

LECTURE 06:

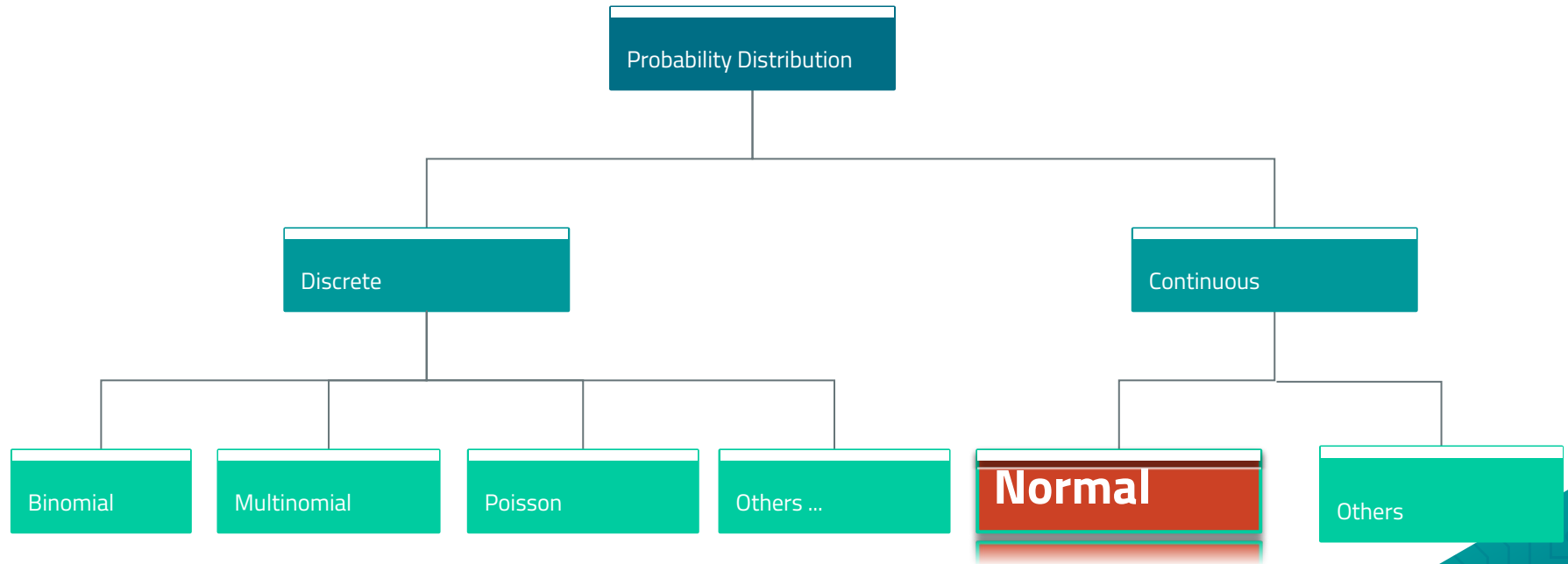
Normal Probability Distribution

Applied Statistics - PKN STAN - Class 5-37 & 5-38
3 & 5 November 2020
By Erika Siregar, SST, MS.

Reviews

1. Apa itu probability distribution?
2. How to visualize prob dist?
3. Random variable? And 2 types of random variable.
4. What can we measure from a probability distribution?
5. What is unusual?
6. Binomial → outcomes 2, discrete
7. Multinomial
8. Poisson
9. Kapan binomial bisa didekati dengan Poisson?

Probability Distribution

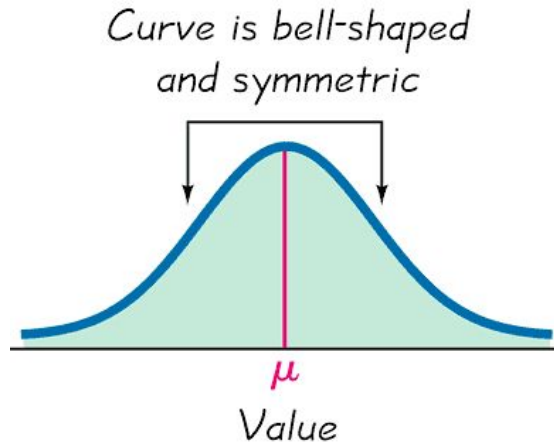


More details: https://en.wikipedia.org/wiki/Probability_distribution

NORMAL DISTRIBUTION

Normal Distributions

- A distribution that represent the probabilities of a **continuous** random variable where its histogram form a bell-shaped, **symmetric** curve.



Formula for prob dist:

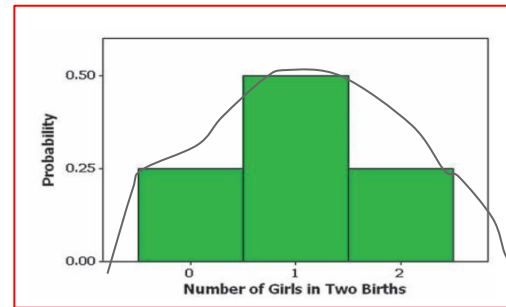
$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

Computing the Probability of Normal Dist.

- Utilize the density curve

Table 5-1 Probability Distribution for the Number of Girls in Two Births

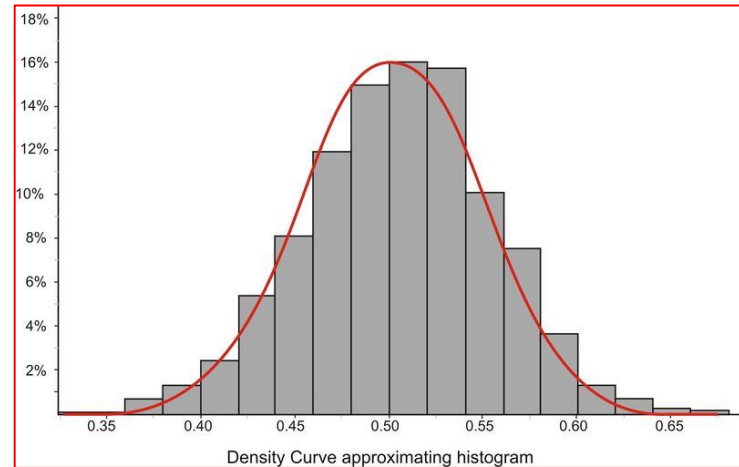
Number of Girls x	$P(x)$
0	0.25
1	0.50
2	0.25



- Remember that: $\sum P(x_i) = 1$
- Each bar in the histogram represents a probability for one particular value of random variable.
 - It implies that the total area of all bars in histogram = 1
- The curve enclose the whole histogram →
 - meaning: area of histogram \approx area under the curve
- The curve is called "**the density curve**".

Density Curve

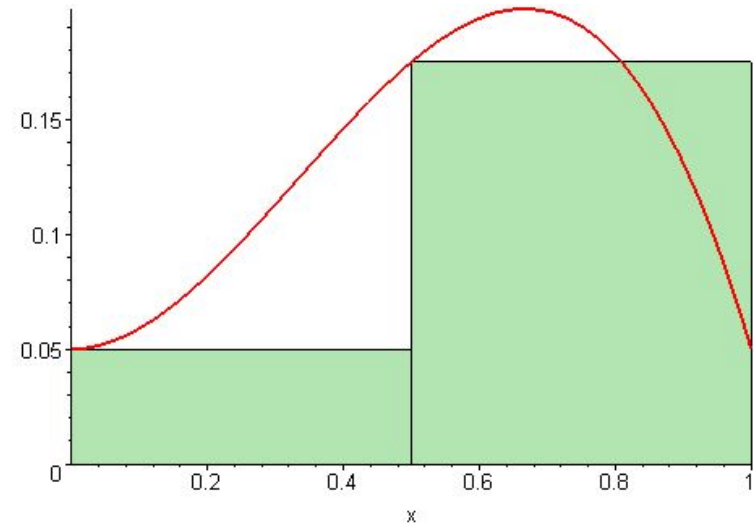
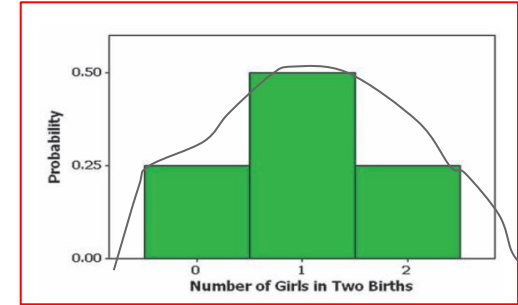
- **A curve represent prob dist.**
- **Requirements:**
 - ⊙ The **total area** under the curve must equal **1**.
 - ⊙ Every point on the curve must have a **vertical height ≥ 0** . (the curve cannot fall below the x-axis.)
- **Conclusion:**
 - ⊙ there is a correspondence between area and probability
 - ⊙ So we can compute the probability by **finding area under the curve instead of using the formula**



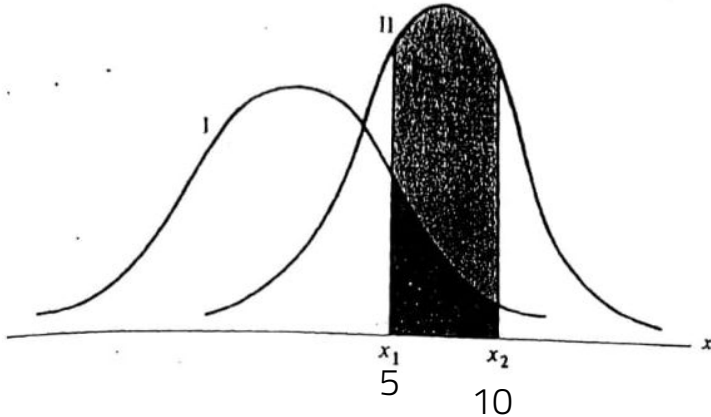
$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

Area under the curve

- Area under the curve can be used to estimate area of histogram
 - ◎ Meaning: we can use area under the curve to calculate probability
- The **more bars we have, the more precise our estimation is.**
- Semakin banyak = semakin tak terhingga
- Menjumlahkan sesuatu hingga jumlah tak terhingga = **integral**



Area under the curve (2)



$P(x_1 < X < x_2)$ = luas area di bawah kurva → **integral**

$$P(x_1 < X < x_2) = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} e^{-(1/2)((x-\mu)/\sigma)^2} dx$$

No need to worry about this complex formula → there's a **z-score table**, and of course, **R**, to save your life.

Standardized Normal Prob. Dist

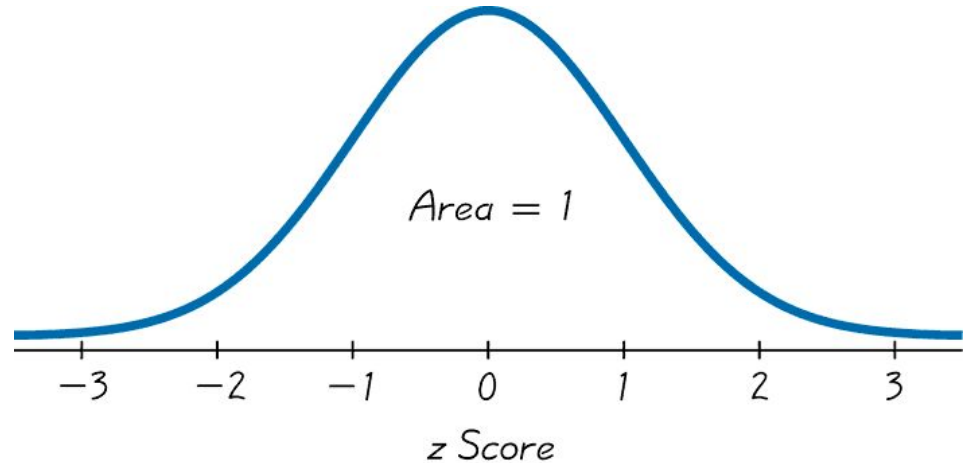
- Computing probability using normal dist. formula will be cumbersome
- It will be easier if the **$\mu = 0$** and **$\sigma = 1$** , so that we'll have

$$\begin{aligned}f(x) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} & -\infty \leq x \leq \infty \\&= \frac{1}{1 \times \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-0}{1}\right)^2} & -\infty \leq x \leq \infty \\&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} & -\infty \leq z \leq \infty \quad \left(\frac{x-0}{1} = z\right)\end{aligned}$$

Standardized Normal Prob. Dist

What is standardized?

- A **special case** of normal distribution where:
 - graph is bell-shaped.
 - $\mu = 0$
 - $\sigma = 1$
- What's so special about it?
 - We can easily find its probability using the **z**-score table (Triola p.585) → corresponding to the area under the graph.



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

Example

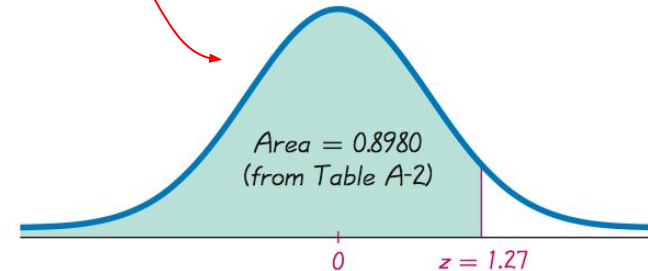
The Precision Scientific Instrument Company **manufactures thermometers** that are supposed to give readings of **0°C at the freezing point of water**. Tests on a large sample of these instruments reveal that at the freezing point of water, **some thermometers give readings below 0°** (denoted by negative numbers) and some give readings **above 0° (denoted by positive numbers)**.

Assume that the **mean reading is 0°C** and the **standard deviation of the readings is 1.00°C**. Also assume that the readings are **normally distributed**. If one thermometer is randomly selected, **find the probability** that, at the freezing point of water, the reading is **less than 1.27°**.

Answer:

TABLE A-2 (continued) Cumulative Area from the LEFT								
z	.00	.01	.02	.03	.04	.05	.06	.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292

$$P(z < 1.27)$$



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

How to Read the Normal Dist. Table

NEGATIVE z Scores

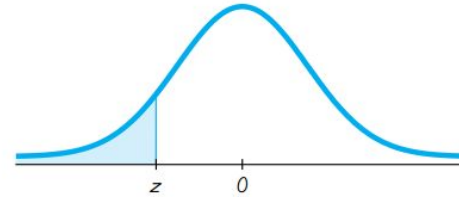
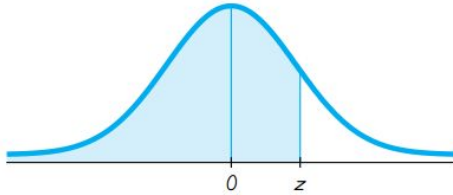


Table A-2 Standard Normal (z) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.50 and lower	.0001									
-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

How to Read the Normal Dist. Table



POSITIVE z Scores

$$Z = 0.33$$

Table A-2 (continued) Cumulative Area from the LEFT

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

Triola, halaman 585

Tabel z-score menunjukkan nilai kumulatif

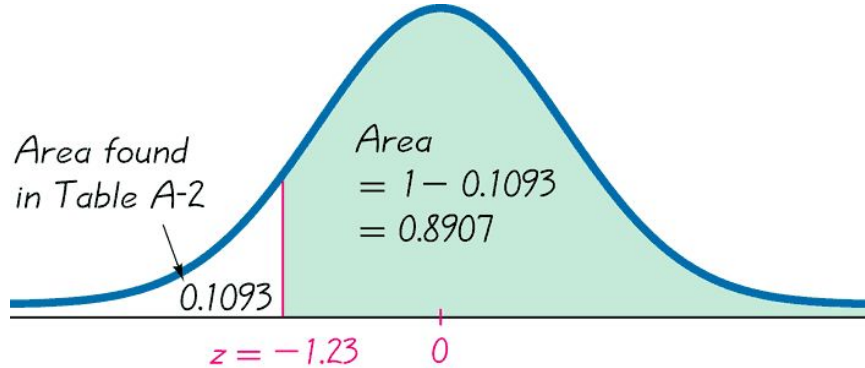
Try it with R

- `dnorm()` → peluang satu titik
- `pnorm()` → peluang cumulative
- `qnorm()` → given prob, what's the z-score?

Interpretation:

- The probability of randomly selecting a thermometer with a reading less than 1.27° is 0.8980
- Or 89.80% will have readings below 1.27°

How about $P(z > -1.23)$?



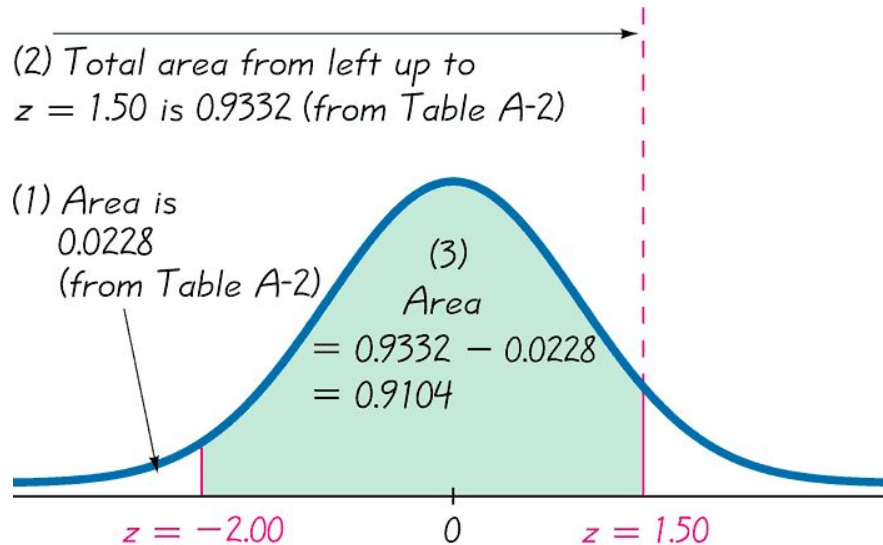
lower.tail=FALSE

Interpretation:

Probability of randomly selecting a thermometer with a reading above -1.23° is 0.8907.

$$1 - P(z < -1.23) = 1 - 0.1093 = 0.8907$$

How about $P(-2.00 < z < 1.50)$

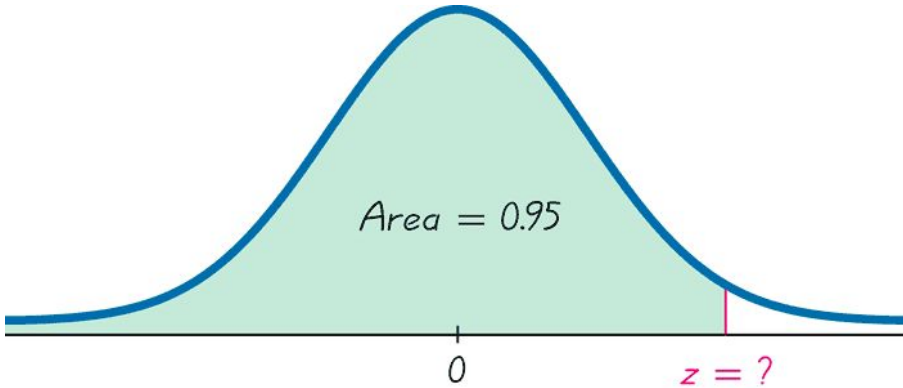


- $P(z < -2.00) = 0.0228$
- $P(z < 1.50) = 0.9332$
- $P(-2.00 < z < 1.50) = 0.9332 - 0.0228 = 0.9104$

The probability that the chosen thermometer has a reading between -2.00 and 1.50 degrees is 0.9104. \rightarrow 91.04%

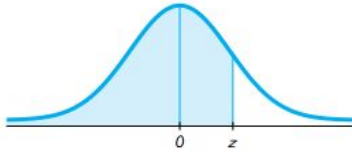
0.9104427

Given Probability, What's the z-score?



What is the z-score? → Finding the 95th Percentile?

Solution



POSITIVE z Scores

Table A-2 (continued) Cumulative Area from the LEFT

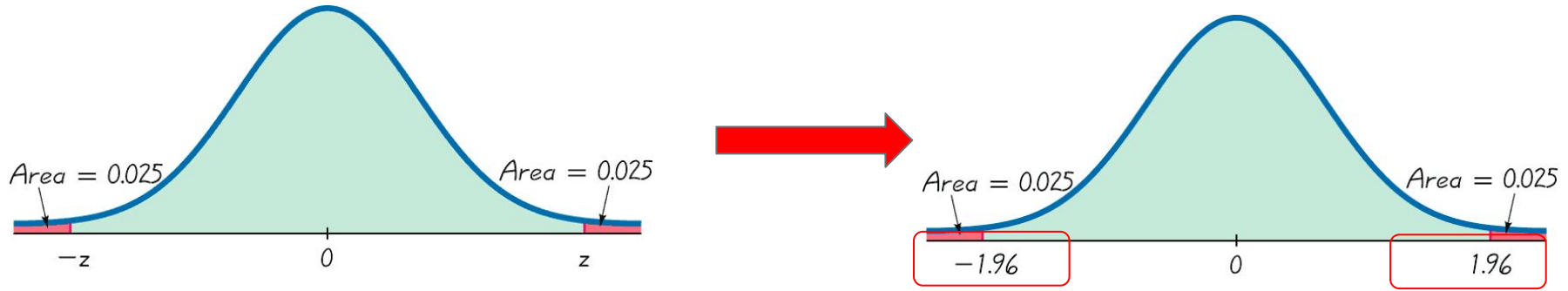
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.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

Dengan R:

```
> qnorm(0.95)  
[1] 1.644854
```

Dengan table

Finding the Bottom 2.5% and Upper 2.5%



One z score will be negative and the other positive

Dengan R

```
> qnorm(0.025)
[1] -1.959964
> qnorm(1 - 0.025)
[1] 1.959964
```

Cara lain: menggunakan tabel

DEALING WITH NON-STANDARDIZED NORMAL DISTRIBUTION

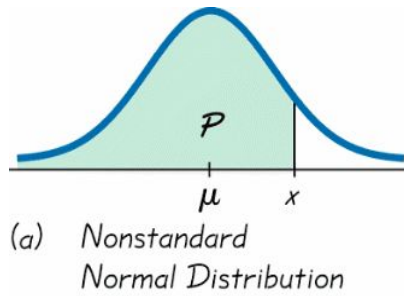
$Z \rightarrow \text{normal distribution} \rightarrow \mu = 0, \sigma = 1 \rightarrow \text{tabel}$

Dealing with non-standardized normal dist.

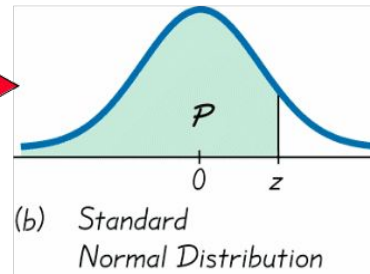
- Given a non-standardized normal distribution, how to find its probability using the z-score table?
- The answer: standardized it
- How to standardized?
 - Find its z-score

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

→ round to **2** decimal places

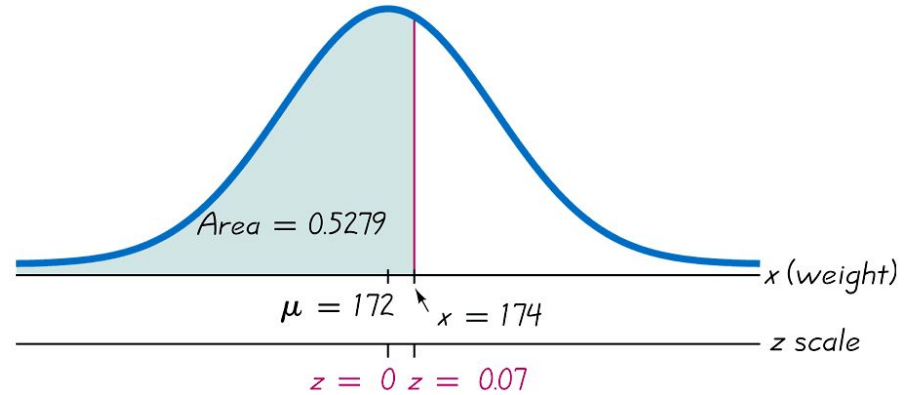


$$z = \frac{x - \mu}{\sigma}$$



Example

Assume that the weights of the **men** are **normally** distributed with a **mean of 172 pounds** and **standard deviation of 29 pounds**. If one man is randomly selected, what is the probability he weighs **less than 174 pounds**?



Try with R

```
> pnorm(zscore, mean = 0, sd = 1)
[1] 0.5274915
> pnorm(174, mean = 172, sd = 29)
[1] 0.5274915
> # with standardizing
> z <- (174-172)/29
> pnorm(z, mean = 0, sd = 1)
[1] 0.5274915
```

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

$$Z = \frac{174 - 172}{29}$$

$$z = \frac{174 - 172}{29} = 0.06896552 = 0.07$$

Cari probaility dari z-score = 0.07 di tabel → **0.5279**

$$P(x < 174 \text{ lb.}) = P(z < 0.07) = 0.5279$$

Reverse Thinking

Use the data from the previous example to determine **what weight separates the lightest 99.5% from the heaviest 0.5%?**

Answer:

For probability 0.995 $\rightarrow z = 2.575$

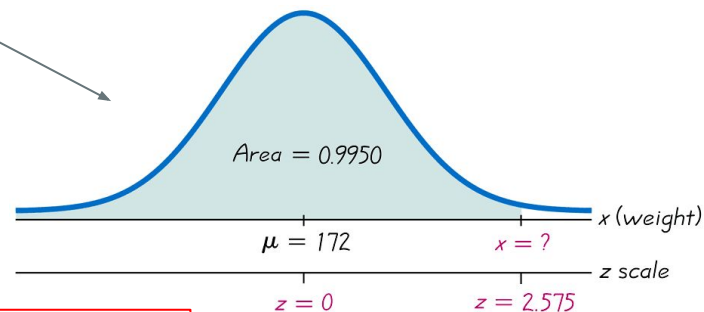
$$z = \frac{x - \mu}{\sigma}$$



$$\begin{aligned} x &= \mu + (z * \sigma) \\ x &= 172 + (2.575 * 29) \\ x &= 246.675 \text{ (247 rounded)} \end{aligned}$$

Try it with R

```
> qnorm(0.995)
[1] 2.575829
> z_0995 <- qnorm(0.995)
> print(z_0995)
[1] 2.575829
> x_0995 <- 172 + (z_0995 * 29)
> print(x_0995)
[1] 246.699
```




Interpretation:

The weight of 247 pounds separates the lightest 99.5% from the heaviest 0.5%

Practical Rules Commonly Used

1. For samples of size $n > 30$, the distribution of the sample means **can be approximated reasonably well by a normal distribution**. The approximation **gets closer to a normal distribution** as the sample size n becomes larger.

Normal as Approximation to Binomial

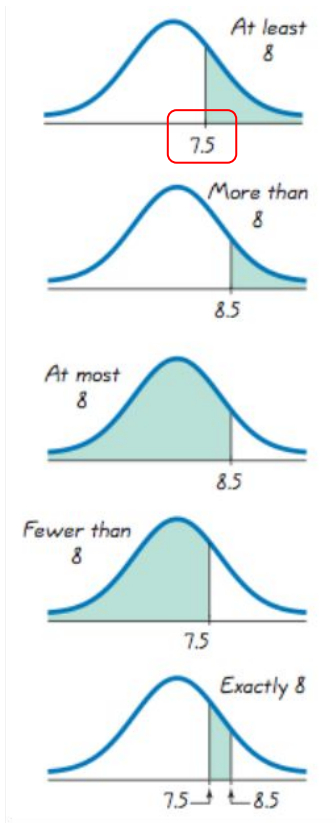
- ⊙ Requirements:
 - ⊙ $np \geq 5$
 - ⊙ $nq \geq 5$
- ⊙ Then binomial can be approximated by normal, where:
 - ⊙ $\mu = np$
 - ⊙ $\sigma = \sqrt{npq}$
- ⊙  $z = \frac{x - np}{\sqrt{npq}}$
- ⊙

Caution

When we use the **normal distribution** (which is a **continuous** probability distribution) as an approximation to the **binomial** distribution (which is **discrete**), a continuity **correction** is made to a discrete whole number x in the binomial distribution by representing the discrete whole number x by the interval from

$$x - 0.5 \text{ to } x + 0.5$$

(that is, adding and subtracting 0.5).



$X =$ at least 8
(includes 8 and above)

$X =$ more than 8
(doesn't include 8)

$X =$ at most 8
(includes 8 and below)

$X =$ fewer than 8
(doesn't include 8)

$X =$ exactly 8

Example:

$$z = \frac{x - np}{\sqrt{npq}}$$

Peluang seseorang sembuh dari suatu penyakit adalah **0.4**. Bila diketahui ada **100** orang yang telah terserang penyakit ini, berapa peluangnya bahwa kurang dari **30** orang yang sembuh?

$$p(x < 30) \rightarrow z = (29.5 - 40) / \sqrt{24}$$

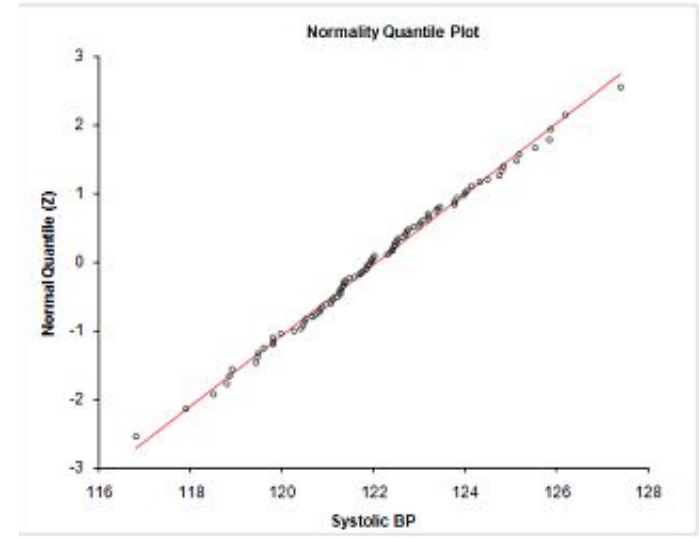
Answer:

Kasus binomial $\rightarrow np = 100 * 0.4 = 40$ dan $nq = 100 * 0.6 = 60 \rightarrow \geq 5$, bisa didekati dengan normal.

$$\begin{aligned} \mu &= np = 100 * 0.4 = 40 \\ \sigma &= \sqrt{npq} = \sqrt{100 * 0.4 * 0.6} = 4.899 \\ z &= \frac{29.5 - 40}{4.899} = -2.14 \\ P(x < 30) &\approx P(z < -2.14) = 0.0162 \end{aligned}$$

Nice to know: Assessing Normality

1. **Histogram:** Construct a histogram. Reject normality if the histogram departs dramatically from a **bell shape**.
2. **Outliers:** Identify outliers. Reject normality if there is **more than one outlier present**.
3. **Normal Quantile Plot (NQP):** If the histogram is basically **symmetric** and there is **at most one outlier**, points is reasonably **close to a straight line**
 - a. NQP = a graph of points (x,y) , where x = **original set of sample data**, and y = the corresponding **z score**



Refreshing Questions (1)

1. Apa ciri utama dari distribusi normal?
Kurvanya simetris, mean-nya ada di tengah
2. Apakah distribusi normal **selalu** memiliki $\mu = 0$ dan $\sigma = 1$? \rightarrow Tidak
3. Berapakah skewness dari distribusi normal? $\rightarrow 0$
4. Waktu yang dibutuhkan untuk memperbaiki mesin pengepak makanan mengikuti distribusi **normal** dengan rata-rata 120 menit dan **varians** 16 menit. Berapakah probabilitas bahwa mesin tsb dapat diperbaiki dalam waktu kurang dari 125 menit?

`pnorm(125, 120, 4)`

$$P(x < 125) \rightarrow z = (125 - 120) / 4 = 1.25 \rightarrow P(z < 1.25) = 0.8944$$

Refreshing Questions (2)

5. Banyaknya suatu produk terjual per hari mengikuti distribusi normal dengan rata-rata 20 dan simpangan baku 3. Berapakah:

a. Peluang bahwa penjualan akan kurang dari 16?

$$P(x < 16) \rightarrow z = (16 - 20) / 3 = -1.33 \rightarrow P(z < -1.33) = 0.0918$$

b. Peluang bahwa penjualan akan lebih besar dari 18?

$$P(x > 18) \rightarrow z = (18 - 20) / 3 = -0.67 \rightarrow P(z < -0.67) = 0.2514 \rightarrow 1 - 0.2514 = 0.7486.$$

```
> pnorm(18, 20, 3, lower.tail = FALSE)
```

```
[1] 0.7475075
```

c. Peluang bahwa penjualan akan terletak antara 17 dan 22?

$$p(17 < x < 22) \rightarrow z_{17} = (17 - 20) / 3 = -1 \rightarrow z_{22} = (22 - 20) / 3 = 0.67 \rightarrow P(z_{17}) =$$

$$0.1586553 \rightarrow P(z_{22}) = 0.7475075 \rightarrow 0.5909$$

d. Peluang bahwa penjualan akan berada antara 15 dan 19?

Refreshing Questions (3)

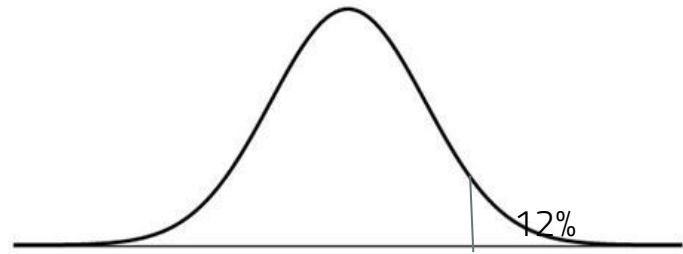
5. Nilai **rata-rata** dalam suatu nilai ujian adalah **74** dan **simpangan bakunya 7**. Bila **12%** dari peserta ujian mendapat **A** dan nilai ujian mengikuti distribusi **normal**, berapakah **kemungkinan** nilai A yang terkecil dan nilai B yang tertinggi?

`qnorm(0.88)`

$$Z = 1.175 \rightarrow x = \mu + (z * \sigma) = 74 + (1.175 * 7) = 82.225$$

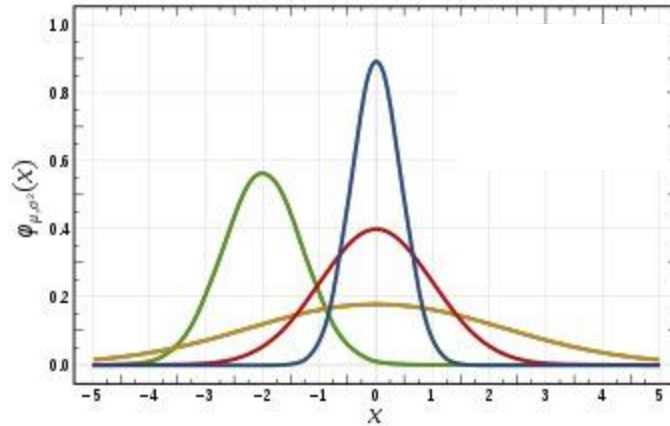
$$z = \frac{x - \mu}{\sigma}$$

Nilai A terkecil = 83
Nilai B terbesar = 82



Refreshing Questions (4)

6. Mana yang bukan merupakan kurva normal?



Refreshing Question (5)

7. Sebuah ujian pilihan ganda terdiri atas 80 soal, masing-masing dengan 4 pilihan dan hanya 1 jawaban yg benar. Tanpa memahami sedikitpun masalahnya dan hanya menerka saja, berapa peluang seorang murid menjawab **25 sampai 30** soal dengan benar?

Kasusnya binomial.

$n=80, p=0.25, q=0.75 \rightarrow np = 20, nq = 60$

$$z = \frac{x - np}{\sqrt{npq}}$$

$P(25 \leq x \leq 30)$

$Z_{25} = (24.5 - 20) / 3.87 = 1.16$

$Z_{30} = (30.5 - 20) / 3.87 = 2.71$

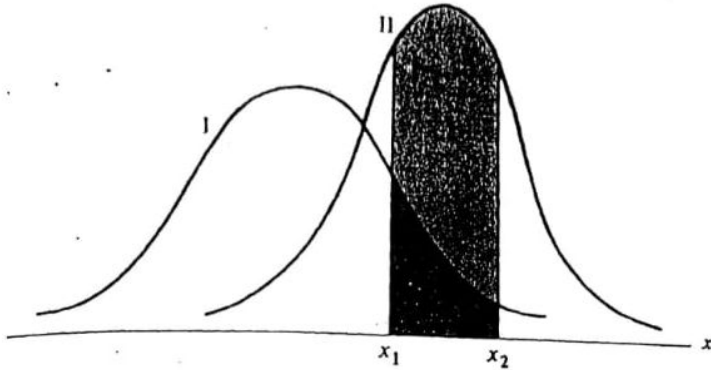
$P(z < 1.16) = 0.8770 \rightarrow P(z < 2.71) = 0.9965 \rightarrow P(1.16 \leq z \leq 2.71) = 0.9965 - 0.8770 = \mathbf{0.1195}$

Discrete didekati dengan continuous.



THANKS!

Area under the curve (2)



- Distribusi peluang, bentuknya seperti apa?
- Bisa digambarkan sbg apa?

$P(x_1 < X < x_2)$ = luas area di bawah kurva → **integral**

$$P(x_1 < X < x_2) = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} e^{-(1/2)((x-\mu)/\sigma)^2} dx$$

No need to worry bout this complex formula → there's a **z-score table**, and of course, **R**, to save your life.

Distribusi peluang → histogram peluang → tarik garis khayal yg menghubungkan puncak histogram → smooth it → jadi kurva → jika kurvanya simetris dan berbentuk lonceng → ini adalah kurva normal → ingat bahwa $x.p(x) = 1$ dan karena area dibawah kurva ini menggambarkan total peluang untuk semua → berarti luas area dibawah kurva = peluang. → bgmn menghitung area dibawah kurva? Inget bahwa sesungguhnya kurva ini bisa kita bagi menjadi beberapa bars. Jumlahkan luas untuk semua bar → dapat estimasi luas kurva → keep decreasing the bar width → semakin kecil lebar bar, semakin tepat estimasi luas kurvanya → jumlah dari sesuatu yg mendekati tak terhingga = integral