STA221

Neil Montgomery

Last edited: 2017-07-04 18:43

admin

contact, notes

date format	YYYY-MM-DD – *All Hail ISO8601!!!*
instructor	Neil Montgomery
email	`neilmontg@gmail.com`
office	TBA
office hours	Tuesday and Thursday 18:10 to 19:00
website	portal (announcements, grades, suggested exercises, etc.)
github	`https://github.com/sta221-summer-2017` (lecture material, code, etc.)

Note: I will be in **this room** from 18:10 to 19:00 to answer any questions and solve any problems in an open setting. If you need a private meeting, please make an appointment for some other time.

evaluation, readings, tutorials

what	when	how much
midterm 1	2017-07-18	25%
midterm 2	2017-08-03	25%
exam	TBA	50%

I will provide readings that will contain some suggested exercises, throughout the course.

We are not a big enough course to merit a TA, so there will be no tutorials.

Any thick and comprehensive "Stats 101" book could also be a good resource.

software

Data analysis requires a computer. Also, some concepts can be illustrated using simulation, which also requires a computer. I will be using `R`. It's pretty good at data analysis. You should use it too, but I can't force you.

language	interpreter	integrated development environment
`R`	`R`	`RStudio`

Some detailed instructions and suggestions for installation and configuration will appear on the course website. I will try to impart some data analysis workflow wisdom throughout the course. Some already appears in the detailed instructions.

A really thorough resource for learning `R` is here:

Grolemund, G., Wickham, H., *R for Data Science* **available free at `http://r4ds.had.co.nz/`**

pre-preliminaries—what is a dataset?

most datasets are rectangles

Columns are the *variables*.

The top row has the names of the variables; possibly chosen wisely.

Rows are the *observations* of measurements taken on *units*.

There are no averages, no comments (unless in a “comment” variable), no colors, no formatting, no plots!

not a dataset

A	B	C	D	E
Irrelevant commentary				
	HUGE TITLE ACROSS THREE MERGED LINES			
Some God-forsaken Date Format	Column Title Which Is Very Long And Has Spaces And @$#^ Special Characters!	time2	status	
November 12 2003	2.575817169	27.43610042	censored	
November 12 2003	7.405809497	29.34394097	censored	
November 12 2003	0.372988356	27.33832542	censored	
November 12 2003	3.195281626	12.87646771	pr_fail	
November 12 2003	6.555084512	13.83875584	censored	
November 12 Average	**4.020996232**	**22.16671807**		
November 13 2003	0	11.64588809	censored	
November 13 2003	5.371449791	15.38626237	tx_fail	
November 13 2003	3.928454966	11.40722991	censored	
November 13 2003	4.90945976	20.55325312	censored	
November 13 2003	0	19.44576571	censored	
November 13 Average	**2.841872903**	**15.68767984**		

Neil:
Hey Bob, check out this cell! It's yellow!

not a dataset

ASSETNUM	MOVEDATE_1	FROM_LOCATION1	TO_LOCATION1	MOVEDATE_2	FROM_LOCATION2	TO_LOCATION2	MOVEDATE_3	FRO
0201011	2005-12-16	NO_LOCATION	RSREPAIR					
0209679	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN4VNCR	2014-02-14	DN
0209680	2005-05-17	NO_LOCATION	RSREPAIR	2005-08-03	RSREPAIR	WY172UCR	2013-11-08	WY
0209709	2005-05-20	NO_LOCATION	WY92WEPR	2011-10-07	WY92WEPR	RSREPAIR	2013-11-08	RSR
0209711	2011-10-07	WY91WEPR	RSREPAIR	2013-11-08	RSREPAIR	WY174VNCR		
0209714	2003-12-15	NO_LOCATION	RSREPAIR					
0209720	2011-10-07	WY95WEPR	RSREPAIR	2013-06-25	RSREPAIR	WY70ASPR		
0209722	2011-10-07	WY106WEPR	RSREPAIR	2013-06-27	RSREPAIR	WY144BSUSR		
0209728	2011-10-07	WY94WEPR	RSREPAIR	2013-11-08	RSREPAIR	WY143NWCPR		
0209729	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN12ASRA	2014-04-04	DN
0209737	2005-01-11	NO_LOCATION	DN15NWCRB	2006-03-21	DN15NWCRB	RSREPAIR	2006-03-31	RSR
0209739	2011-10-07	WY144WEPR	RSREPAIR	2013-12-09	RSREPAIR	WY178TPR		
0209740	2011-10-07	WY143WEPR	RSREPAIR	2012-09-12	RSREPAIR	DNSPARE	2014-05-30	DN
0209741	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN10BHR	2014-09-05	DN

an oil readings dataset (wide version)

```
## # A tibble: 612 × 17
##     Ident                Date WorkingAge   TakenBy    Fe    Al    Cu
##     <chr>              <dttm>      <dbl>     <chr> <dbl> <dbl> <dbl>
## 1  448576 1999-05-10 19:00:00        243 EMPL_0917    13     5    14
## 2  448576 1999-07-26 19:00:00        569 EMPL_0917    18     6    25
## 3  448576 1999-09-29 19:00:00        830 EMPL_9375    26     6    35
## 4  448576 1999-10-08 19:00:00        862 EMPL_0917    15     9    14
## 5  448576 1999-11-02 19:00:00        946 EMPL_9375    14     4    19
## 6  448576 1999-12-09 19:00:00       1088 EMPL_0917    18     5    23
## 7  448576 1999-12-27 19:00:00       1157 EMPL_9375    24     8    25
## 8  448576 2000-01-14 19:00:00       1238 EMPL_9375    27     9    34
## 9  448576 2000-02-15 19:00:00       1376 EMPL_9375    16     8    17
## 10 448576 2000-03-11 19:00:00       1492 EMPL_0917    20     8    20
## # ... with 602 more rows, and 10 more variables: Cr <dbl>, Si <dbl>,
## #   Pb <dbl>, Ph <dbl>, Ca <dbl>, Zn <dbl>, Mg <dbl>, Mo <dbl>,
## #   Sn <dbl>, Na <dbl>
```

oil readings with Ident and TakenBy properly treated

```
## # A tibble: 612 × 17
##      Ident                Date WorkingAge   TakenBy    Fe    Al    Cu
##     <fctr>              <dttm>      <dbl>    <fctr> <dbl> <dbl> <dbl>
## 1   448576 1999-05-10 19:00:00        243 EMPL_0917    13     5    14
## 2   448576 1999-07-26 19:00:00        569 EMPL_0917    18     6    25
## 3   448576 1999-09-29 19:00:00        830 EMPL_9375    26     6    35
## 4   448576 1999-10-08 19:00:00        862 EMPL_0917    15     9    14
## 5   448576 1999-11-02 19:00:00        946 EMPL_9375    14     4    19
## 6   448576 1999-12-09 19:00:00       1088 EMPL_0917    18     5    23
## 7   448576 1999-12-27 19:00:00       1157 EMPL_9375    24     8    25
## 8   448576 2000-01-14 19:00:00       1238 EMPL_9375    27     9    34
## 9   448576 2000-02-15 19:00:00       1376 EMPL_9375    16     8    17
## 10  448576 2000-03-11 19:00:00       1492 EMPL_0917    20     8    20
## # ... with 602 more rows, and 10 more variables: Cr <dbl>, Si <dbl>,
## #   Pb <dbl>, Ph <dbl>, Ca <dbl>, Zn <dbl>, Mg <dbl>, Mo <dbl>,
## #   Sn <dbl>, Na <dbl>
```

oil readings dataset (long version)

```
## # A tibble: 7,956 × 6
##      Ident                Date WorkingAge   TakenBy element   ppm
##     <fctr>              <dttm>      <dbl>    <fctr>   <chr> <dbl>
## 1  448576 1999-05-10 19:00:00        243 EMPL_0917      Fe    13
## 2  448576 1999-07-26 19:00:00        569 EMPL_0917      Fe    18
## 3  448576 1999-09-29 19:00:00        830 EMPL_9375      Fe    26
## 4  448576 1999-10-08 19:00:00        862 EMPL_0917      Fe    15
## 5  448576 1999-11-02 19:00:00        946 EMPL_9375      Fe    14
## 6  448576 1999-12-09 19:00:00       1088 EMPL_0917      Fe    18
## 7  448576 1999-12-27 19:00:00       1157 EMPL_9375      Fe    24
## 8  448576 2000-01-14 19:00:00       1238 EMPL_9375      Fe    27
## 9  448576 2000-02-15 19:00:00       1376 EMPL_9375      Fe    16
## 10 448576 2000-03-11 19:00:00       1492 EMPL_0917      Fe    20
## # ... with 7,946 more rows
```

a (simulated) “gas pipeline” dataset

```
## # A tibble: 1,000 × 4
##       Leak  Size Material Pressure
##     <fctr> <ord>   <fctr>   <fctr>
## 1       No  1.75  Aldyl A     High
## 2       No  1.75  Aldyl A      Med
## 3       No     1  Aldyl A      Low
## 4      Yes   1.5    Steel      Med
## 5       No     1    Steel     High
## 6      Yes     1    Steel     High
## 7      Yes  1.75  Aldyl A      Low
## 8       No  1.75    Steel      Med
## 9       No   1.5  Aldyl A     High
## 10      No  1.75    Steel     High
## # ... with 990 more rows
```

important questions

- where did the data come from?

important questions

- where did the data come from?
  - were the units chosen randomly from a population?

important questions

- where did the data come from?
  - were the units chosen randomly from a population?
  - were the units randomly assigned into groups?

important questions

- where did the data come from?
  - were the units chosen randomly from a population?
  - were the units randomly assigned into groups?
- what are the (joint) *distributions* of the data?

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

(Not a focus of this course.)

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

(Not a focus of this course.)

Sometimes data come from an *experiment* where units are randomly assigned to different *levels* of one or more *factors*, in which cause cause-and-effect can be inferred using laws of probability.

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

(Not a focus of this course.)

Sometimes data come from an *experiment* where units are randomly assigned to different *levels* of one or more *factors*, in which cause cause-and-effect can be inferred using laws of probability.

Often the data are just some records of what happened. Grander inferences might be made, but only on a subject-matter basis.

distribution (informally)

- A *distribution* is a

distribution (informally)

- A *distribution* is a
  - Complete description of...

distribution (informally)

- A *distribution* is a
  - Complete description of…
  - …the possible values of one or more variables…

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed

distribution (informally)

- A *distribution* is a
  - Complete description of. . .
  - . . . the possible values of one or more variables. . .
  - . . . and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed
  - numerically

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed
  - numerically
  - graphically

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed
  - numerically
  - graphically
- We can also consider probability models for one or more variables or a relationship among variables. (Focus of this course.)

important concepts from probability

independence

independence - definition and example

Two *events* A and B are *independent* if:

$$P(A \cap B) = P(A)P(B),$$

(where $\cap$ means *and*.)

For example, roll a fair die. Let $A = \{1, 2, 3\}$ and $B = \{2, 4\}$.

independence - definition and example

Two *events* A and B are *independent* if:

$$P(A \cap B) = P(A)P(B),$$

(where $\cap$ means *and*.)

For example, roll a fair die. Let $A = \{1, 2, 3\}$ and $B = \{2, 4\}$.

$P(A) = 1/2$ and $P(B) = 1/3$, so $P(A)P(B) = 1/6$.

independence - definition and example

Two *events* A and B are *independent* if:

$$P(A \cap B) = P(A)P(B),$$

(where $\cap$ means *and*.)

For example, roll a fair die. Let $A = \{1, 2, 3\}$ and $B = \{2, 4\}$.

$P(A) = 1/2$ and $P(B) = 1/3$, so $P(A)P(B) = 1/6$.

Also, $A \cap B = \{2\}$ so $P(A \cap B) = 1/6 = P(A)P(B)$

independence - definition and example

Two *events* A and B are *independent* if:

$$P(A \cap B) = P(A)P(B),$$

(where $\cap$ means *and*.)

For example, roll a fair die. Let $A = \{1, 2, 3\}$ and $B = \{2, 4\}$.

$P(A) = 1/2$ and $P(B) = 1/3$, so $P(A)P(B) = 1/6$.

Also, $A \cap B = \{2\}$ so $P(A \cap B) = 1/6 = P(A)P(B)$

Conclude: A and B are independent (short form: $A \perp B$.)

independence - definition and example

Two *events* A and B are *independent* if:

$$P(A \cap B) = P(A)P(B),$$

(where $\cap$ means *and*.)

For example, roll a fair die. Let $A = \{1, 2, 3\}$ and $B = \{2, 4\}$.

$P(A) = 1/2$ and $P(B) = 1/3$, so $P(A)P(B) = 1/6$.

Also, $A \cap B = \{2\}$ so $P(A \cap B) = 1/6 = P(A)P(B)$

Conclude: A and B are independent (short form: $A \perp B$.)

Exercise: if $C = \{2, 5, 6\}$ then $B \perp C$ and $A \not\perp C$

independence - comments

Independence is normally something that is *assumed* and not something that is demonstrated.

independence - comments

Independence is normally something that is *assumed* and not something that is demonstrated.

Undisciplined use of language (e.g. "A has nothing to do with B") is the leading cause of error. Use the definition.

independence - comments

Independence is normally something that is *assumed* and not something that is demonstrated.

Undisciplined use of language (e.g. "A has nothing to do with B") is the leading cause of error. Use the definition.

The opposite of independent is "not independent." (Avoid "dependent", which has misleading connotations.)

independence - comments

Independence is normally something that is *assumed* and not something that is demonstrated.

Undisciplined use of language (e.g. "A has nothing to do with B") is the leading cause of error. Use the definition.

The opposite of independent is "not independent." (Avoid "dependent", which has misleading connotations.)

$A \perp B \iff A \perp B^c \iff A^c \perp B \iff A^c \perp B^c$

random variables and distributions

concept of random variable

A *random variable* is a rule that assigns a number to any outcome of a random process.

Example: "Roulette". There are 38 slots on a wheel coloured as follows:

Colour	# of slots	Slot labels
Green	2	0, 00
Red	18	1, 3, 5, 7, 9, 12, 14, 16, 18, 19, 21, 23, 25, 27, 30, 32, 34, 36
Black	18	2, 4, 6, 8, 10, 11, 13, 15, 17, 20, 22, 24, 26, 28, 29, 31, 33, 35

roulette - II

If bet \$100 on “Red”, then these are the possibilities:

Result	I receive
Red	200
Not Red	0

Stated another way, here is my net “gain”, which I will call X, after the play:

Result	X
Red	100
Not Red	-100

roulette - III

Technically the random variable is this the *rule*:

$$X(1) = X(3) = X(5) = \cdots = X(36) = 100$$

$$X(00) = X(0) = X(2) = \cdots = X(35) = -100$$

roulette - III

Technically the random variable is this the *rule*:

$$X(1) = X(3) = X(5) = \cdots = X(36) = 100$$

$$X(00) = X(0) = X(2) = \cdots = X(35) = -100$$

But this is often a useless technicality. This is all we care about:

x	$P(X = x)$
100	$18/38$
-100	$20/38$

This table is the *distribution* of X, i.e. the possible outcomes and their probabilities.

distribution and independence

The distribution of a random variable X is, roughly, all information about the values of X and their probabilities.

distribution and independence

The distribution of a random variable X is, roughly, all information about the values of X and their probabilities.

There's the odd (or maybe not?) fact that when X is *continuously measured* then we have $P(X = x) = 0$ for any particular x. In this case we're concerned with intervals of values and not particular values.

distribution and independence

The distribution of a random variable X is, roughly, all information about the values of X and their probabilities.

There's the odd (or maybe not?) fact that when X is *continuously measured* then we have $P(X = x) = 0$ for any particular x. In this case we're concerned with intervals of values and not particular values.

X and Y can be independent when *knowing the outcome of X does not change the distribution of Y - a very strong statement (usually assumed when appropriate.)

expected value

Random variables can have expected values (averages, means), variances, and standard deviations, that follow these rules:

$$E(aX + b) = aE(X) + b$$

expected value

Random variables can have expected values (averages, means), variances, and standard deviations, that follow these rules:

$$E(aX + b) = aE(X) + b$$
$$E(X + Y) = E(X) + E(Y)$$

expected value

Random variables can have expected values (averages, means), variances, and standard deviations, that follow these rules:

$$E(aX + b) = aE(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

expected value

Random variables can have expected values (averages, means), variances, and standard deviations, that follow these rules:

$$
\begin{aligned}
E(aX + b) &= aE(X) + b \\
E(X + Y) &= E(X) + E(Y) \\
\mathrm{Var}(aX + b) &= a^2 \mathrm{Var}(X) \\
\mathrm{Var}(X + Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) \text{ when } X \perp Y
\end{aligned}
$$

normal distributions and the central limit theorem

Normal distributions are an important family of symmetric, bell-shaped distributions, parametrized by mean μ and standard deviation σ.

normal distributions and the central limit theorem

Normal distributions are an important family of symmetric, bell-shaped distributions, parametrized by mean μ and standard deviation σ.

They are so widely used *in statistics* because the distribution of a sample average will be approximately normal if the sample size is "large enough".

normal distributions and the central limit theorem

Normal distributions are an important family of symmetric, bell-shaped distributions, parametrized by mean μ and standard deviation σ.

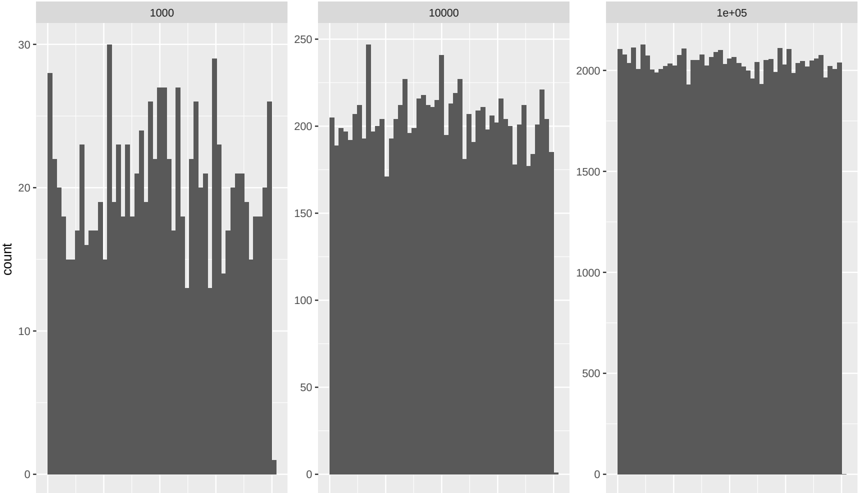
They are so widely used *in statistics* because the distribution of a sample average will be approximately normal if the sample size is "large enough".

"Large enough" is not fixed, but depends on the shape of the underlying population distribution, with more skewness requiring a larger sample size.

normal approximation illustration through simulation - I

I can simulate picking numbers uniformly at random between 0 and 1.

Here are histograms of 1000, 10000, and 100000 picks:



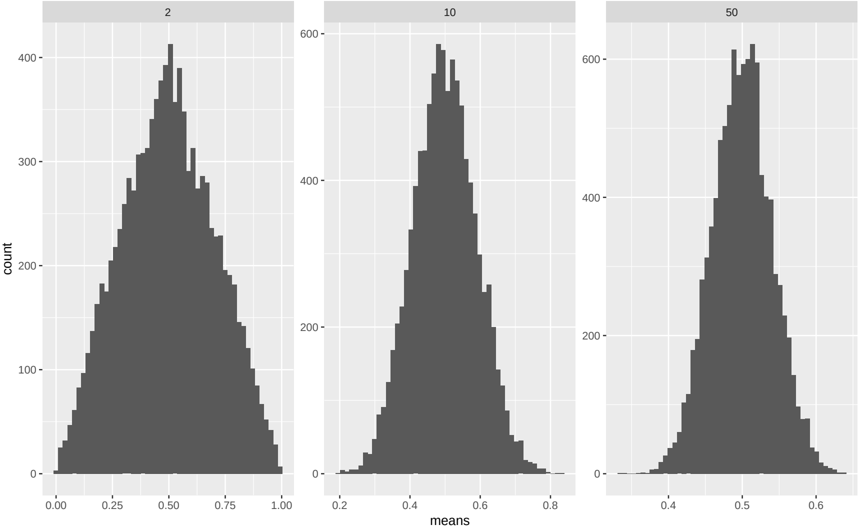
normal approximation illustration through simulation - II

I'll settle on $k = 10000$ "replications" of my simulation.

My simulation will actually consist of: * picking n numbers uniformly at random * calculating the average of those n numbers * doing this k times * making a histogram of the results.

I will choose n to be 2, 10, and 50.

normal approximation illustration through simulation - III



t distributions

If a population is being modeled with a $N(\mu, \sigma)$ probability model and you are going to gather a sample $X_1, X_2, \ldots, X_n$, then the following are true:

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$$

t distributions

If a population is being modeled with a $N(\mu, \sigma)$ probability model and you are going to gather a sample $X_1, X_2, \ldots, X_n$, then the following are true:

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$$

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

t distributions

If a population is being modeled with a $N(\mu, \sigma)$ probability model and you are going to gather a sample $X_1, X_2, \ldots, X_n$, then the following are true:

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

We usually don't know σ, but we can estimate it from the data using s, but then:

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

$n - 1$ is called "degrees of freedom".

degress of freedom

"Degrees of freedom" comes from the denominator $s/\sqrt{n}$. Let's look at (the square of) s:

degress of freedom

"Degrees of freedom" comes from the denominator $s/\sqrt{n}$. Let's look at (the square of) s:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_i)^2}{n-1}$$

There's $n - 1$ again!

degress of freedom

"Degrees of freedom" comes from the denominator $s/\sqrt{n}$. Let's look at (the square of) s:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_i)^2}{n-1}$$

There's $n-1$ again!

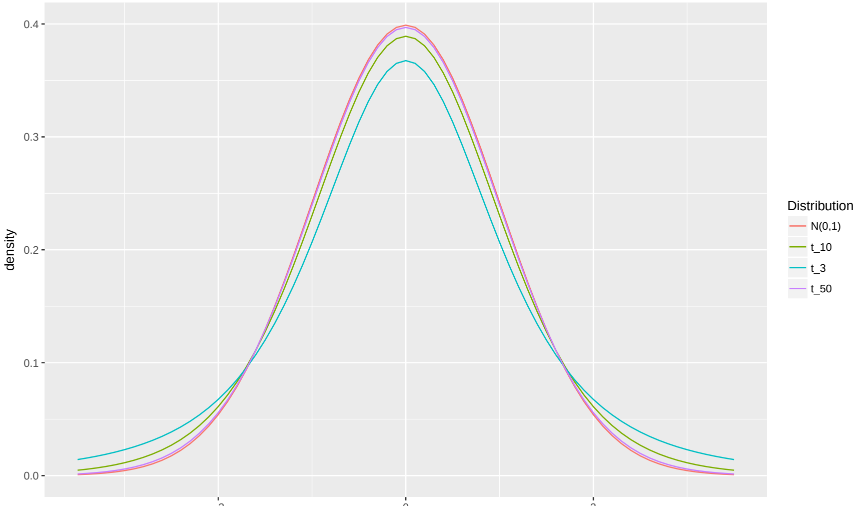
The phrase "degrees of freedom" comes from the realization that *given the value of* $\overline{x}$ the following list of number is redundant:

$$\{x_1, x_2, x_3, \ldots, x_n\}$$

From *any* $n-1$ of them, along with $\overline{x}$, you could calculate the missing value.

t distributions - II

The t distributions are (another) family of symmetric and bell-shaped distributions that look very much like $N(0,1)$ distributions.



estimation - confidence intervals

From the following:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \text{and} \qquad \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

which are approximately true for "large enough" n we get the usual 95% confidence intervals:

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}} \qquad \text{and} \qquad \overline{X} \pm \text{"2"}\frac{s}{\sqrt{n}}$$

I put "2" because the value (for a 95% interval) is always close to 2.

hypothesis testing - some very opinionated hints

Sometimes particular values of a population parameters have an obvious meaning along the lines of “no difference”, “no relationship”, or something similar.

hypothesis testing - some very opinionated hints

Sometimes particular values of a population parameters have an obvious meaning along the lines of “no difference”, “no relationship”, or something similar.

This obvious parameter value can be given the grand title “null hypothesis”, such as in:

$$H_0 : \mu_1 = \mu_2$$

hypothesis testing - some very opinionated hints

Sometimes particular values of a population parameters have an obvious meaning along the lines of “no difference”, “no relationship”, or something similar.

This obvious parameter value can be given the grand title “null hypothesis”, such as in:

$$H_0 : \mu_1 = \mu_2$$

The “alternative” is the negation of the null. (No selecting alternatives based on hopes and dreams!), such as in:

$$H_a : \mu_1 \neq \mu_2$$

hypothesis testing - some very opinionated hints

Sometimes particular values of a population parameters have an obvious meaning along the lines of "no difference", "no relationship", or something similar.

This obvious parameter value can be given the grand title "null hypothesis", such as in:

$$H_0 : \mu_1 = \mu_2$$

The "alternative" is the negation of the null. (No selecting alternatives based on hopes and dreams!), such as in:

$$H_a : \mu_1 \neq \mu_2$$

Modern inference is done using "p-values", which are defined as *the probability of observing a summary of the data that is more extreme than what was observed.*

p-values

More extreme than what?

p-values

More extreme than what?

More extreme than where the null hypothesis "lives"

p-values

More extreme than what?

More extreme than where the null hypothesis "lives"

Hypothesis testing and p-values are controversial, due to misuse, misunderstanding, and lots of other issues.

p-values

More extreme than what?

More extreme than where the null hypothesis "lives"

Hypothesis testing and p-values are controversial, due to misuse, misunderstanding, and lots of other issues.

Required reading: the ASA Statement on Statistical Significance and P-Values (pdf with lecture materials.)

example ("eye drops")

Which eye drop (A or B) for pupil dilation wears off faster?

40 people are each given both eye drops on different days. The wear-out times are recorded for each person.

```
## # A tibble: 40 × 3
##          A        B Difference
##      <dbl>    <dbl>      <dbl>
## 1 107.4709 115.8900  -8.419056
## 2 123.6729 128.5384  -4.865533
## 3 103.2874 146.1660 -42.878535
## 4 151.9056 189.1721 -37.266528
## 5 126.5902 114.0399  12.550228
## # ... with 35 more rows
```

example "eye drops"

Mean and standard deviation of `Difference` are:

x-bar	sd
-19.81	37.03

example “eye drops”

Mean and standard deviation of `Difference` are:

x-bar	sd
-19.81	37.03

The “standard error” of $\overline{x}$ is $s/\sqrt{n} = 5.8554612$

the *t* test in R

```
##
##  One Sample t-test
##
## data:  eyedrops$Difference
## t = -3.3833, df = 39, p-value = 0.001642
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -31.654462  -7.966885
## sample estimates:
## mean of x
## -19.81067
```

goodness-of-fit testing

detour 1 - what tends to have a Normal distribution?

In Stats 101 you will have encountered (at least) Binomial and Normal distributions.

detour 1 - what tends to have a Normal distribution?

In Stats 101 you will have encountered (at least) Binomial and Normal distributions.

And we've learned that $\overline{X}$ is always approximately Normal when the sample is large enough.

Sums of random things also tend to be approximately Normal.

detour 1 - what tends to have a Normal distribution?

In Stats 101 you will have encountered (at least) Binomial and Normal distributions.

And we've learned that $\overline{X}$ is always approximately Normal when the sample is large enough.

Sums of random things also tend to be approximately Normal.

A special case is the so-called "Normal approximation to the Binomial".

detour 1 - what tends to have a Normal distribution?

In Stats 101 you will have encountered (at least) Binomial and Normal distributions.

And we've learned that $\overline{X}$ is always approximately Normal when the sample is large enough.

Sums of random things also tend to be approximately Normal.

A special case is the so-called "Normal approximation to the Binomial".

Why? Recall that a Binomial probability model is used to *count* the number of "*successes*" in n "trials".

detour 1 - what tends to have a Normal distribution?

In Stats 101 you will have encountered (at least) Binomial and Normal distributions.

And we've learned that $\overline{X}$ is always approximately Normal when the sample is large enough.

Sums of random things also tend to be approximately Normal.

A special case is the so-called "Normal approximation to the Binomial".

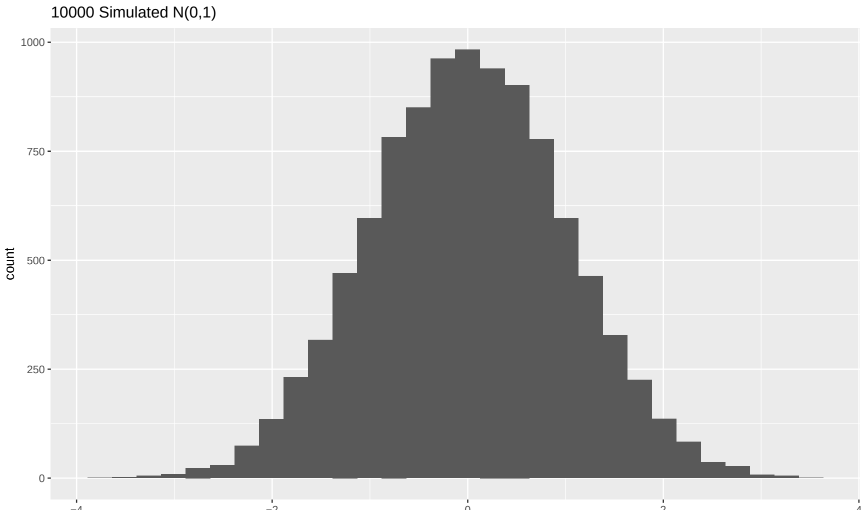
Why? Recall that a Binomial probability model is used to *count* the number of "*successes*" in n "trials".

Let's map "success" to the number 1 and "failure" to the number 0.

Counting 1s in a sequence of 0s and 1s is exactly equivalent to adding up all the 0s and 1s

detour 2.1 - what happens when you look at the square of a normal?

My computer can simulate random "draws" from a standard normal (N(0,1)) distribution, resulting in a histogram such as:

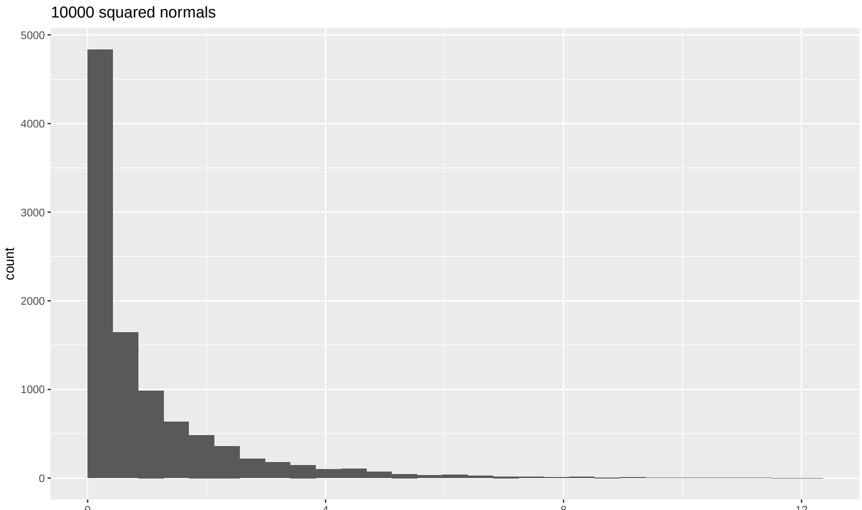


detour 2.2 - what what happens when you look at the square of a normal?

I could take all of those simulated standard normals and square them, and make a histogram of the result, which would give:

detour 2.2 - what what happens when you look at the square of a normal?

I could take all of those simulated standard normals and square them, and make a histogram of the result, which would give:

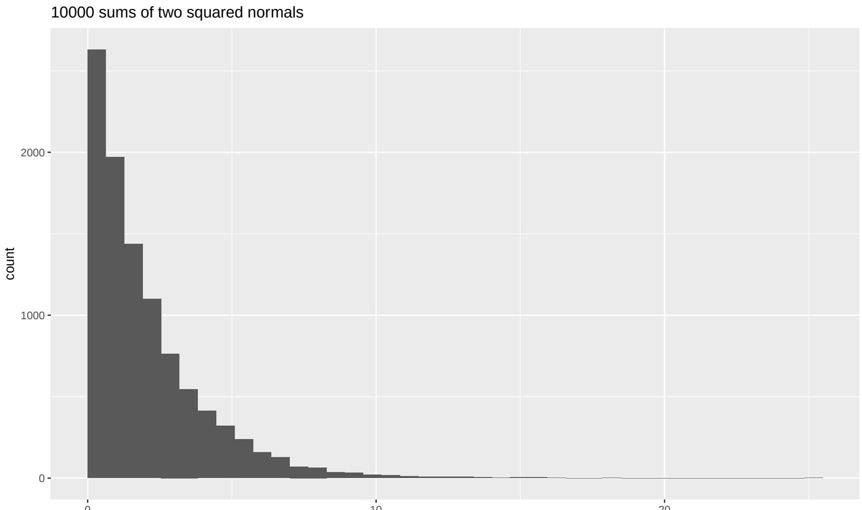


detour 2.3 - sum of squared normals?

I can simulate *two* columns of standard normals, square them *both*, add the results, and make a histogram of the result:

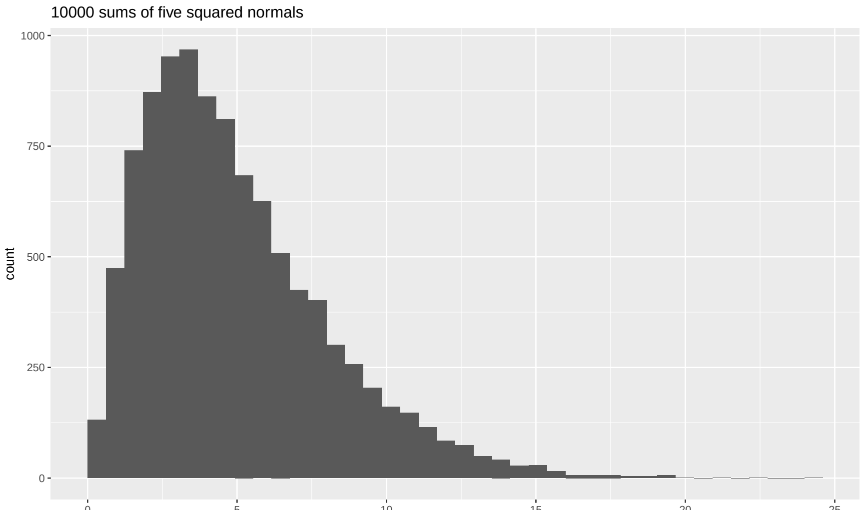
detour 2.3 - sum of squared normals?

I can simulate *two* columns of standard normals, square them *both*, add the results, and make a histogram of the result:



detour 2.4 - sum of many squared normals?

I can make several columns of normals, square them, add them up, and make a histogram. Here's the histogram with 5 columns of normals:



detour - the χ^2 family of distributions

If you have n independent standard normals, the sum of their squares will have a χ_n^2 distribution.

detour - the χ^2 family of distributions

If you have n independent standard normals, the sum of their squares will have a χ_n^2 distribution.

The n is a parameter going by the name "degrees of freedom."

detour - the χ^2 family of distributions

If you have n independent standard normals, the sum of their squares will have a χ^2_n distribution.

The n is a parameter going by the name "degrees of freedom."

If you have n general $N(\mu, \sigma)$, say called $X_1, X_2, \ldots, X_n$, you could *standardize them*:

$$Z_i = \frac{X_i - \mu}{\sigma}$$

and then the sums of the squares of these Z_i will have a χ^2_n distribution.

detour - the χ^2 family of distributions

If you have n independent standard normals, the sum of their squares will have a χ^2_n distribution.

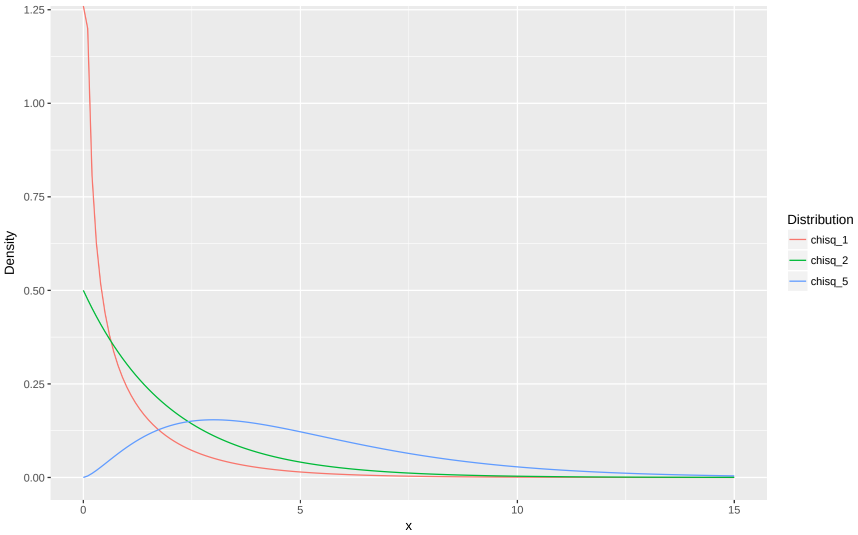
The n is a parameter going by the name "degrees of freedom."

If you have n general $N(\mu, \sigma)$, say called $X_1, X_2, \ldots, X_n$, you could *standardize them*:

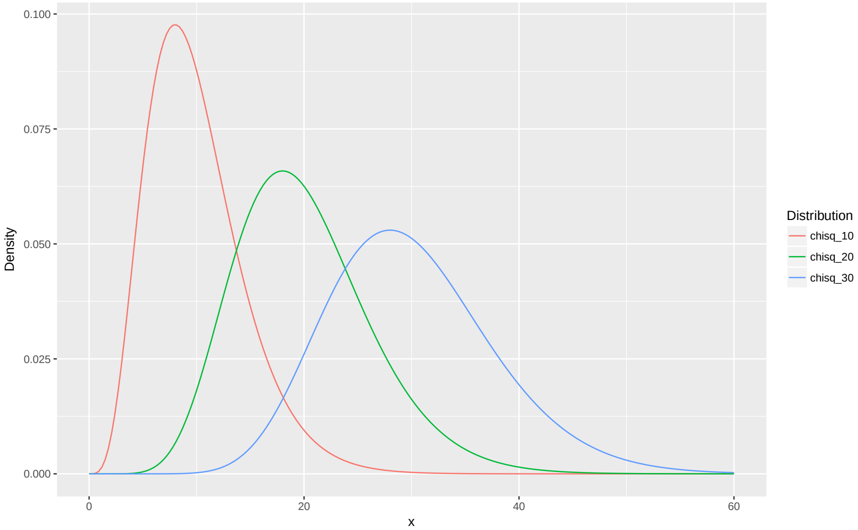
$$Z_i = \frac{X_i - \mu}{\sigma}$$

and then the sums of the squares of these Z_i will have a χ^2_n distribution.

detour - pictures of some χ^2_n distributions



detour - pictures of more χ^2_n distributions



Note: the average of a χ^2_n distribution is just n.

ever wonder why the sample variance is divided by $n-1$?

Look at the formula for sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

The numerator is a sum of n squares, but the denominator is $n-1$. Why?

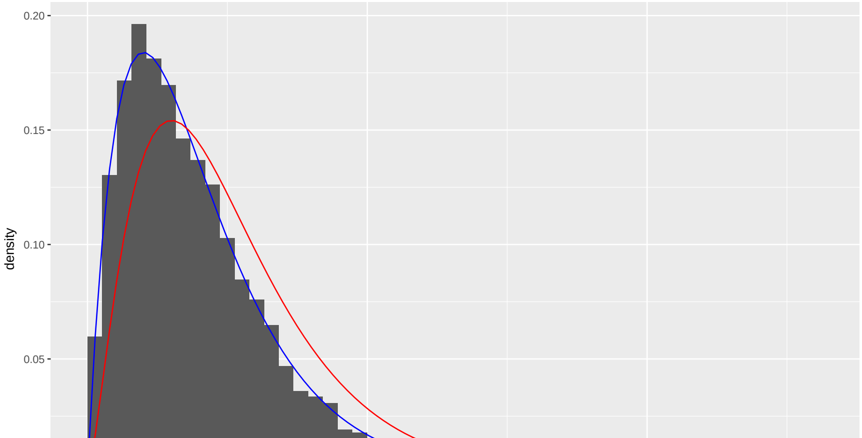
pictures of $\sum_{i=1}^{5}(x_i - \overline{x})^2$

I can simulate samples of size, say, 5 and compute that numerator, and make a histogram.

pictures of $\sum_{i=1}^{5}(x_i - \overline{x})^2$

I can simulate samples of size, say, 5 and compute that numerator, and make a histogram.

Here it is, with the χ^2_4 distribution in blue and the χ^2_5 in red:



a heuristic explanation

s^2 is calculated after fixing the value of $\overline{x}$

a heuristic explanation

s^2 is calculated after fixing the value of $\overline{x}$

So given $\overline{x}$ and *any* $n - 1$ of the n raw values, I can calculate that other raw value.

a heuristic explanation

s^2 is calculated after fixing the value of $\overline{x}$

So given $\overline{x}$ and *any* $n - 1$ of the n raw values, I can calculate that other raw value.

We say s^2 (given $\overline{x}$) only has $n - 1$ degrees of freedom.

is there evidence that something doesn't follow a given distribution?

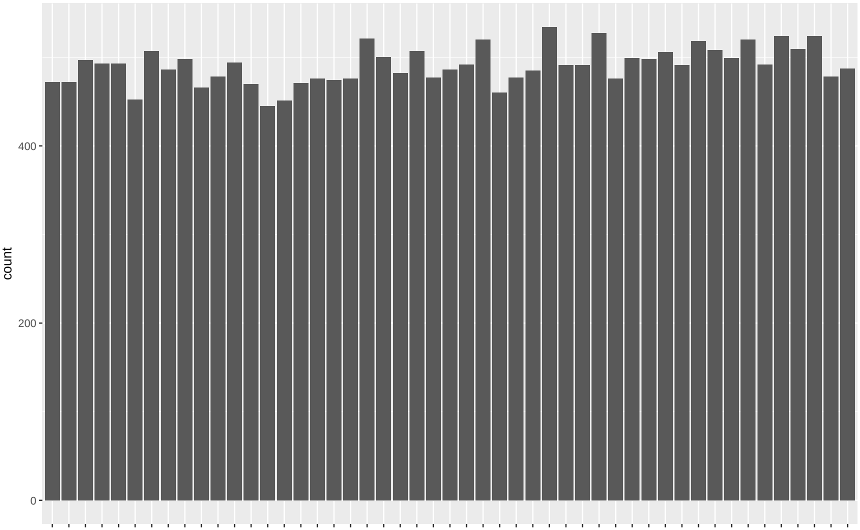
is a lottery “fair”

Lotto 6/49 is a Canadian lottery in which 49 identical balls are mixed together and 7 are selected, now twice per week. People can win money based on how many of the numbers they have out of the 6 on their ticket.

I found a list of every number ever picked here:
`http://portalseven.com/lottery/canada_lotto_649.jsp`

```
## # A tibble: 3,437 × 8
##                 date  num1  num2  num3  num4  num5  num6 bonus
##                <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Sat, Jan 14, 2017     1     6    19    30    32    44    33
## 2 Wed, Jan 11, 2017    24    34    36    38    42    43    30
## 3  Sat, Jan 7, 2017     1    10    18    19    23    27    48
## 4  Wed, Jan 4, 2017     2    11    13    23    35    48    30
## 5 Sat, Dec 31, 2016     3     5    14    18    26    28    40
## # ... with 3,432 more rows
```

all 49 numbers should appear with roughly the same frequency



categorical data, cells, observed cell counts

The dataset (now) consists of one variable called `numbers`. This is a *categorical*, or *factor* variable with 49 possible *levels*. There are 24050 observations.

categorical data, cells, observed cell counts

The dataset (now) consists of one variable called `numbers`. This is a *categorical*, or *factor* variable with 49 possible *levels*. There are 24050 observations.

A categorical variable is summarized by producing a table of *observed cell counts* (notation: O_i). In this case:

```
## # A tibble: 49 × 2
##   numbers   O_i
##    <fctr> <int>
## 1       1   472
## 2       2   472
## 3       3   497
## 4       4   493
## 5       5   493
## # ... with 44 more rows
```

expected cell counts

If Lotto $6/49$ is actually fair, each number would appear with probability $1/49 = 0.0204$ each.

After 24050 numbers have been selected, we would expect to see:

$$24050 \cdot \frac{1}{49} = 490.82$$

of each number.

These are called *expected cell counts* — calculated under the assumption of fairness as defined in this example. (Notation: E_i)

measuring the deviation from the assumption of fairness

Each O_i is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

measuring the deviation from the assumption of fairness

Each O_i is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

How many? As long as $E_i \geq 10$, the approximation will be good.

measuring the deviation from the assumption of fairness

Each O_i is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

How many? As long as $E_i \geq 10$, the approximation will be good.

(Note: this is exactly like the $np \geq 10$ or $np \geq 5$ suggestion that is given for the accuracy of a normal approximation to a binomial.)

measuring the deviation from the assumption of fairness

Each O_i is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

How many? As long as $E_i \geq 10$, the approximation will be good.

(Note: this is exactly like the $np \geq 10$ or $np \geq 5$ suggestion that is given for the accuracy of a normal approximation to a binomial.)

The overall deviation is measured as:

$$\sum_{i=1}^{n} \left(\frac{O_i - E_i}{\sqrt{E_i}} \right)^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

measuring the deviation - compared to what?

The expected cell counts E_i are computed *for a given fixed sample size* N.

measuring the deviation - compared to what?

The expected cell counts E_i are computed *for a given fixed sample size* N.

So given n along with any of the $n-1$ expected cell counts, we could compute that other expected cell count.

measuring the deviation - compared to what?

The expected cell counts E_i are computed *for a given fixed sample size* N.

So given n along with any of the $n-1$ expected cell counts, we could compute that other expected cell count.

(This might *seem* trivial in this example because all the expected cell counts are the same - but this is only because our hypothesis is that all the cell probabilities are the same.)

measuring the deviation - compared to what?

The expected cell counts E_i are computed *for a given fixed sample size* N.

So given n along with any of the $n-1$ expected cell counts, we could compute that other expected cell count.

(This might *seem* trivial in this example because all the expected cell counts are the same - but this is only because our hypothesis is that all the cell probabilities are the same.)

We say

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

has $n-1$ degrees of freedom, and it follows (approximately) a χ^2_{n-1} distribution.

let's measure the deviation

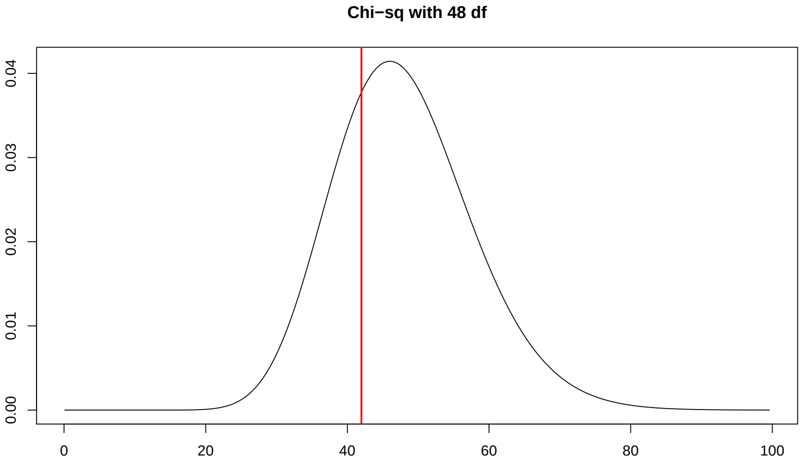
Here are the first few deviations (with $(O_i - E_i)^2/E_i$ called `D_i` for short):

```
## # A tibble: 49 × 4
##   numbers   O_i    E_i         D_i
##    <fctr> <int>  <dbl>       <dbl>
## 1       1   472 490.82 0.721634000
## 2       2   472 490.82 0.721634000
## 3       3   497 490.82 0.077813455
## 4       4   493 490.82 0.009682572
## 5       5   493 490.82 0.009682572
## # ... with 44 more rows
```

The sum of the `D_i` column is 41.99. Is this number surprising?

surprising, compared to what?

We know we should compare this number with the χ^2_{48} distribution. Here we can see we are not surprised. There is no evidence that Lotto 6/49 is unfair.



goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribution of interest.

goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribution of interest.

In the Lotto example, technically this statement is:

$$H_0 : p_1 = p_2 = \cdots = p_{49} = \frac{1}{49}$$

But usually we just make H_0 a simple written statement:

$$H_0 : \text{ the probabilities are all the same.}$$

goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribution of interest.

In the Lotto example, technically this statement is:

$$H_0 : p_1 = p_2 = \cdots = p_{49} = \frac{1}{49}$$

But usually we just make H_0 a simple written statement:

$$H_0 : \text{ the probabilities are all the same.}$$

The "distribution of interest" is technically the "discrete uniform distribution on the outcomes $\{1, 2, 3, \ldots, 49\}$"

goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribution of interest.

In the Lotto example, technically this statement is:

$$H_0 : p_1 = p_2 = \cdots = p_{49} = \frac{1}{49}$$

But usually we just make H_0 a simple written statement:

$$H_0 : \text{ the probabilities are all the same.}$$

The "distribution of interest" is technically the "discrete uniform distribution on the outcomes $\{1, 2, 3, \ldots, 49\}$"

The alternative hypothesis is the negation of the null. We don't normally bother to write it down.

goodness of fit as formal hypothesis test - II

Given a sample size N and the null hypothesis probabilities, compute the n expected cell counts. In this case:

$$E_i = Np_i = 24050 \cdot \frac{1}{49} = 490.82$$

goodness of fit as formal hypothesis test - II

Given a sample size N and the null hypothesis probabilities, compute the n expected cell counts. In this case:

$$E_i = Np_i = 24050 \cdot \frac{1}{49} = 490.82$$

Compute the *observed value of the test statistic*:

$$\chi^2_{\text{obs}} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 41.99$$

goodness of fit as formal hypothesis test - II

Given a sample size N and the null hypothesis probabilities, compute the n expected cell counts. In this case:

$$E_i = Np_i = 24050 \cdot \frac{1}{49} = 490.82$$

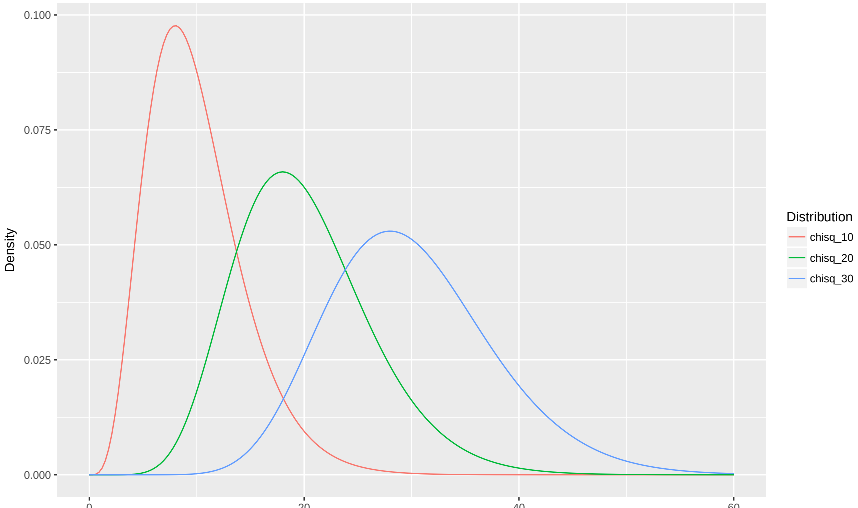
Compute the *observed value of the test statistic*:

$$\chi^2_{\text{obs}} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 41.99$$

Calculate the p-value based on χ^2_{obs} being approximately χ^2_{n-1}.

goodness-if-fit testing p-value

A p-value is the *probability of observing a more extreme value*, in the sense of being further from where the null hypothesis "lives", which is where in this case?



goodness-of-fit testing p-value

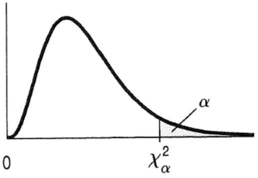
The p-value is $P(\chi^2_{48} \geq 41.99) = 0.7165747$

goodness-of-fit testing p-value

The p-value is $P(\chi^2_{48} \geq 41.99) = 0.7165747$

On tests you'll need to use a table. Here's a close-up of a table I found in a book:

Table X
Values of χ^2_α



Right-tail probability df	0.10	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819

goodness-of-fit testing p-value (from table)

28	37.916	41.337	44.461	48.278	50.994
29	39.087	42.557	45.722	59.588	52.336
30	40.256	43.773	46.979	50.892	53.672
40	51.805	55.759	59.342	63.691	66.767
50	63.167	67.505	71.420	76.154	79.490
60	74.397	79.082	83.298	88.381	91.955
70	85.527	90.531	95.023	100.424	104.213

From a table the best you can do is to estimate the p-value.

All this together is called the "χ^2 goodness-of-fit test."

applications of χ^2 goodness-of-fit testing to two-way tables

contingency tables

Recall the gas pipelines data:

```
## # A tibble: 1,000 × 4
##      Leak  Size Material Pressure
##    <fctr> <ord>   <fctr>   <fctr>
## 1      No  1.75  Aldyl A     High
## 2      No  1.75  Aldyl A      Med
## 3      No     1  Aldyl A      Low
## 4     Yes   1.5    Steel      Med
## 5      No     1    Steel     High
## # ... with 995 more rows
```

The (only?) suitable numerical summary for two categorical/factor variables at a time is a so-called contingency table, or two-way table.

two-way table for “Leak” and “Pressure”

	High	Low	Med	Sum
No	277	278	247	802
Yes	71	66	61	198
Sum	348	344	308	1000

the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

Question 2: are the rows and columns *independent*?

the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

Question 2: are the rows and columns *independent*?

This question is answered using a *test of independence*.

the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

Question 2: are the rows and columns *independent*?

This question is answered using a *test of independence*.

The mechanics of both tests are identical. Only the interpretation is (slightly) different.

two-way table again

Count version:

	High	Low	Med	Sum
No	277	278	247	802
Yes	71	66	61	198
Sum	348	344	308	1000

Proportion version:

	High	Low	Med	Sum
No	0.277	0.278	0.247	0.802
Yes	0.071	0.066	0.061	0.198
Sum	0.348	0.344	0.308	1.000