Using R built-in distribution functions to solve sample mean problems

Examples from Stanford OLI ProbStat reading

Challenge – how to do it?

- You have data from a random sample and want to estimate a population mean within a tolerance.
- You know a population mean and want to find out how unlikely a given sample mean is.
- What do you do?

Use central limit theorem

- View **sample mean** and **sample proportion** as random variables
  - Your sample's mean (or proportion) is one realization of that random variable
- Both are **normally distributed** with
  - mean equal to population mean or proportion
  - standard deviation **equal to population standard deviation divided by square root of sample size**

Using CLT - sample mean

- If μ is population mean, σ is population standard deviation, N is sample size
  - then sample mean is normally distributed with mean μ and standard deviation σ/sqrt(N)
- This lets you estimate population mean using sample mean

Using CLT - sample proportion

- If p is population proportion and N is sample size
  - Population standard deviation is sqrt(p*(1-p))
  - Sample proportion is normally distributed with mean p and standard deviation sqrt(p*(1-p)/N)
- This lets you estimate population proportion using sample proportion

dnorm

qnorm

rnorm

pnorm

http://seankross.com/notes/dpqr/

Introduction to dnorm, pnorm, qnorm, and rnorm for new biostatisticians

Sean Kross

October 1, 2015

Today I was in Dan's office hours and someone asked, "what is the equivalent in R of the back of the stats textbook table of probabilities and their corresponding Z-scores?" (This is an example of the kind of table the student was talking about.) This question indicated to me that although we've been asked to use some of the distribution functions in past homeworks, there may be some misunderstanding about how these functions work.

Right now I'm going to focus on the functions for the normal distribution, but you can find a list of all distribution functions by typing `help(Distributions)` into your R console.

`dnorm`

As we all know the probability density for the normal distribution is:

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Read it now...

Distribution functions in R - `dnorm, pnorm, qnorm, rnorm`

- dnorm - density function; for a given x, find the corresponding y value on the normal curve

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- use it to plot normal curves

Can use `dnorm` to plot normal curves

- dnorm - density function; for a given x, find the corresponding y value on the normal curve

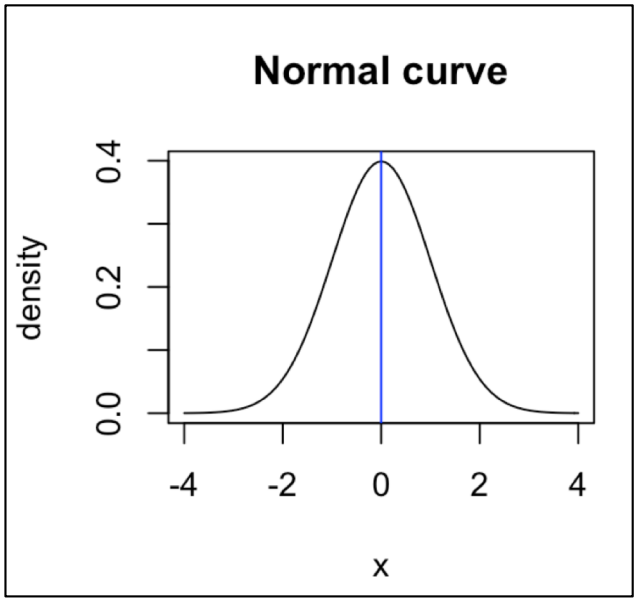
$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

```
1  x=seq(-4,4,by=0.01)
2  y=dnorm(x)
3  plot(x,y,
4       ylab="density",
5       pch=".",
6       main="Normal curve")
7
8  plot(function(x) dnorm(x),
9       ylab="density",
10      from=-4,
11      to=4,
12      main="Normal curve")
```

`pnorm` - cumulative distribution function finds P(X< x), area under curve to the **left of x**

```
 8  plot(function(x) dnorm(x),
 9       ylab="density",
10       from=-4,
11       to=4,
12       main="Normal curve")
13  x=0
14  abline(v=x,col="blue")
15  pnorm(x)
```

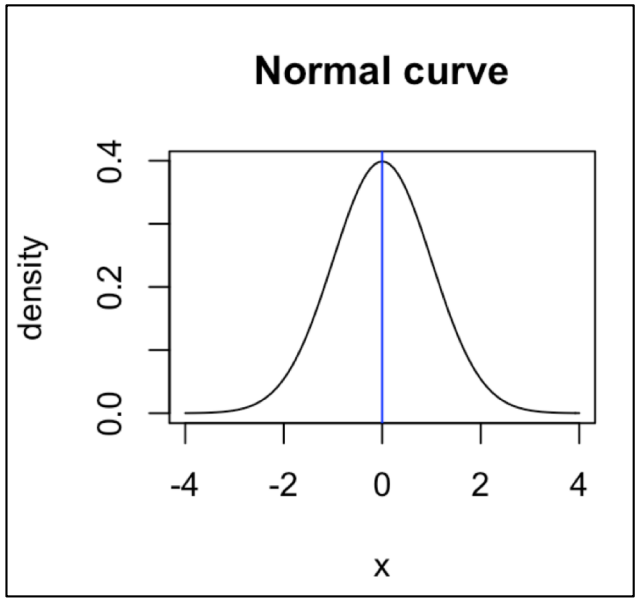
```
> pnorm(x)
[1] 0.5
```



`qnorm` - finds x value immediately to the right of given area

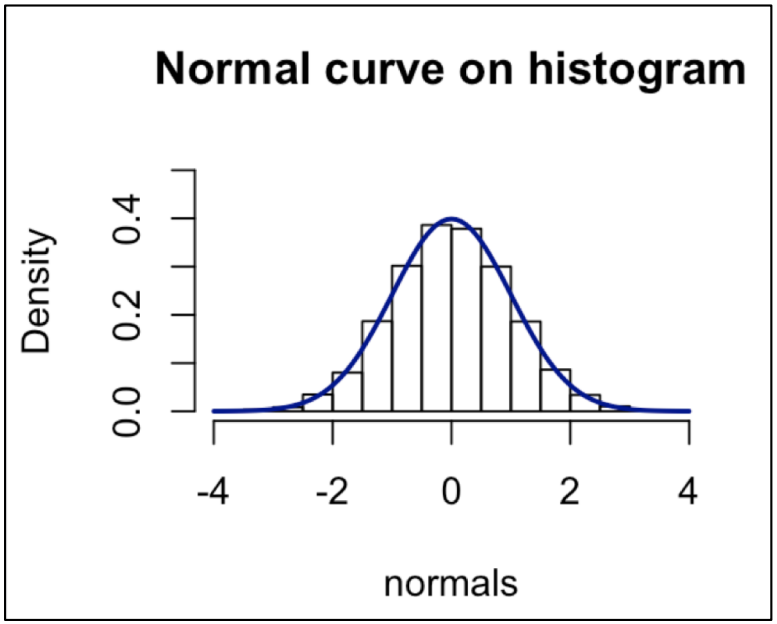
```
 8  plot(function(x) dnorm(x),
 9       ylab="density",
10       from=-4,
11       to=4,
12       main="Normal curve")
13  area=0.5
14  x=qnorm(area)
15  abline(v=x,col="blue")
```

```
> qnorm(area)
[1] 0
```



`rnorm` - simulate taking a random sample from the normal distribution

```
25  normals = rnorm(10000)
26  hist(normals, breaks=20, prob=TRUE,
27       ylim=c(0, 0.5),
28       main="Normal curve on histogram")
29  curve(dnorm(x),from=-4,to=4,
30        col="darkblue", lwd=2,
31        add=TRUE,
32        yaxt="n")
```



There are other "dpqr" functions for other distributions

- Use help to get info

```
> help("distribution")
```

Distributions {stats} R Documentation

Distributions in the stats package

Description

Density, cumulative distribution function, quantile function and random variate generation for many standard probability distributions are available in the **stats** package.

Details

The functions for the density/mass function, cumulative distribution function, quantile function and random variate generation are named in the form `dxxx`, `pxxx`, `qxxx` and `rxxx` respectively.

For the beta distribution see `dbeta`.

For the binomial (including Bernoulli) distribution see `dbinom`.

For the Cauchy distribution see `dcauchy`.

For the chi-squared distribution see `dchisq`.

How to solve CLT-related problems

1. Draw normal curve
2. Add vertical lines
   - mean or sample proportion
   - threshold lines
3. Shade relevant area

Problem-solving tip: First draw the problem. This will give you insight!

4. Pick R function
   - `pnorm` if finding area
   - `qnorm` if finding x value (location of vertical lines)
5. Define variables:
   - standard deviation
   - mean/proportion
   - test parameter
6. Run R function
7. Check result using inverse
   - pnorm if used qnorm
   - qnorm is used pnorm

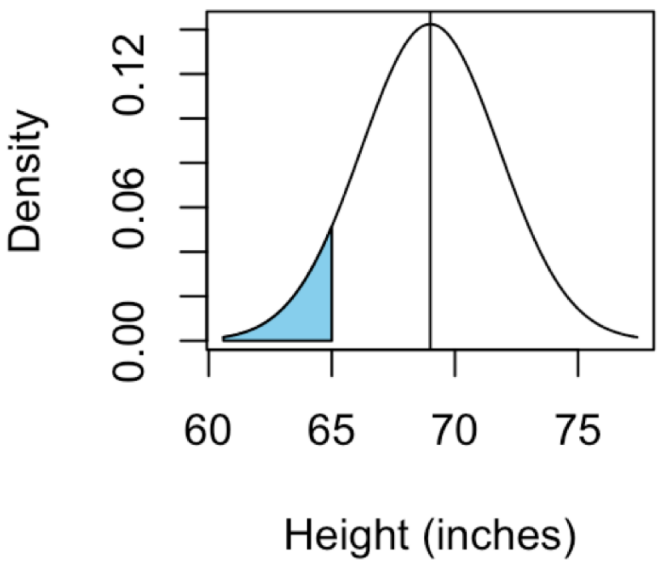
Adult male height (X) follows (approximately) a normal distribution with a mean of 69 inches and a standard deviation of 2.8 inches.

(a) What proportion of males are less than 65 inches tall? In other words, what is $P(X < 65)$?

from Stanford OLI reading: Probability: Continuous Random Variables > Normal Random Variables > Statistics Package Exercise: Using the Normal Distribution

Tip: Draw by hand before using R!

1,2,3 - draw, add lines & shading



4 - pick a function - pnorm or dnorm.
You know X , want area (blue) under curve.
Use pnorm (not qnorm)

5 - define function arguments: standard deviation & mean & test parameter

```
pop.mean=69
pop.stdev=2.8
test.height=65
```

6 - run function

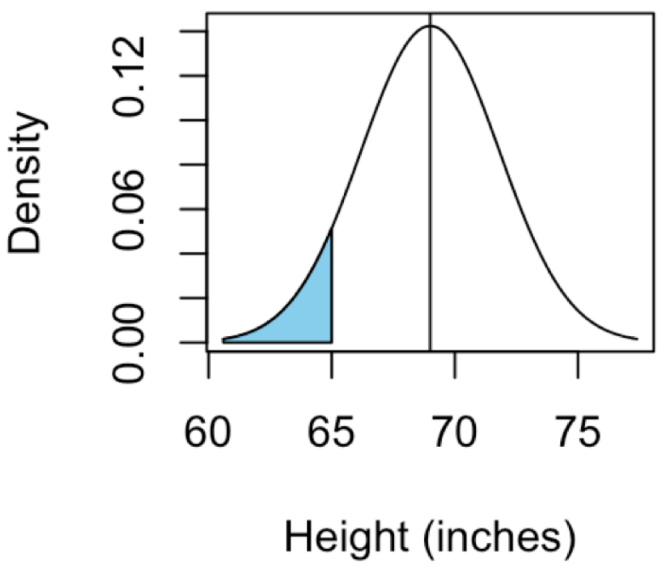
```
> pnorm(test.height,sd=pop.stdev,mean=pop.mean)
[1] 0.07656373
```

7 - run inverse to check result

```
> result=pnorm(test.height,sd=pop.stdev,mean=pop.mean)
> qnorm(result,sd=pop.stdev,mean=pop.mean)
[1] 65
```

1,2,3 - draw, add lines & shading

Advanced R plotting – how to draw the curve + shading




```
34  pop.mean=69
35  pop.stdev=2.8
36  test.height=65
37  curve(dnorm(x,mean=pop.mean,sd=pop.stdev),
38          xlim=c(pop.mean-3*pop.stdev,
39                 pop.mean+3*pop.stdev),
40        ylab="Density",xlab="Height (inches)")
41  cord.x=c(pop.mean-3*pop.stdev,
42          seq(pop.mean-3*pop.stdev,test.height,0.01),
43          test.height)
44  cord.y=c(0,
45           dnorm(seq(pop.mean-3*pop.stdev,test.height,0.01),
46                 mean=pop.mean,
47                 sd=pop.stdev),
48           0)
49  abline(v=pop.mean,col="black")
50  # Add the shaded area.
51  polygon(cord.x,cord.y,col='skyblue')
```

Adult male height (X) follows (approximately) a normal distribution with a mean of 69 inches and a standard deviation of 2.8 inches.

(b) What proportion of males are more than 75 inches tall? In other words, what is P(X > 75)?

In-class Problem 1

You solve it

from Stanford OLI reading: Probability: Continuous Random Variables > Normal Random Variables > Statistics Package Exercise: Using the Normal Distribution

Adult male height (X) follows (approximately) a normal distribution with a mean of 69 inches and a standard deviation of 2.8 inches.

(c) What proportion of males are between 66 and 72 inches tall? In other words, what is $P(66 < X < 72)$?

In-class Problem 2

You solve it

from Stanford OLI reading: Probability: Continuous Random Variables > Normal Random Variables > Statistics Package Exercise: Using the Normal Distribution

A random sample of 100 students is taken from the population of all part-time students in the United States, for which the overall proportion of females is 0.6.

(a) There is a 70% chance that the sample proportion falls between what two values?

In-class Problem 3

You solve it

adapted from Stanford OLI reading: Probability: Sampling Distributions > Sample Proportion > Behavior of Sample Proportion: Applying the Standard Deviation Rule

The proportion of left-handed people in the general population is about 0.1. Suppose a random sample of 225 people is observed.

(a) What is the probability that 40 or more people in the sample are left-handed?

In-class Problem 4

You solve it

adapted from Stanford OLI reading: Probability: Sampling Distributions > Sample Proportion > Behavior of Sample Proportion: Applying the Standard Deviation Rule

Upload your answers to Canvas

- Upload a knitted Markdown
- Show your code
  - How you answered the problem.
  - How you checked your answer.
  - Illustrative plots
    - Use R (best!) or draw them by hand and take a picture