STAT22000 Autumn 2013 Lecture 3

Yibi Huang

March 23, 2014

1.3 Density Curves and Normal Distributions

Outline

- Density curves
  - area under a density curve
  - mean and median for density curves
- Normal distributions
  - The 68-95-99.7 rule
  - Using the standard normal table
  - Standardization
  - Inverse normal calculations
  - Normal quantile plots

Recall in a histogram

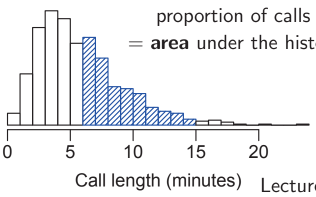
(**area** of a bin) $\propto$ (number of obs. in that bin).

By rescaling the height of bars as

$$\frac{\text{percentage of obs. in the bin}}{\text{bin width}},$$

we can make the **area** of bars *equal* to the percentages of of observations in the bins.

E.g., histogram of the lengths of 800 calls to a customer service center



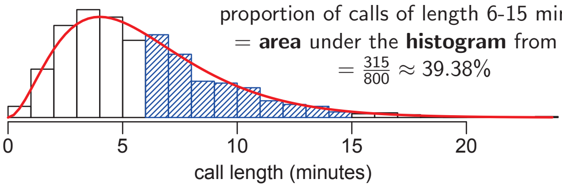
proportion of calls of length 6-15 mins $= \textbf{area}$ under the histogram from 6 to 15 $= \frac{315}{800} \approx 39.38\%$

call length	count of calls
0 - 1	12
1 - 2	52
2 - 3	99
3 - 4	116
4 - 5	108
5 - 6	83
6 - 7	89
7 - 8	68
8 - 9	39
9 - 10	37
10 - 11	32
11 - 12	18
12 - 13	14
13 - 14	12
14 - 15	6
15 - 16	4
16 - 17	5
17 - 18	3
18 - 19	1
19 - 20	0
20 - 21	1
21 - 22	0
22 - 23	0
23 - 24	1
total	800

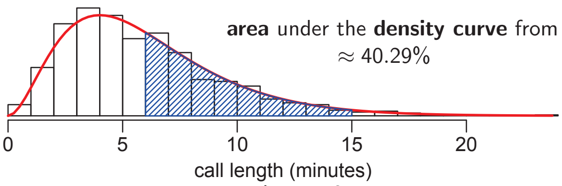
Density Curves

A **density curve** is a smoothed approximation of a histogram.

E.g., here is the histogram of the lengths of 800 calls to a customer service center.

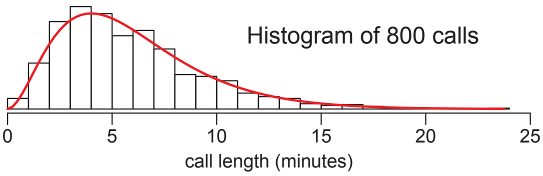


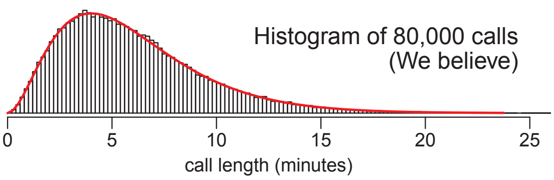
proportion of calls of length 6-15 mins $= \textbf{area}$ under the **histogram** from 6 to 15 $= \frac{315}{800} \approx 39.38\%$



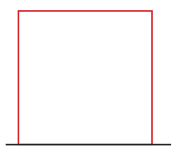
area under the **density curve** from 6 to 15 $\approx 40.29\%$

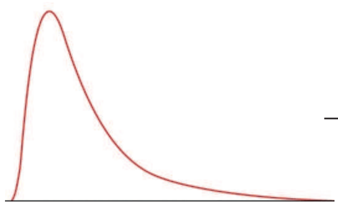
A density curve is also a **mathematical model** of a distribution. By "a model" we mean that if the data can be generated in the same way as in the original data to a larger size (e.g., by taking a larger sample, or repeating an experimental procedure more times, etc), we *believe* the histogram of the data will *approach the density curve*.

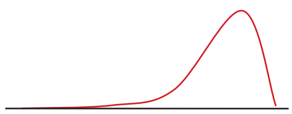




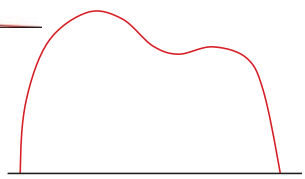
Density curves come in any imaginable shape.







Some are well known mathematically and others aren't.

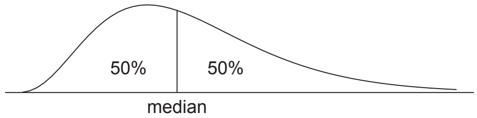


Properties of Density Curves

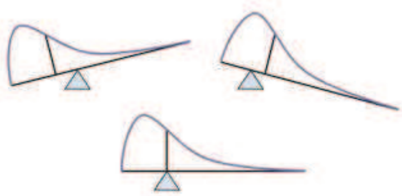
- A density curve is *nonnegative*,
  i.e., always on or above the zero line.
- The total area under the density curve is always 1, or 100%.

Mean and Median of a Density Curve (1)

The **median** of a density curve is the *equal-areas point*: the point that divides the area under the curve in half

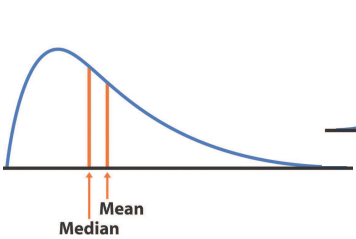


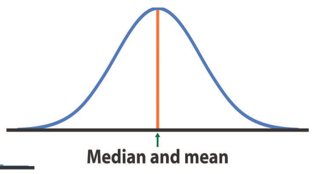
The **mean** of a density curve is the *balance point*, at which the curve would balance if it were made of solid material.



Mean and Median of a Density Curve (2)

The median and mean are the same for a symmetric density curve.
The mean of a skewed curve is pulled in the direction of the long tail.

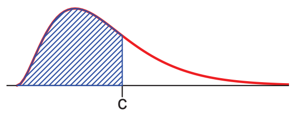




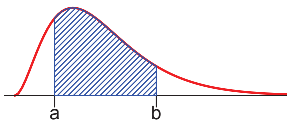
Area Under A Density Curve

For a density, not only the centers (mean, median) are important, in many cases, the distribution itself is more important, e.g., the percentage of 65+ people in a country is directly related to the social security budget of the government.

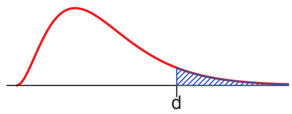
For a distribution, the **percentage of cases in a range** is represented by the **area** under the density curve for a range of values.



area of the shaded region
$=$ proportion of cases $< c$



area of the shaded region
$=$ proportion of cases
between a and b



area of the shaded region
$=$ proportion of cases $> d$

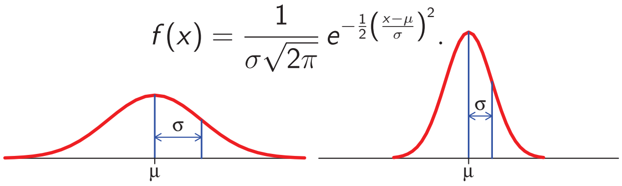
Normal Distributions

Normal distributions (aka. Gaussian distributions) are a family of *symmetric*, *bell- shaped* density curves defined by

a mean μ, and an SD σ

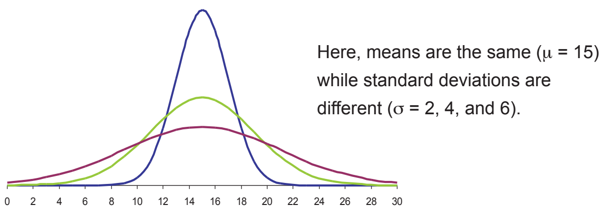
denoted as $N(\mu, \sigma)$. The formula for the $N(\mu, \sigma)$ curve is

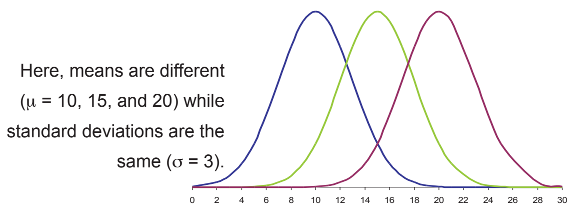
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$



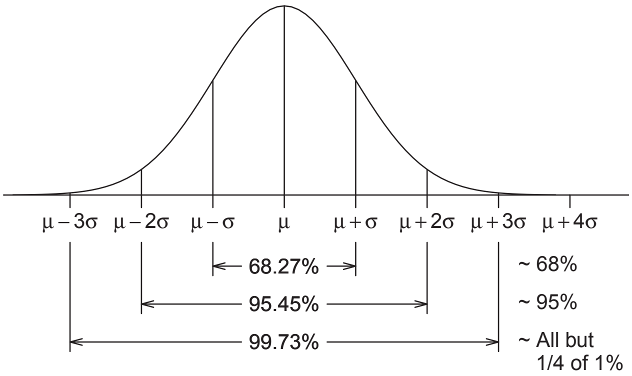
A normal distribution with $\mu = 0$, and $\sigma = 1$ is called the **standard normal distribution**, denoted as $N(0, 1)$.

Normal Family



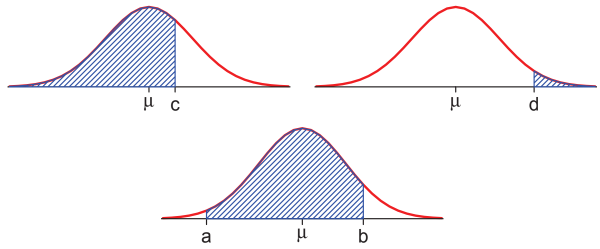


68-95-99.7% Rule for Normal Distributions



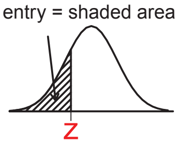
Normal Calculation

In statistics, a calculation that we will do from time to time is to find **areas** under a normal curve $N(\mu, \sigma)$, which represent proportions of observations from that Normal distribution.



Unfortunate there is no simple formula for areas under a Normal curve. We need to use either softwares or the **standard normal table** in the next 2 slides.

table entry = shaded area



The standard normal table gives the areas under the $N(0, 1)$ curve to the left of z.

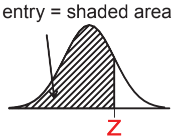
E.g., for $z = -0.83$, look at the row -0.8 and the column 0.03.

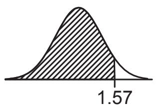
shaded area = 0.2033

Standard Normal Table (Table A at the end of the Textbook)

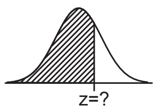
Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

table entry = shaded area





when $z = 1.57$
shaded area = 0.9418

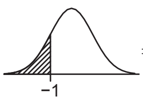


if shaded area = 0.75
then z = 0.675

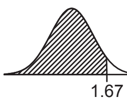
Standard Normal Table (continued)

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

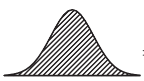
All the following curves are the standard normal. Use the standard normal table to find the area of the shaded regions.



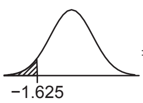
= 0.1587,



= 0.9525



= 1,

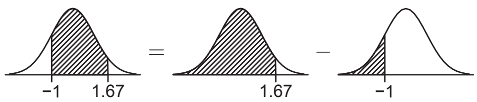


$= \frac{0.0526 + 0.0516}{2} = 0.0521$

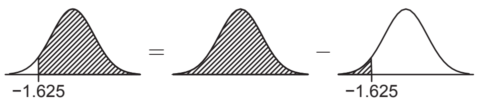
Alternatively, one can use the R command `pnorm()` to find areas under the standard normal $N(0, 1)$ curve.

```
> pnorm(-1)
[1] 0.1586553
> pnorm(1.67)
[1] 0.9525403
> pnorm(-1.625)
[1] 0.05208128
```


All the following curves are the standard normal $N(0, 1)$. Find the area of the shaded regions.



$= 0.9525 - 0.1587 = 0.7938$



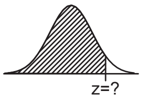
$= 1 - 0.0521 = 0.9479$

```
> pnorm(1.67) - pnorm(-1)
[1] 0.7938851
> 1-pnorm(-1.625)
[1] 0.9479187
```

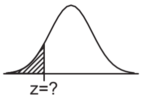
Alternatively, we can ask R to find the area in the UPPER tail.

```
> pnorm(-1.625,lower.tail=FALSE)
[1] 0.9479187
```

Conversely, we sometimes want to find the z for a given area.

If 

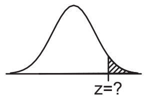
$= 0.95$, then $z =$ 1.645

If 

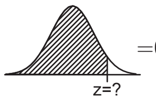
$= 0.10$, then $z =$ 1.28

The R command to find z for a given area under the $N(0,1)$ curve is `qnorm()`.

```
> qnorm(0.95)
[1] 1.644854
> qnorm(0.1)
[1] -1.281552
```


If 

$= 0.05$, then $z = ?$

Solution: This implies that 

$= 0.95$, so $z = 1.645$.

```
> qnorm(1-0.05)
[1] 1.644854
```

Alternatively, one can specify that 0.05 is the upper-tail area.

```
> qnorm(0.05,lower.tail=F)
[1] 1.644854
```

Now we know how to find area under a $N(0,1)$ curve using the normal table or R. How about a general normal curve $N(\mu, \sigma)$? This has to do with a **standardized value** or a z-**score**.
See next slide.

Standardized Value (aka. z-Scores) (1)

- For an observation x from a distribution with mean μ and SD σ, the **standardized value** of x says how many SDs x is above (+ sign) or below (− sign) the mean, i.e.,

$$z = \frac{x - \mu}{\sigma}.$$

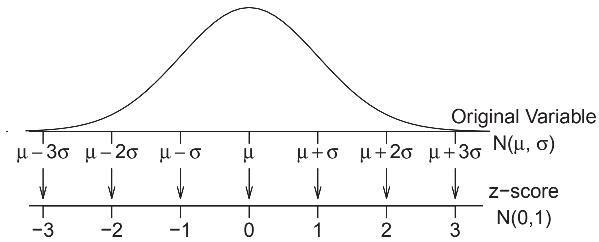
- A standardized value is often called a z-**score**.

E.g., Men's heights follow a $N(69'', 3'')$ distribution.

- A height of 66" has a z-score of $\frac{66''-69''}{3''} = -1$, i.e., a 66" tall man is 1 SD <u>below</u> the mean.
- A height of 74" has a z-score of $\frac{75''-69''}{3''} \approx 1.67$, i.e., a 74" tall man is 1.67 SD above the mean.
- What height has a z-score of +2? $69'' + 2 \times 3'' = 72''$.

Standardized Value (aka. z-Scores) (2)

Standardization is simply a change of units.

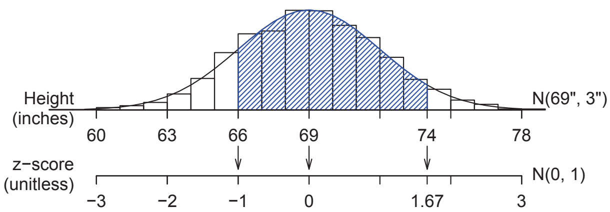


For a variable X with a normal distribution $N(\mu, \sigma)$, after standardization, its z-score $Z = \frac{X-\mu}{\sigma}$ has a standard normal distribution $N(0,1)$.

This is because all normal distributions have the same *shape*; differ only in center and scale.

Example: Men's Height

Men's heights follow a N(69", 3") distribution. What percent of men with heights between 66" and 74" (that is, 5'6" and 6'2"?)



- The percentage is equal to the area under the N(69", 3") curve between 66" and 74" (shaded region).
- The z-scores of $x = 66''$ and $x = 74''$ are respectively

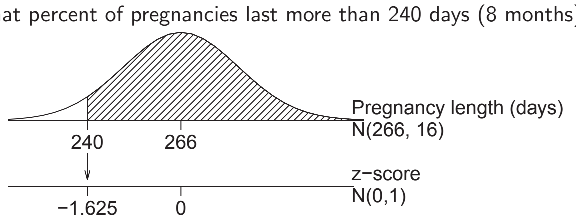
$$\frac{x - \mu}{\sigma} = \frac{66'' - 69''}{3''} = -1, \text{ and } \frac{74'' - 69''}{3''} = \frac{5}{3} = 1.67$$

- So ans = 

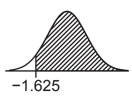
$= 0.7938 = 79.38\%$ (See Lecture 3-17)

Example: Length of Pregnancy

The length of the human pregnancy is not fixed. It is known that it varies according to a distribution which is roughly normal, with a mean of 266 days, and an SD of 16 days.
What percent of pregnancies last more than 240 days (8 months)?



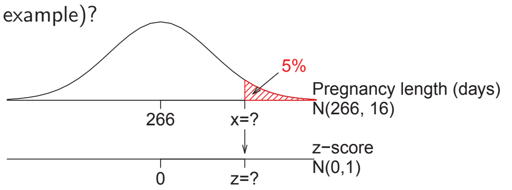
The z-score of 240 is $\dfrac{240 - 266}{16} = -1.625$.

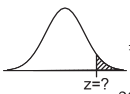
So the proportion is 

$= 0.9479$ (See Lecture 3-17)

Inverse Normal Calculation

How long are the longest 5% of pregnancies (in the pregnancy length example)?



Must find a z such that  $=0.05$, which was found in Lecture 3-18 to be 1.645. As $z = \frac{x-266}{16}$ is the z-score of the unknown x, we can find the value of x as

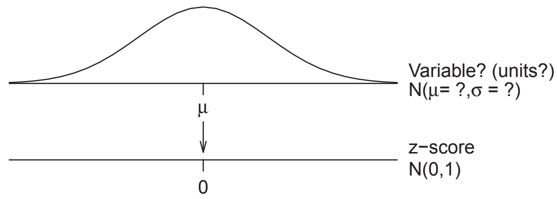
$$x = 266 + 16 \times z = 266 + 16 \times 1.645 = 292.32 \approx 292 \text{ days.}$$

The method:
standard normal curve $\rightarrow$ z-scores $\rightarrow$ original variable.

General Procedure for Normal Calculation

Draw the picture!

- Sketch the normal curve
- Put in the axis for the original variable
- Put in the axis for the z-scores
- Shade the area of interest
- Proceed



Follow this procedure on the HW, exercises, and exams!

Normal Quantile Plots (aka. Normal QQ Plots)

How to make a normal quantile plot?

1. Given data $(x_1, x_2, \ldots, x_n)$, arrange the data in increasing order: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.
2. Find quantiles of the $\mathrm{N}(0,1)$ distribution: $z_{(\frac{1}{n+1})}$, $z_{(\frac{2}{n+1})}$, $\ldots, z_{(\frac{n}{n+1})}$.
   That is, $z_{(\frac{i}{n+1})}$ is a value such that $\mathrm{P}(Z \le z_{(\frac{i}{n+1})}) = \frac{i}{n+1}$ for $Z \sim N(0,1)$.
3. Plot the $x_{(i)}$ values against the $z_{(\frac{i}{n+1})}$ values.
   That is, plot the points $(z_{(\frac{i}{n+1})}, x_{(i)})$ for $i = 1, 2, \ldots, n$

Interpreting Normal Probability Plots

- If the data are approximately normal, the plot will be close to a straight line.
- Systematic deviations from a straight line indicate a non-normal distribution.
- Outliers appear as points that are far away from the overall pattern of the plot.
- In R, use `qqnorm()`