STA286 Lecture 19

Neil Montgomery

Last edited: 2017-03-09 14:08

other distributions

There are lots of other distributions (continuous and discrete) that have many applications, but we'll stop here.

There will be a few more special-purpose distributions specific for data analysis that I'll introduce when necessary.

statistics

the general problem, and its solution

For every probability problem done so far, the distribution was completely known, along with all paramater values.

the general problem, and its solution

For every probability problem done so far, the distribution was completely known, along with all paramater values.

The problem is: often we encounter one or more distributions that are only partially specified, such as:

- $N(\mu, 2)$, μ unknown.

the general problem, and its solution

For every probability problem done so far, the distribution was completely known, along with all paramater values.

The problem is: often we encounter one or more distributions that are only partially specified, such as:

- $N(\mu, 2)$, μ unknown.
- $N(\mu, \sigma)$, both parameters unknown.

the general problem, and its solution

For every probability problem done so far, the distribution was completely known, along with all paramater values.

The problem is: often we encounter one or more distributions that are only partially specified, such as:

- $N(\mu, 2)$, μ unknown.
- $N(\mu, \sigma)$, both parameters unknown.
- "Some distribution that has unknown mean and variance."

the general problem, and its solution

For every probability problem done so far, the distribution was completely known, along with all paramater values.

The problem is: often we encounter one or more distributions that are only partially specified, such as:

- $N(\mu, 2)$, μ unknown.
- $N(\mu, \sigma)$, both parameters unknown.
- "Some distribution that has unknown mean and variance."

the general problem, and its solution

For every probability problem done so far, the distribution was completely known, along with all paramater values.

The problem is: often we encounter one or more distributions that are only partially specified, such as:

- $N(\mu, 2)$, μ unknown.
- $N(\mu, \sigma)$, both parameters unknown.
- "Some distribution that has unknown mean and variance."

etc.

the general problem, and its solution

For every probability problem done so far, the distribution was completely known, along with all paramater values.

The problem is: often we encounter one or more distributions that are only partially specified, such as:

- $N(\mu, 2)$, μ unknown.
- $N(\mu, \sigma)$, both parameters unknown.
- "Some distribution that has unknown mean and variance."

etc.

The solution to this problem is to obtain a dataset, and use it to *infer* statements about the underlying distribution.

a probabilistic model for the phrase “to obtain a dataset”

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.

a probabilistic model for the phrase "to obtain a dataset"

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.

X
X_1
X_2
X_3
X_4
$\vdots$
X_n

a probabilistic model for the phrase “to obtain a dataset”

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.

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

a probabilistic model for the phrase “to obtain a dataset”

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.

"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

a probabilistic model for the phrase “to obtain a dataset”

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.

"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

a probabilistic model for the phrase “to obtain a dataset”

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.

"Subject ID"	X	Y	"Group ID"	"InputVar"
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

the dataset, once observed, is fixed

The model for a dataset is a model for the plan you make to collect it.

Every method of analysis we will discuss is based on this plan to collect.

the dataset, once observed, is fixed

The model for a dataset is a model for the plan you make to collect it.

Every method of analysis we will discuss is based on this plan to collect.

. . . because once the dataset is collected, there is nothing random about it. It is a fixed rectangle of numbers/etc.

definition of sample | definition of statistic

The basic model for what we'll call a *sample* is a sequence of random variables that are independent with the same distribution (abbreviation: i.i.d.):

$$X_1, X_2, \ldots, X_n$$

definition of sample | definition of statistic

The basic model for what we'll call a *sample* is a sequence of random variables that are independent with the same distribution (abbreviation: i.i.d.):

$$X_1, X_2, \ldots, X_n$$

This is a column of the dataset.

definition of sample | definition of statistic

The basic model for what we'll call a *sample* is a sequence of random variables that are independent with the same distribution (abbreviation: i.i.d.):

$$X_1, X_2, \ldots, X_n$$

This is a column of the dataset.

Some practical generalizations we won't touch: the elements are not independent; the elements do not have the same distribution (e.g. change as a function of time.)

definition of sample | definition of statistic

The basic model for what we'll call a *sample* is a sequence of random variables that are independent with the same distribution (abbreviation: i.i.d.):

$$X_1, X_2, \ldots, X_n$$

This is a column of the dataset.

Some practical generalizations we won't touch: the elements are not independent; the elements do not have the same distribution (e.g. change as a function of time.)

A *statistic* is defined as *a function of the sample.* So, a statistic is a random variable.

example of statistic

The *sample mean* (or sample average):

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

example of statistic

The *sample mean* (or sample average):

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Source of confusion: the function $\overline{x} = \sum_{i=1}^{n} x_i / n$ of a column $\{x_1, \ldots, x_n\}$ in an *observed* dataset is also called "sample mean", but this is not a random quantity.

example of statistic

The *sample mean* (or sample average):

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Source of confusion: the function $\overline{x} = \sum_{i=1}^{n} x_i / n$ of a column $\{x_1, \ldots, x_n\}$ in an *observed* dataset is also called "sample mean", but this is not a random quantity.

I might call $\overline{x}$ the "observed" sample mean, but you really just need to get used to the terminology.

example of statistic

The *sample mean* (or sample average):

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Source of confusion: the function $\overline{x} = \sum_{i=1}^{n} x_i/n$ of a column $\{x_1, \ldots, x_n\}$ in an *observed* dataset is also called "sample mean", but this is not a random quantity.

I might call $\overline{x}$ the "observed" sample mean, but you really just need to get used to the terminology.

You might have a plan to guess the unknown "true" mean of a random variable X using the sample mean $\overline{X}$ of a sample $X_1, \ldots, X_n$ (whose properties can be studied using probability), and when you actually observe the sample $x_1, \ldots, x_n$ your actual guess will be $\overline{x}$.

more "statisics"

The *sample variance*:

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

(source of confusion: the (observed) sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$)

more “statisics”

The *sample variance*:

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

(source of confusion: the (observed) sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$)

The *sample standard deviation*: $S = \sqrt{S^2}$

more "statisics"

The *sample variance*:

$$S^2 = \frac{\sum\limits_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{n-1}$$

(source of confusion: the (observed) sample variance $s^2 = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$)

The *sample standard deviation*: $S = \sqrt{S^2}$

More:

- the sum $\sum\limits_{i=1}^{n} X_i$

more "statisics"

The *sample variance*:

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

(source of confusion: the (observed) sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$)

The *sample standard deviation*: $S = \sqrt{S^2}$

More:

- the sum $\sum_{i=1}^{n} X_i$
- the *minimum* $X_{(1)}$

more “statisics”

The *sample variance*:

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

(source of confusion: the (observed) sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$)

The *sample standard deviation*: $S = \sqrt{S^2}$

More:

- the sum $\sum_{i=1}^{n} X_i$
- the *minimum* $X_{(1)}$
- the *maximum* $X_{(n)}$

more "statisics"

The *sample variance*:

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

(source of confusion: the (observed) sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$)

The *sample standard deviation*: $S = \sqrt{S^2}$

More:

- the sum $\sum_{i=1}^{n} X_i$
- the *minimum* $X_{(1)}$
- the *maximum* $X_{(n)}$
- the *sample median* $\tilde{X}$

more “statisics”

The *sample variance*:

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

(source of confusion: the (observed) sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$)

The *sample standard deviation*: $S = \sqrt{S^2}$

More:

- the sum $\sum_{i=1}^{n} X_i$
- the *minimum* $X_{(1)}$
- the *maximum* $X_{(n)}$
- the *sample median* $\tilde{X}$
- the *sample range* $X_{(n)} - X_{(1)}$

“sampling distributions”

Statistics are random variables, so we mainly care about their distributions, which are (unnecessarily) called the “sampling distributions”.

“sampling distributions”

Statistics are random variables, so we mainly care about their distributions, which are (unnecessarily) called the “sampling distributions”.

Example, if $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p), the statistic $\sum_i X_i$ has a Binomial(n, p) distribution. (What about $\overline{X}$ in this case?)

“sampling distributions”

Statistics are random variables, so we mainly care about their distributions, which are (unnecessarily) called the “sampling distributions”.

Example, if $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p), the statistic $\sum_i X_i$ has a Binomial(n, p) distribution. (What about $\overline{X}$ in this case?)

Example: if $X_1, \ldots, X_n$ are i.i.d. Exponential(λ), the sample mean $\overline{X}$ has a Gamma($n, n\lambda$) distribution.

“sampling distributions”

Statistics are random variables, so we mainly care about their distributions, which are (unnecessarily) called the “sampling distributions”.

Example, if $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p), the statistic $\sum_i X_i$ has a Binomial(n, p) distribution. (What about $\overline{X}$ in this case?)

Example: if $X_1, \ldots, X_n$ are i.i.d. Exponential(λ), the sample mean $\overline{X}$ has a Gamma($n, n\lambda$) distribution.

Important example: if $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma)\ldots$

basic result for $\overline{X}$ in general

For a sample $X_1, \ldots, X_n$ i.i.d. from *any* distribution with mean μ and variance σ^2, the following are always true:

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

basic result for $\overline{X}$ in general

For a sample $X_1, \ldots, X_n$ i.i.d. from *any* distribution with mean μ and variance σ^2, the following are always true:

$$
\begin{aligned}
E\left(\overline{X}\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\
\mathrm{Var}\left(\overline{X}\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)
\end{aligned}
$$

basic result for $\overline{X}$ in general

For a sample $X_1, \ldots, X_n$ i.i.d. from *any* distribution with mean μ and variance σ^2, the following are always true:

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

basic result for $\overline{X}$ in general

For a sample $X_1, \ldots, X_n$ i.i.d. from *any* distribution with mean μ and variance σ^2, the following are always true:

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i)$$

basic result for $\overline{X}$ in general

For a sample $X_1, \ldots, X_n$ i.i.d. from *any* distribution with mean μ and variance σ^2, the following are always true:

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

basic result for $\overline{X}$ in general

For a sample $X_1, \ldots, X_n$ i.i.d. from *any* distribution with mean μ and variance σ^2, the following are always true:

$$
\begin{aligned}
E\left(\overline{X}\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu \\
\mathrm{Var}\left(\overline{X}\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \\
SD(\overline{X}) &= \frac{\sigma}{\sqrt{n}}
\end{aligned}
$$

full distribution of $\overline{X}$ when sample is normal

For a sample $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma)$ we have:

$$M_{\sum_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

so that $\sum_{i=1}^{n} X_i \sim N\left(n\mu, \sqrt{n}\sigma\right)$

full distribution of $\overline{X}$ when sample is normal

For a sample $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma)$ we have:

$$M_{\sum_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(e^{\mu t + \sigma^2 t^2/2}\right)^n$$

so that $\sum_{i=1}^{n} X_i \sim N\left(n\mu, \sqrt{n}\sigma\right)$

full distribution of $\overline{X}$ when sample is normal

For a sample $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma)$ we have:

$$M_{\sum_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(e^{\mu t + \sigma^2 t^2/2}\right)^n = e^{n\mu t + n\sigma^2 t^2/2}$$

so that $\sum_{i=1}^{n} X_i \sim N\left(n\mu, \sqrt{n}\sigma\right)$

full distribution of $\overline{X}$ when sample is normal

For a sample $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma)$ we have:

$$M_{\sum_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(e^{\mu t + \sigma^2 t^2/2}\right)^n = e^{n\mu t + n\sigma^2 t^2/2}$$

so that $\sum_{i=1}^{n} X_i \sim N\left(n\mu, \sqrt{n}\sigma\right)$

full distribution of $\overline{X}$ when sample is normal

For a sample $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma)$ we have:

$$M_{\sum\limits_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(e^{\mu t + \sigma^2 t^2/2}\right)^n = e^{n\mu t + n\sigma^2 t^2/2}$$

so that $\sum\limits_{i=1}^{n} X_i \sim N\left(n\mu, \sqrt{n}\sigma\right)$

Using the rules for normal distributions we also get:

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

full distribution of $\overline{X}$ when sample is normal

For a sample $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma)$ we have:

$$M_{\sum_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(e^{\mu t + \sigma^2 t^2/2}\right)^n = e^{n\mu t + n\sigma^2 t^2/2}$$

so that $\sum_{i=1}^{n} X_i \sim N\left(n\mu, \sqrt{n}\sigma\right)$

Using the rules for normal distributions we also get:

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

full distribution of $\overline{X}$ when sample is normal

For a sample $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma)$ we have:

$$M_{\sum_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(e^{\mu t + \sigma^2 t^2/2}\right)^n = e^{n\mu t + n\sigma^2 t^2/2}$$

so that $\sum_{i=1}^{n} X_i \sim N\left(n\mu, \sqrt{n}\sigma\right)$

Using the rules for normal distributions we also get:

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

full distribution of $\overline{X}$ when sample is normal

For a sample $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma)$ we have:

$$M_{\sum_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(e^{\mu t + \sigma^2 t^2/2}\right)^n = e^{n\mu t + n\sigma^2 t^2/2}$$

so that $\sum_{i=1}^{n} X_i \sim N\left(n\mu, \sqrt{n}\sigma\right)$

Using the rules for normal distributions we also get:

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad\qquad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

The seemingly impossible task is to determine the distribution of $\overline{X}$ when the underlying distribution is not normal (i.e., almost always.)