Math 183
Statistical Methods

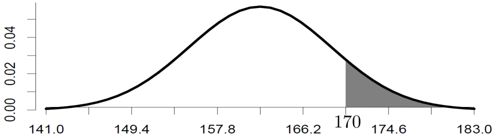
Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today: Chapter 3 (end)

- z-score as a way to standardize data
- Finding areas under normal curves using: R, standardization, and z-tables
- The $68 - 95 - 99.7\%$ rule
- Explore the reverse area problem

Areas under the Gauss Curve



Last class: R provided us with an answer with the function `pnorm`.

```
> pnorm(170, mean = 162, sd = 7, lower.tail = F)
[1] 0.126549
```

$$P(X \geq 170) \simeq 12.65\%.$$

This class: We'll do without technology, using the so-called z-table.

z-score

You are in charge of admissions at UCSD. One applicant took the SAT and got a 1775. Another took the ACT and got 27. Which would you admit?

When data are measured on different scales (= have different units), we need a common way to compare them that is unitless.

The z-score of a data point y from a dataset is $\dfrac{y - \bar{y}}{s_y}$.

The z-score:

- Is a unitless idea (units in numerator and denominator cancel)
- Tells you how many standard deviations above the mean some piece of data is.

z-score: Example

You see on Google that the SAT has mean 1500 with SD 250, and the ACT has mean 20.8 with SD 4.8. Which do you admit, the 1775 SAT or 27 ACT?

$$z_{SAT} = \frac{1775 - 1500}{250} = 1.1 \qquad z_{ACT} = \frac{27 - 20.8}{4.8} \simeq 1.29.$$

Assuming the two tests are equally difficult, you'd ratther admit the ACT person.

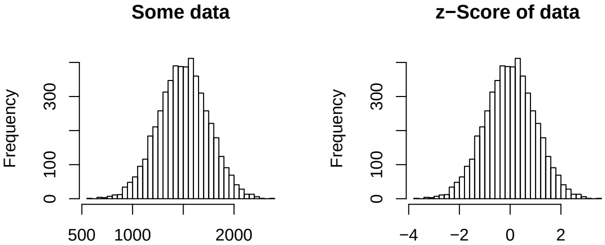
z-scores provide a single "ruler idea" to measure all phenomena, erasing the effect of units.

The z-score says how extreme a data point is relative to its own data set.

z-score

Let's take a data set and find the z-score for every data point.

$$\mu_{SAT} = 1500, \qquad \sigma_{SAT} = 250$$



x

$$z_x = \frac{x - \mu_{SAT}}{\sigma_{SAT}}$$

It appears:

- The new histogram is similar
- The new mean is 0.
- The new standard deviation is 1.

Are These Claims True?

Suppose you data set collects random variables X with mean μ and standard deviation σ.

By moving to z-scores, we create a new random variable $Z = \dfrac{X - \mu}{\sigma}$.

Note that

$$\begin{aligned} E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}E(X-\mu) &= \frac{1}{\sigma}(E(X)-\mu) \\ &= \frac{1}{\sigma}(\mu-\mu) = 0. \end{aligned}$$

Also,

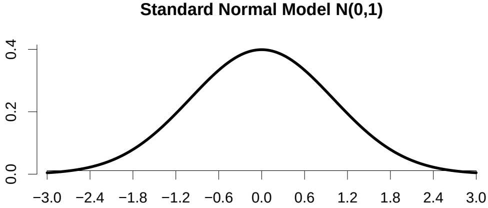
$$\begin{aligned} SD(Z) = SD\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}SD(X-\mu) &= \frac{1}{\sigma}SD(X) \\ &= \frac{1}{\sigma}\sigma = 1. \end{aligned}$$

z-Score Summary

- z-scores allow us to compare two data points from different data sets (with different centers and spreads) and get a sense for which datum is more extreme relative to its own data set.
- z-scores allow us to rescale a given data set so it has mean 0 and standard deviation 1.
  In the case of probability models, this allows us to think about whole families of curves using a single "standardized" model.

The Standard Normal Model

By rescaling data with the z-score, we turn all Normal models $N(\mu, \sigma)$ into a single one: the **Standard Normal Model** $N(0, 1)$.



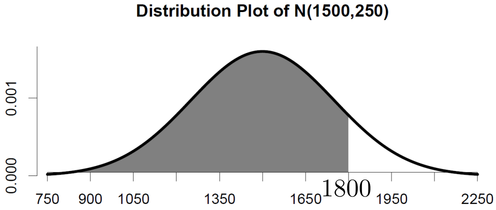
Any question in the original setting can be reframed as a question on the standard Normal model $N(0, 1)$.

If we understand $N(0, 1)$, we understand all the Normal models.

One Question, Many Ways to Solve

Find the percentage of students that has an SAT score below 1800.

A priori, we could find this probability by finding the area under the density curve using an integral. For normal distributions, this is too difficult to do by hand.

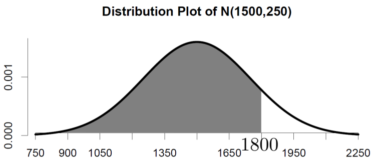


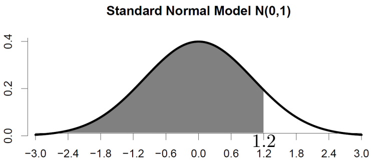
Option 1: Original setting + technology. $X = N(1500, 250)$, and R outputs

```
> pnorm(1800, mean = 1500, sd = 250, lower.tail = T)
 [1] 0.8849303
```

$$P(X \leq 1800) \simeq 88.49\%.$$

One Question, Many Ways to Solve


Distribution Plot of N(1500,250)


Standard Normal Model N(0,1)

Option 2: Standardized setting + technology.

Let $X = N(1500, 250)$, and define $Z = \dfrac{X - 1500}{250}$. We see that

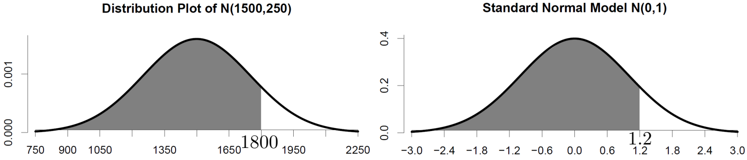
$$P(X \leq 1800) = P\left(Z \leq \frac{1800 - 1500}{250}\right) = P(Z \leq 1.2)$$

$$\simeq 88.49\%.$$

```
> pnorm(1.2, mean = 0, sd = 1, lower.tail = T)
 [1] 0.8849303
```

Remark: If you have technology, this method is silly. You can stay in the original setting.

One Question, Many Ways to Solve



Option 3: Standardized setting + tables.

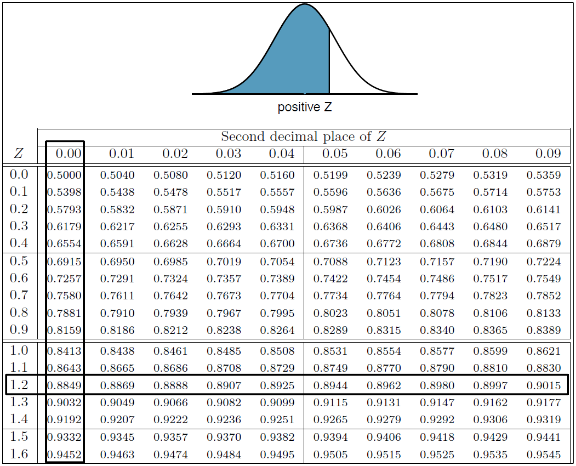
We found $P(X \leq 1800) = P(Z \leq 1.2)$.

Any question abount Normal curves can be converted to an equivalent question on the standard normal curve. We just need a lookup table of "areas under the curve" for the standard normal!

On tests, you won't have access to R, so you **will have to** use this approach.

We want the area less than $z = 1.20$.

z-Table (see textbook p. 427-429)

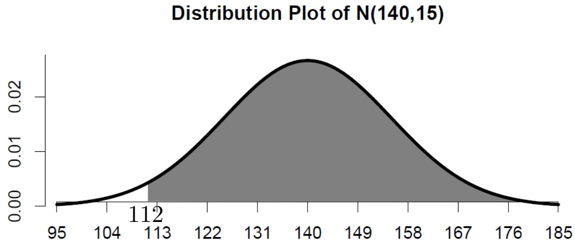


	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545

The values in the table are all areas (probabilities) The number along the top and left side are the $z - value$ broken into its two parts.

Here, $z = 1.20 = 1.2 + 0.00$, so we find $P(Z \leq 1.20) \simeq 0.8849$.

10-year-olds, regardless of gender, have heights (in cm) well-modeled by $N(140, 15)$. What percentage of 10-year-olds can ride Disneyland's Space Mountain, which has a height requirement of 112cm?

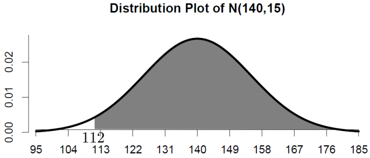


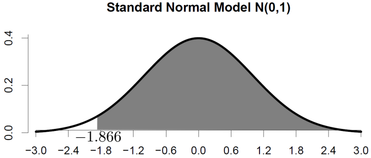
Remark: Always define your random variable and draw a picture!

Let $X = N(140, 15)$. We want $P(X \geq 112)$.

```
> pnorm(112, mean = 140, sd = 15, lower.tail = F)
 [1] 0.9690259
```

Practice With Tables

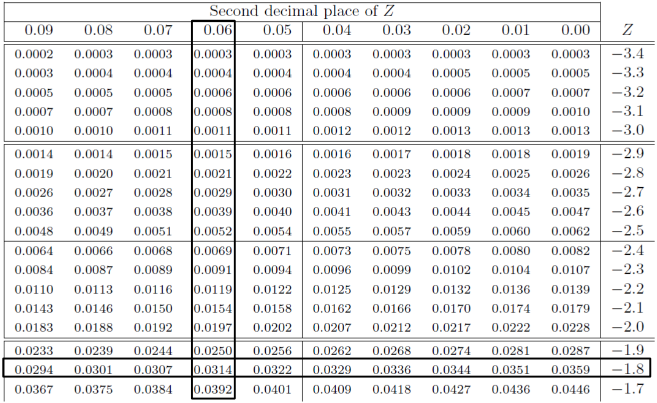




Writing $Z = \dfrac{X - 140}{15}$, we see that we want

$$\begin{aligned} P(X \geq 112) &= P\left(Z \geq \frac{112 - 140}{15}\right) \\ &= P(Z \geq -1.866) \\ &= 1 - P(Z < -1.866) \end{aligned}$$

Practice With Tables



Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7

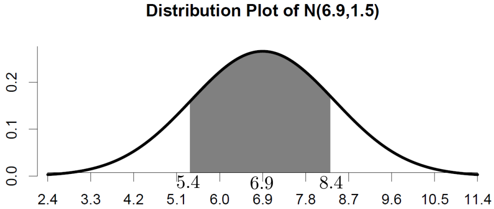
We read $P(Z < -1.86) \simeq 0.0314$, so

$$P(X \geq 112) \simeq 1 - P(Z < -1.866) \simeq 1 - 0.0314 = 0.9686.$$

Remark: On these types of problems, never worry about rounding issues or slight difference in answers, we just want approximations.

Sleep Time

The sleep times (in hours) of American men on a weekday are well-modeled by $N(6.9, 1.5)$. What percentage of American men are within 1 standard deviation of the mean sleep time?



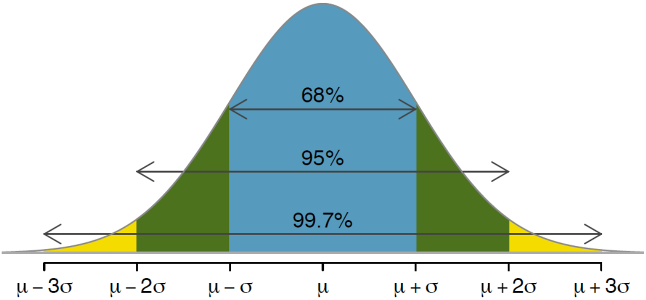
Let $X = N(6.9, 1.5)$. We want $P(5.4 \leq X \leq 8.4)$.
To find this, we solve

$$P(5.4 \leq X \leq 8.4) = P(X \leq 8.4) - P(X < 5.4).$$

```
> pnorm(8.4,6.9,1.5)-pnorm(5.4,6.9,1.5)
[1] 0.6826895
```

About 68% American men are within one σ of the mean.

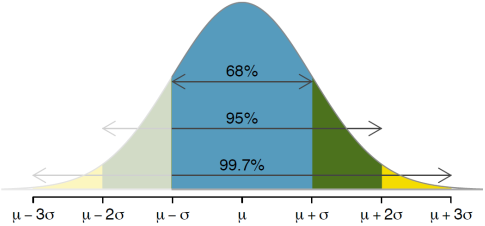
The 68 – 95 – 99.7% Rule



This holds for any data set that is normally distributed.

```
> pnorm(1,0,1)-pnorm(-1,0,1)
 [1] 0.6826895
> pnorm(2,0,1)-pnorm(-2,0,1)
 [1] 0.9544997
> pnorm(3,0,1)-pnorm(-3,0,1)
 [1] 0.9973002
```

A Question using the 68 – 95 – 99.7% Rule



What percentage of students score above a 1250 on the SAT? ($\mu = 1500$, $\sigma = 250$)

Notice that 1250 is one SD below the mean.

The area outside the 1 SD windows is $100 - 68 = 32\%$.

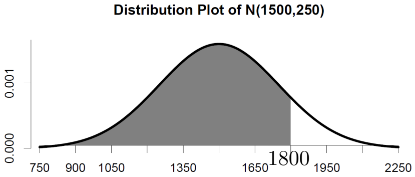
So, the area to the left of -1 SD is $32/2 = 16\%$ (by symmetry)

The desired percentage is $100 - 16 = \boxed{84\%}$.

Percentiles and “Going Backwards”

For any x value (or z-score, if you convert to a standard normal), the **percentile** is simply the area to the left of this value.

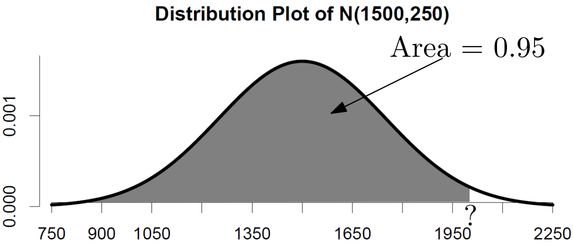
Example: the value $x = 1800$ on the SAT is about the 88th percentile.



```
> pnorm(1800, mean = 1500, sd = 250, lower.tail = T)
[1] 0.8849303
```

Percentiles and “Going Backwards”

Suppose a college only takes students who reach the 95th percentile (or better) on the SAT. What cutoff must you attain?



In R, the `qnorm` function solves the “reverse area problem”.

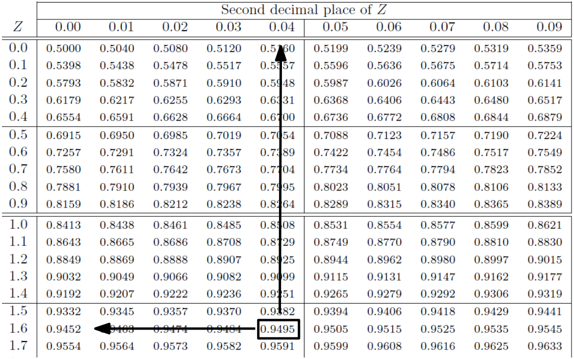
`pnorm`(x value) = area under curve up to that x value.

`qnorm`(area) = x value to the left of which, is the given area.

```
> qnorm(0.95,mean = 1500, sd = 250)
 [1] 1911.213
```

You must score above $\boxed{1911.22}$ to get into this college!

“Going Backwards” with Tables



	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

We find the area in the table that gives 95%: $z_{95\%} = 1.64$
Since the initial model is $X = N(1500, 250)$, we get

$$z_{95\%} = \frac{x_{95\%} - 1500}{250},$$

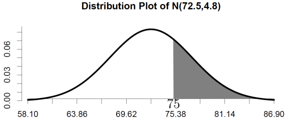
which we solve to get $x_{95\%} = 250 \times z_{95\%} + 1500 = \boxed{1910}$ points.

San Diego Drivers

The speed of drivers on highway 162 (speed limit 65 mph) is normally distributed with mean 72.5 mph and standard deviation 4.8 mph. On average, how many cars must a cop speed-check before finding the first person going 10 mph about the limit?

Creating the Bernoulli trial {10 mph above limit,not 10 mph above}, the problems wants the expected value of a Geometric random variable.

Let $X = N(72.5, 4.8)$ be the speeds of drivers
Let
$p = P(10 \text{ mph above limit}) = P(X \geq 75)$.



```
> pnorm(75, mean = 72.5, sd = 4.8, lower.tail = F)
 [1] 0.3012414
```

Let $Y = Geom(p)$, then

$$E(Y) = \frac{1}{p} = \frac{1}{0.3012} \simeq 3.32 \text{ cars.}$$