

# Lecture 7 - Continuous Distributions (Normal)

---

Sta102 / BME 102

February 8, 2016

Colin Rundel

## Types of Distributions

---

# Discrete Probability Distributions

A *discrete probability distribution* lists all possible outcomes and the probabilities with which they occur.

Rules for discrete probability distributions:

- The outcomes must be disjoint

$$P(X \cap Y) = 0 \text{ if } X \neq Y$$

- The probability of each outcome must be between 0 and 1

$$0 \leq P(X) \leq 1$$

- The sum of the probabilities of outcomes must total 1

$$\sum_{\text{all } x} P(X = x) = 1$$

# Continuous Probability Distributions

A *continuous probability distribution* differs from a discrete probability distribution in several ways:

- The probability that a continuous RV will equal to any specific value is always zero.

$$P(X = x) = 0 \quad \forall x$$

- Distribution is instead described by a *probability density function* -  $f(x)$ .

$$f(x) = \lim_{\epsilon \rightarrow 0} P(X \in \{x, x + \epsilon\})$$

- Probabilities can only be calculated for ranges of values, and are given by the area under the pdf curve.

$$P(a < X < b) = \int_a^b f(x) dx$$

## Rules for density functions

A proper *probability density function* has the following properties:

- The density is positive everywhere

$$f(x) \geq 0 \text{ for all } x \in (-\infty, \infty).$$

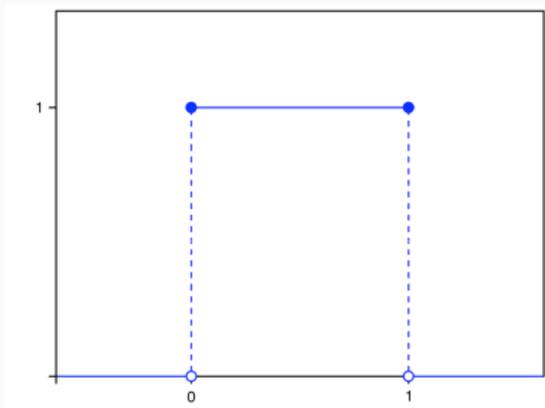
- The area under the density function (from  $-\infty$  to  $\infty$ ) must be equal to 1,

$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

## Example - Uniform Distribution

If a random variable  $X$  has constant pdf over a range  $(0, 1)$  then  $X$  has a standard uniform distribution,  $X \sim \text{Unif}(0, 1)$ .

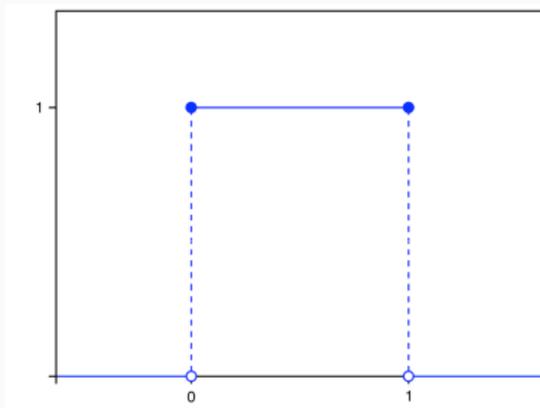
$$f(x) = \begin{cases} 1 & \text{for } x \in (0, 1) \\ 0 & \text{otherwise} \end{cases}$$



## Example - Uniform Distribution

If a random variable  $X$  has constant pdf over a range  $(0, 1)$  then  $X$  has a standard uniform distribution,  $X \sim \text{Unif}(0, 1)$ .

$$f(x) = \begin{cases} 1 & \text{for } x \in (0, 1) \\ 0 & \text{otherwise} \end{cases}$$



Properties:

$$E(X) = 1/2$$

$$\text{Var}(X) = 1/12$$

$$P(X = x) = 0$$

$$P(a < X < b) = b - a$$

for  $a, b \in (0, 1)$

## Example - Uniform(0,2) Distribution

What are the properties of a random variable  $Y$  if it has a constant pdf over a range  $(0, 1/2)$ ?

# Normal distribution

---

# Normal distribution

The normal distribution is a continuous distribution that is unimodal and symmetric with a distinctive bell shaped density function.

Many variables are nearly normal, but very few are exactly normal - we will see why next time.

Density given by

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right]$$

# Normal distribution

The normal distribution is a continuous distribution that is unimodal and symmetric with a distinctive bell shaped density function.

Many variables are nearly normal, but very few are exactly normal - we will see why next time.

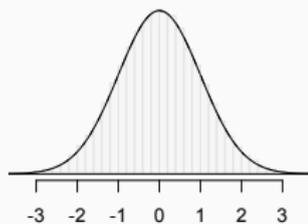
Density given by

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right]$$

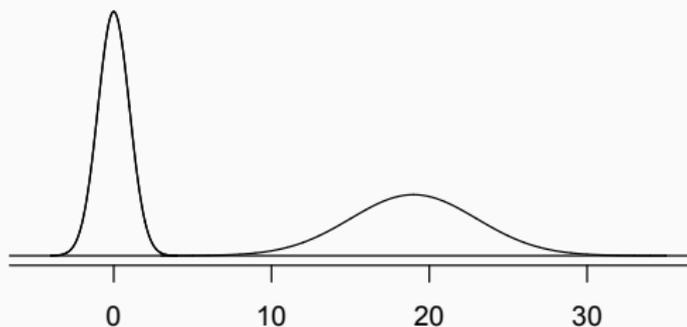
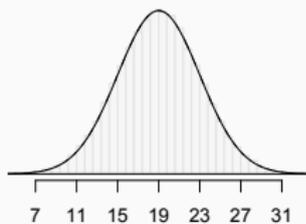
$$X \sim N(\mu, \sigma) \quad E(X) = \mu \quad \text{Var}(X) = \sigma^2 \quad \text{SD}(X) = \sigma$$

# Normal distribution parameters

$$N(\mu = 0, \sigma = 1)$$



$$N(\mu = 19, \sigma = 4)$$

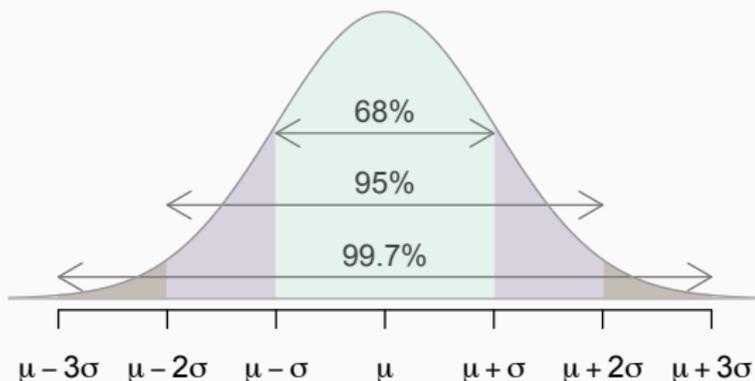


## 68-95-99.7 Rule

For nearly normally distributed data,

- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SD of the mean,
- about 99.7% falls within 3 SD of the mean.

It is possible for observations to fall 4, 5, or 1000 standard deviations away from the mean, but these occurrences are very rare.



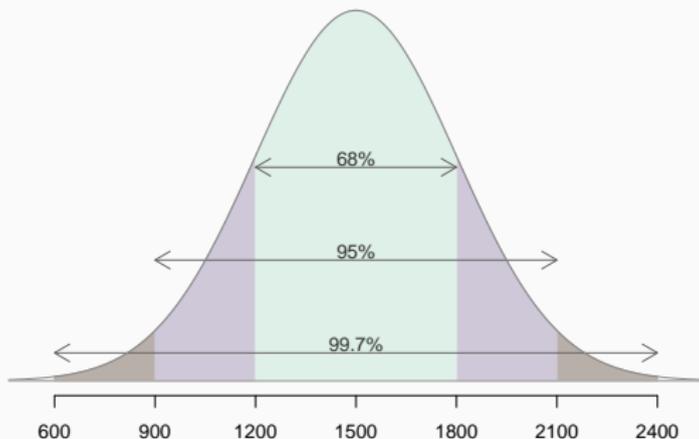
## Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

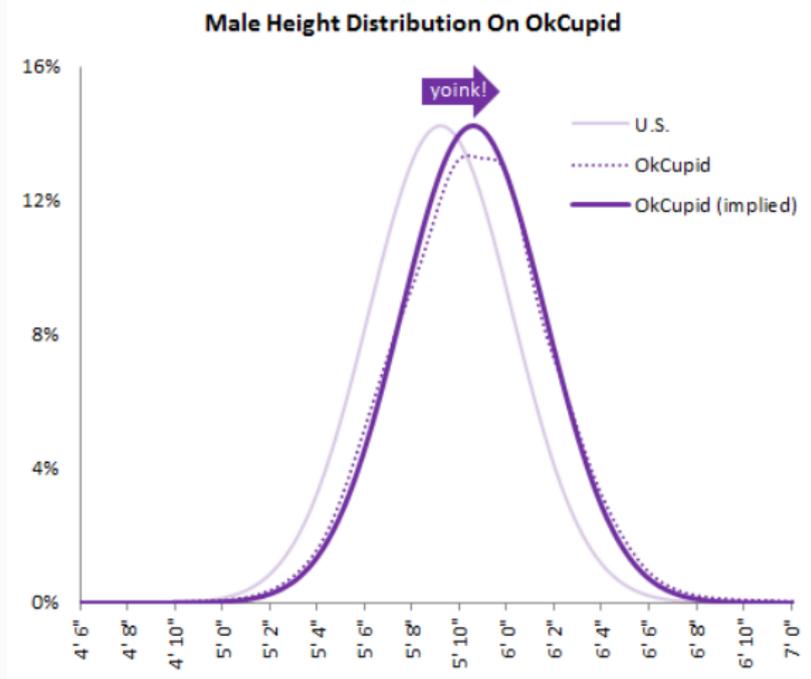
## Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

- ~68% of students score between 1200 and 1800 on the SAT.
- ~95% of students score between 900 and 2100 on the SAT.
- ~99.7% of students score between 600 and 2400 on the SAT.



# Heights of males



<http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/>

## OkCupid's Take

“The male heights on OkCupid very nearly follow the expected normal distribution – except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches.”

## OkCupid's Take

“The male heights on OkCupid very nearly follow the expected normal distribution – except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches.”

“You can also see a more subtle vanity at work: starting at roughly 5' 8", the top of the dotted curve tilts even further rightward. This means that guys as they get closer to six feet round up a bit more than usual, stretching for that coveted psychological benchmark.”

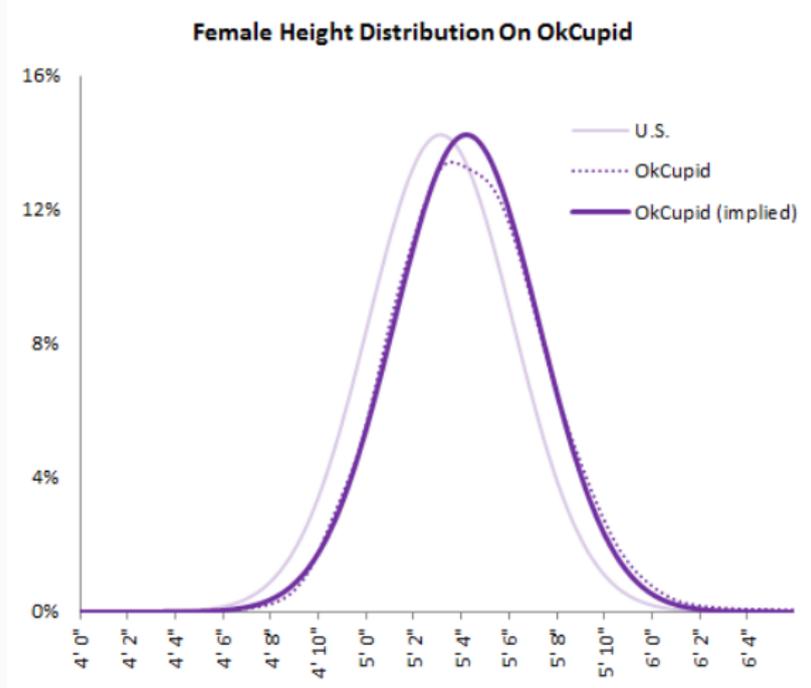
## OkCupid's Take

“The male heights on OkCupid very nearly follow the expected normal distribution – except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches.”

“You can also see a more subtle vanity at work: starting at roughly 5' 8", the top of the dotted curve tilts even further rightward. This means that guys as they get closer to six feet round up a bit more than usual, stretching for that coveted psychological benchmark.”

“When we looked into the data for women, we were surprised to see height exaggeration was just as widespread, though without the lurch towards a benchmark height.”

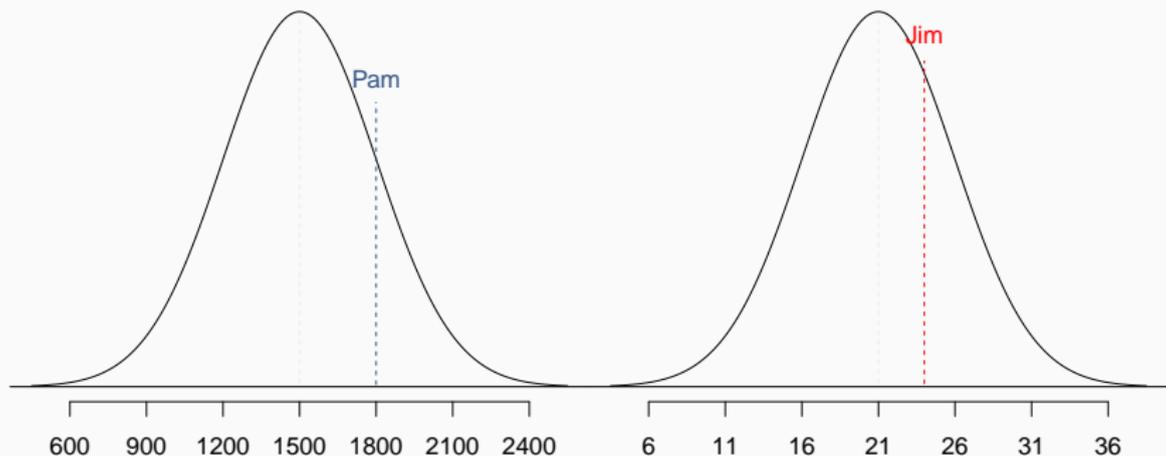
# Heights of females



<http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/>

## Comparing SAT and ACT

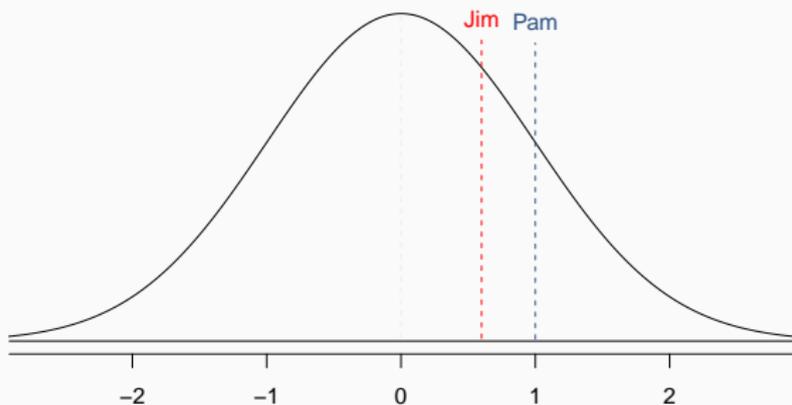
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?



# Standardizing

Since we cannot just compare these two raw scores, we instead compare how many standard deviations above or below the mean each observation is.

- Pam's score is  $\frac{1800-1500}{300} = 1$  standard deviation above the mean.
- Jim's score is  $\frac{24-21}{5} = 0.6$  standard deviations above the mean.



## Standardizing with Z scores (cont.)

*Z* / *standardized* / *normalized* scores

- A measure of the number of standard deviations the data falls above or below the mean.

$$Z = \frac{\textit{observation} - \textit{mean}}{SD}$$

- We can calculate Z scores for distributions of any shape, but with normal distributions we use Z scores to calculate probabilities.
- Observations that are more than 2 SD away from the mean ( $|Z| > 2$ ) are typically considered unusual.

## Z distribution

Another reason we use Z scores is if the distribution of X is nearly normal then the Z scores of X will have a Z distribution (unit normal).

- The Z distribution is a special case of the normal distribution where  $\mu = 0$  and  $\sigma = 1$ .
- Linear transformations of normally distributed random variable are also normally distributed. Hence, if

$$Z = \frac{X - \mu}{\sigma}, \text{ where } X \sim N(\mu, \sigma)$$

## Z distribution

Another reason we use Z scores is if the distribution of X is nearly normal then the Z scores of X will have a Z distribution (unit normal).

- The Z distribution is a special case of the normal distribution where  $\mu = 0$  and  $\sigma = 1$ .
- Linear transformations of normally distributed random variable are also normally distributed. Hence, if

$$Z = \frac{X - \mu}{\sigma}, \text{ where } X \sim N(\mu, \sigma)$$

## Z distribution

Another reason we use Z scores is if the distribution of X is nearly normal then the Z scores of X will have a Z distribution (unit normal).

- The Z distribution is a special case of the normal distribution where  $\mu = 0$  and  $\sigma = 1$ .
- Linear transformations of normally distributed random variable are also normally distributed. Hence, if

$$Z = \frac{X - \mu}{\sigma}, \text{ where } X \sim N(\mu, \sigma)$$

$$E(Z) = E\left(\frac{X - \mu}{\sigma}\right) = E(X/\sigma) - \mu/\sigma = 0$$

## Z distribution

Another reason we use Z scores is if the distribution of  $X$  is nearly normal then the Z scores of  $X$  will have a Z distribution (unit normal).

- The Z distribution is a special case of the normal distribution where  $\mu = 0$  and  $\sigma = 1$ .
- Linear transformations of normally distributed random variable are also normally distributed. Hence, if

$$Z = \frac{X - \mu}{\sigma}, \text{ where } X \sim N(\mu, \sigma)$$

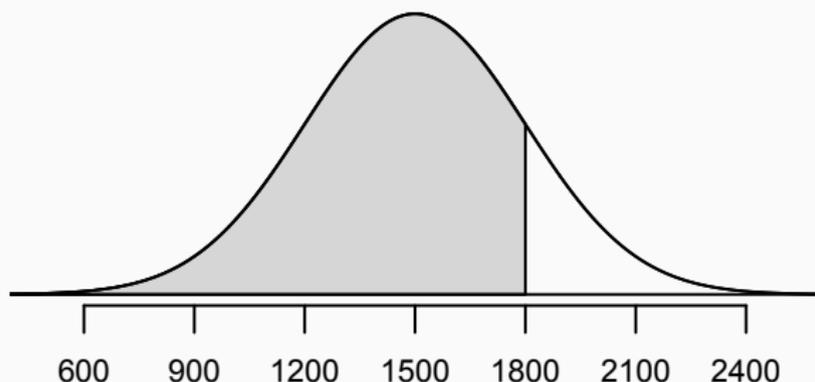
$$E(Z) = E\left(\frac{X - \mu}{\sigma}\right) = E(X/\sigma) - \mu/\sigma = 0$$

$$\text{Var}(Z) = \text{Var}\left(\frac{X - \mu}{\sigma}\right) = \text{Var}(X/\sigma) = \frac{1}{\sigma^2}\text{Var}(X) = 1$$

# Percentiles

- A *Percentile* is the percentage of observations that fall below a given data point.
- Graphically, the percentile is the area below the pdf that is to the left of the given value.

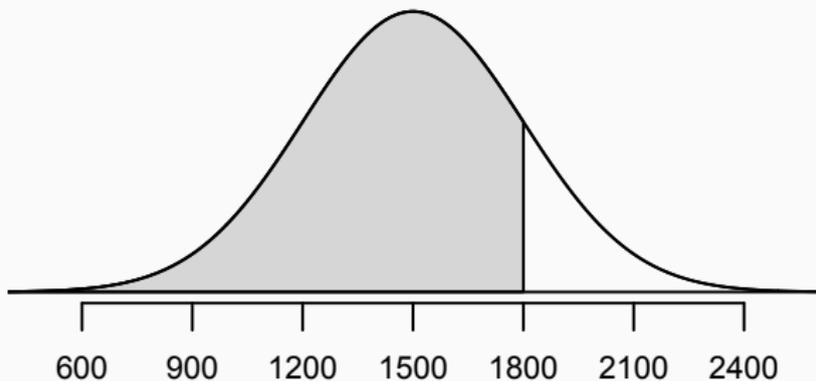
$$P(\text{SAT} < 1800)$$



## Example - SAT

Approximately what percent of students score below 1800 on the SAT?

$$\mu = 1500, \quad \sigma = 300$$



## Calculating Probabilities

We can use the empirical rule to solve some problems, but our Z Score is not -3,-2,-1,2, or 3 we have a problem.

From the definition of a continuous random variable we could find an arbitrary probability using calculus,

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x - \mu)^2/\sigma^2\right] dx.$$

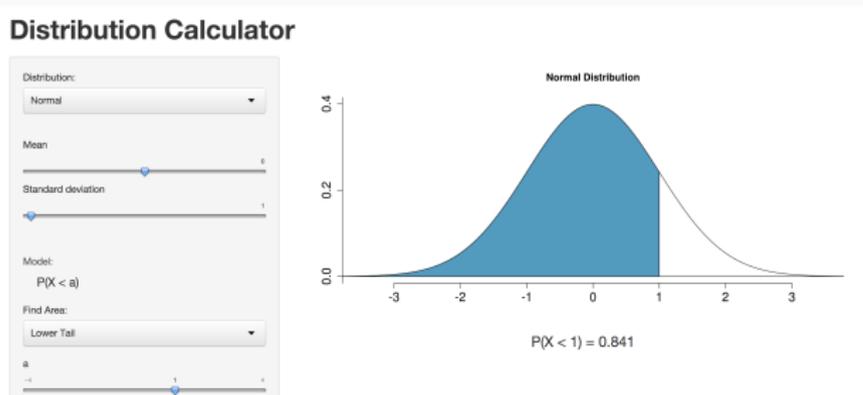
In the case of the SAT example we have to evaluate,

$$P(X \leq 1800) = \int_{-\infty}^{1800} \frac{1}{300\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - 1500}{300}\right)^2\right] dx$$

# Calculating Probabilities

There are additional ways of finding these probabilities/areas under the curve,

- R:  
`pnorm(1800, mean = 1500, sd = 300)`
- Applet:



[https://gallery.shinyapps.io/dist\\_calc/](https://gallery.shinyapps.io/dist_calc/)

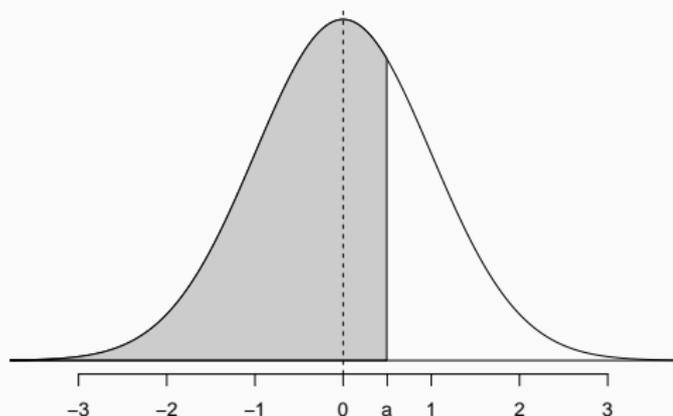
# Calculating Probabilities - Z Table

Z	Second decimal place of 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

## Calculating left tail probabilities

The area under the unit normal curve from  $-\infty$  to  $a$  is given by

$$P(Z \leq a) = \Phi(a)$$

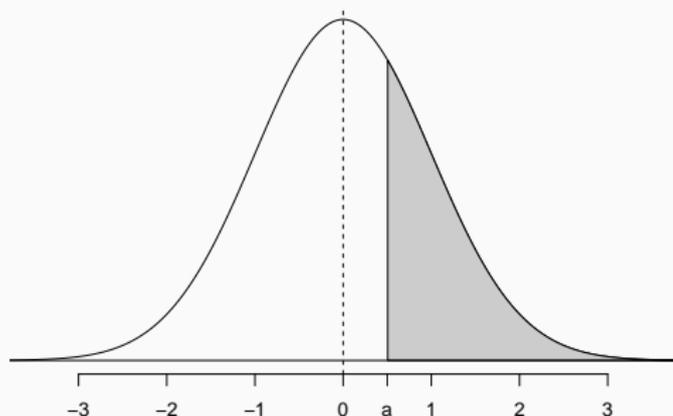


\*These left tail probabilities are sometimes called percentiles

## Calculating right tail probabilities

The area under the unit normal curve from  $a$  to  $\infty$  is given by

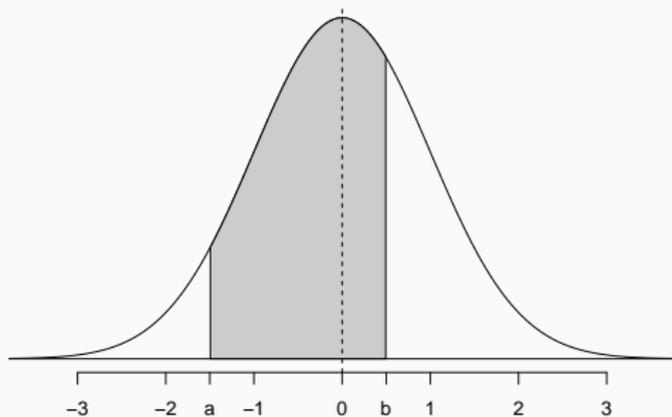
$$P(Z \geq a) = 1 - \Phi(a)$$



## Calculating middle probabilities

The area under the unit normal curve from  $a$  to  $b$  where  $a \leq b$  is given by

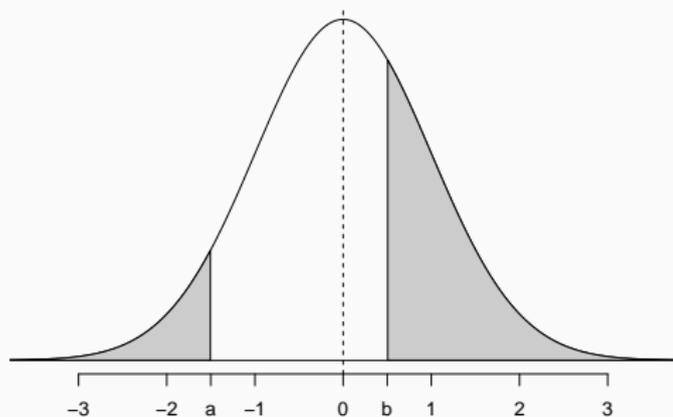
$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$$



## Calculating two tail probabilities

The area under the unit normal curve outside of  $a$  to  $b$  where  $a \leq b$  is given by

$$P(a \geq Z \text{ or } Z \geq b) = \Phi(a) + (1 - \Phi(b)) = 1 - (\Phi(b) - \Phi(a))$$



How would you calculate the following probability?

$$P(Z < -1)$$

How would you calculate the following probability?

$$P(Z > 2.22)$$

How would you calculate the following probability?

$$P(-1.53 \leq Z \leq 2.75)$$

How would you calculate the following probability?

$$P(Z \leq 0.75 \text{ or } Z \geq 1.43)$$

## Probabilities for non-Unit Normal Distributions

Everything we just discussed on the previous 4 slides applies only to the unit normal distribution, but this doesn't come up very often in problems.

Let  $X$  be a normally distributed random variable with mean  $\mu$  and variance  $\sigma^2$  then we define the random variable  $Z$  such that

$$Z = \left( \frac{X - \mu}{\sigma} \right) \sim N(0, 1)$$

## Probabilities for non-Unit Normal Distributions

Everything we just discussed on the previous 4 slides applies only to the unit normal distribution, but this doesn't come up very often in problems.

Let  $X$  be a normally distributed random variable with mean  $\mu$  and variance  $\sigma^2$  then we define the random variable  $Z$  such that

$$Z = \left( \frac{X - \mu}{\sigma} \right) \sim N(0, 1)$$

$$P(a \leq X \leq b) = P\left( \frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma} \right) = \Phi\left( \frac{b - \mu}{\sigma} \right) - \Phi\left( \frac{a - \mu}{\sigma} \right)$$

## Examples

---

## Example - Dosage

At a pharmaceutical factory the amount of the active ingredient which is added to each pill is supposed to be 36 mg. The amount of the active ingredient added follows a nearly normal distribution with a standard deviation of 0.11 mg. Once every 30 minutes a pill is selected from the production line, and its composition is measured precisely. If the amount of the active ingredient in the pill is below 35.8 mg or above 36.2 mg, then that production run of pills fails the quality control inspection. What percent of production runs will fail for having too little active ingredient (less than 35.8 mg)?

# Finding the exact probability

Second decimal place of Z										Z
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	
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
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	-1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	-1.5

## Example - Body Temperature

Body temperatures of healthy humans are distributed nearly normally with mean  $98.2^{\circ}\text{F}$  and standard deviation  $0.73^{\circ}\text{F}$ . What is the cutoff for the lowest 3% of human body temperatures?  
(Mackowiak, Wasserman, and Levine 1992)

## Example - Body Temperature pt. 2

What is the cutoff for the highest 10% of human body temperatures?