result from STA257: $Z \sim N(0,1)$ then $Z^2 \sim \chi^2_1$.

χ^2_ν is a nickname for a Gamma$\left(\frac{\nu}{2}, \frac{1}{2}\right)$ distribution, which has density and m.g.f. respectively:

$$f(x) = \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} \qquad M(t) = (1-2t)^{-\nu/2}$$

Proposition 1.4: If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then:

$$\sum \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_n$$

another chi_ν^2 result

Proposition 1.5: If X and Y are independent with $X \sim \chi_n^2$ and $X + Y \sim \chi_{n+m}^2$, then

$$Y \sim \chi_m^2$$

t distributions

Theorem 1.6: Suppose $Z \sim N(0,1)$ and $U \sim \chi^2_\nu$. Define $T = Z/\sqrt{U/\nu}$. The density of T is:

$$f_T(t) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{t^2}{n}\right)^{-(\nu+1)/2}$$