STM1001 Lecture: Week 4

Amanda Shaker

STM1001 - Making Sense of Data

Introduction

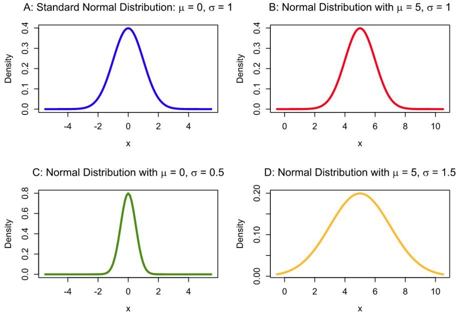
- In this lecture we will cover material that will be very useful for Computer Lab 4, Quiz 4 and Assignment 1
- By the end of this lecture you will know how to:
    - calculate probabilities and quantiles from the Normal Distribution in R using the `norm` functions
    - use the Central Limit Theorem to establish a distribution for $\overline{X}$ and use this distribution to calculate probabilities in R using the `norm` functions
    - calculate probabilities from the Binomial Distribution in R using the `binom` functions

The Normal Distribution

Recall from the Topic 4 readings that we express the distribution of a normally distributed random variable X as

$$X \sim N(\mu, \sigma^2)$$

Notice that we use the variance, σ^2, when specifying the distribution. Some examples from the Topic 4 readings:



Recall that σ is the standard deviation, and it is the square root of the variance, σ^2

Using the `norm` R functions

There are various functions related to the Normal Distribution in R:

`norm` **function**	**What the function does**
`pnorm(q, mean, sd)`	Calculates the **probability** that X is less than some number **q** for a Normal distribution with `mean` and `sd` as specified. That is, calculates $P(X \leq \mathbf{q})$ for $X \sim N($`mean, sd^2`$)$
`qnorm(p, mean, sd)`	Calculates the value **q** (**quantile**) at which we have $P(X \leq \mathbf{q}) = \mathbf{p}$ for $X \sim N($`mean, sd^2`$)$
`rnorm(n, mean, sd)`	Generates **n** **r**andom values from $X \sim N($`mean, sd^2`$)$
`dnorm(x, mean, sd)`	Calculates the **d**ensity at **x** of $X \sim N($`mean, sd^2`$)$

Be careful: When we use the various `norm` functions in R, we specify the standard deviation (`sd`), σ, rather than the variance, σ^2.

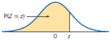
We will consider some examples shortly.

Why R?

Compare this...

Table III **Standard Normal Distribution Cumulative Probabilities**

Let Z be a standard normal random variable: $\mu = 0$ and $\sigma = 1$.
This table contains cumulative probabilities: $P(Z \le z)$.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

To this...

```r
pnorm(-1.5, mean = 0, sd = 1)
```

`pnorm`

Suppose we have $X \sim N(0, 1)$ and we would like to know $P(X \leq -1.5)$.

Before calculating this probability in R, it can be very useful to first draw a picture of the probability we wish to calculate:

pnorm

To calculate the probability, we can use the `pnorm` function as follows:

```
pnorm(-1.5, mean = 0, sd = 1)
```

- You may have noticed that since we have $X \sim N(0, 1)$, we are working with the **Standard Normal Distribution**
- If we do not specify `mean` and `sd`, this is the distribution R will assume we are working with
- That is, R will assume we have `mean = 0` and `sd = 1`
- Therefore, for this example, we could equivalently use the `pnorm` function as follows:

```
pnorm(1.5)
```

pnorm

Still assuming $X \sim N(0,1)$, now suppose we would like to know $P(-1.5 \leq X \leq -1)$.

First, draw a picture of the probability we wish to calculate:

pnorm

To calculate this probability, we can use:

```
pnorm(-1, mean = 0, sd = 1) - pnorm(-1.5, mean = 0, sd = 1)
```

pnorm

Now suppose $X \sim N(2, 5)$, and we would like to know $P(X \geq 3)$.

First, draw a picture of the probability we wish to calculate:

pnorm

To calculate this probability, we can use:

```
1 - pnorm(3, mean = 2, sd = sqrt(5))
```

pnorm

Still assuming $X \sim N(2, 5)$, now suppose we would like to know $P(X \leq 1) + P(X \geq 3)$.

First, draw a picture of the probability we wish to calculate:

pnorm

To calculate this probability, we can use:

```
pnorm(1, mean = 2, sd = sqrt(5)) + (1 - pnorm(3, mean = 2, sd =
    sqrt(5)))
```

Equivalently, we could make use of the symmetry property and calculate the probability as follows:

```
2 * pnorm(1, mean = 2, sd = sqrt(5))
```

qnorm

Still assuming $X \sim N(2, 5)$, now suppose we would like to know the value of x for which we have $P(X \leq x) = 0.6$.

First, draw a picture:

qnorm

The value of x we need to calculate can be thought of as a q**uantile**; hence why we use the `qnorm` function in R.

The probability of 0.6 will be the value of `p` that we input into the function:

```
qnorm(0.6, mean = 2, sd = sqrt(5))
```

Central Limit Theorem

In Topic 4, we considered how to ascertain the **distribution of the sample mean,** $\overline{X}$, under three different scenarios. Consider the following example:

Suppose it is claimed that amongst the population of STM1001 students, the number of hours of sleep in the past 24 hours is normally distributed with a mean of $\mu = 7.34$ and standard deviation of $\sigma = 1.93$. Further suppose that we wish to study the mean of the population by taking a random sample of $n = 34$ STM1001 students. Using this information, write down the distribution of the sample mean, $\overline{X}$.

Central Limit Theorem

Recall the Central Limit Theorem:

The Central Limit Theorem (CLT)

Let $X_1, \ldots, X_n$ be a random sample from a distribution with finite mean μ and finite variance σ^2. For $\overline{X}$ denoting the sample mean, if n is sufficiently large then

$$\overline{X} \overset{\text{approx.}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

where $\overset{\text{approx.}}{\sim}$ denotes 'approximately distributed as'.

Although, from the CLT, it is known that $\overline{X}$ **approximately** follows a normal distribution provided n is sufficiently large, for ease of notation and without loss of generality, from this point onwards we will use $\sim$ in place of $\overset{\text{approx.}}{\sim}$.

Central Limit Theorem

Solution:

- From the question, we know that the underlying distribution is Normal. Hence, the distribution of $\overline{X}$ will also be Normal. This is regardless of sample size
- From the CLT, we have that if $X \sim N(\mu, \sigma^2)$, then $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- From the question, we have that
  - $\mu = 7.34$
  - $\sigma = 1.93$, so that $\sigma^2 = 1.93^2 = 3.7249$
  - $n = 34$
- Now $\frac{\sigma^2}{n} = \frac{3.7249}{34} \approx 0.1096$
- Hence, we can write down the distribution of $\overline{X}$ as

$$\overline{X} \sim N\,(7.34, 0.1096)$$

Central Limit Theorem and `pnorm`

Now assuming $\overline{X} \sim N(7.34, 0.1096)$, suppose we would like to know $P(\overline{X} \leq 6.5)$. First, draw a picture:

Central Limit Theorem and pnorm

To calculate this probability, we can use:

```
pnorm(6.5, mean = 7.34, sd = sqrt(0.1096))
```

Or, equivalently, since $\sqrt{0.1096} \approx 0.3311$, we could use

```
pnorm(6.5, mean = 7.34, sd = 0.3311)
```

Remember that if we wish, we can use the `round` function to round our answer to, say, 4 decimal places:

```
round(pnorm(6.5, mean = 7.34, sd = 0.3311), 4)
```

The Binomial Distribution

- Recall the playing cards example introduced in the Topic 3 Workshop, where we guessed the suit of a playing card 10 times, with replacement, from a standard deck of playing cards
- Since there are 4 suits and each suit has the same number of cards, the probability of a correct guess was 0.25 each time
- We will refer to your number of correct guesses as X, which had a range from 0 up to 10
- We can quantify the probability associated with making a certain number of correct guesses using the *Binomial distribution*

The Binomial Distribution

The Binomial Distribution

Suppose we have n “trials”, each with an outcome of either “success” or “failure”. Further suppose that for each trial, the probability of “success” is equal to p, and that X is the number of “successes” from the n trials. We can model X using the *Binomial Distribution*, defining the distribution as

$$X \sim BIN(n, p),$$

where:

- X is the number of successes
- n is the number of trials
- p is the probability of success for each trial

Given in our playing cards example we have $n = 10$ trials with a probability of “success” of $p = 0.25$, we define the distribution in this example as

$$X \sim BIN(10, 0.25).$$

Using the `binom` R functions

There are various functions related to the Binomial Distribution in R:

`binom` **function**	**What the function does**
`dbinom(x, size, prob)`	Calculates $P(X = \mathbf{x})$ for $X \sim BIN(n = \texttt{size},\ p = \texttt{prob})$. This is the density, or *probability mass* for a discrete distribution, for a given value of **x**.
`pbinom(x, size, prob)`	Calculates $P(X \leq \mathbf{x})$ for $X \sim BIN(n = \texttt{size},\ p = \texttt{prob})$

* In R, the first argument in the `pbinom` function is called `q`. However in the above table, we have referred to it as `x` for ease of notation.

dbinom

Suppose we wish to know our chances of getting 8 correct guesses. That is, suppose we wish to know $P(X = 8)$.

Recalling that $X \sim BIN(10, 0.25)$, we can calculate this probability in R using:

```
dbinom(8, 10, 0.25)
```

pbinom

Now suppose we wish to know the probability of making **8 or less** correct guesses. That is, suppose we wish to know $P(X \leq 8)$.

This time, we will use `pbinom`:

```
pbinom(8, 10, 0.25)
```

pbinom

Now suppose we wish to know $P(X < 8)$.

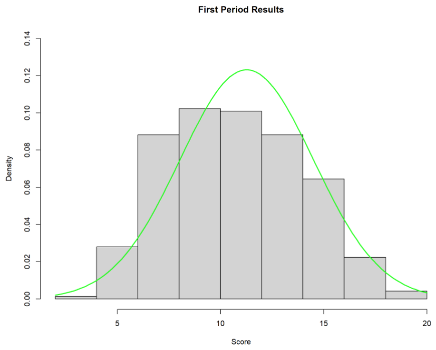
Since we are working with a discrete distribution, we must pay attention to whether we have a *leq* or $<$ sign. Since we want to know $P(X < 8)$, this is the same as $P(X \leq 7)$. Hence, we use:

```
pbinom(7, 10, 0.25)
```

We will consider more examples in Computer Lab 4.

Overlaying a Normal curve on a Histogram

Also in Computer Lab 4, we will learn how to overlay a Normal curve on a histogram:



This may be very useful for Assignment 1.