The normal distribution

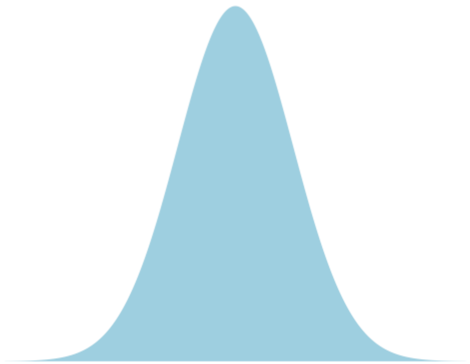
yaniv

March 3

The normal distribution

Much of statistics is based off of the normal distribution.

y tho??



You may have heard much about this before…
What do you know about the normal distribution?

Worksheet / check in

Sums are Normally Distributed

Most quantitative variables are sums of a bunch of things.

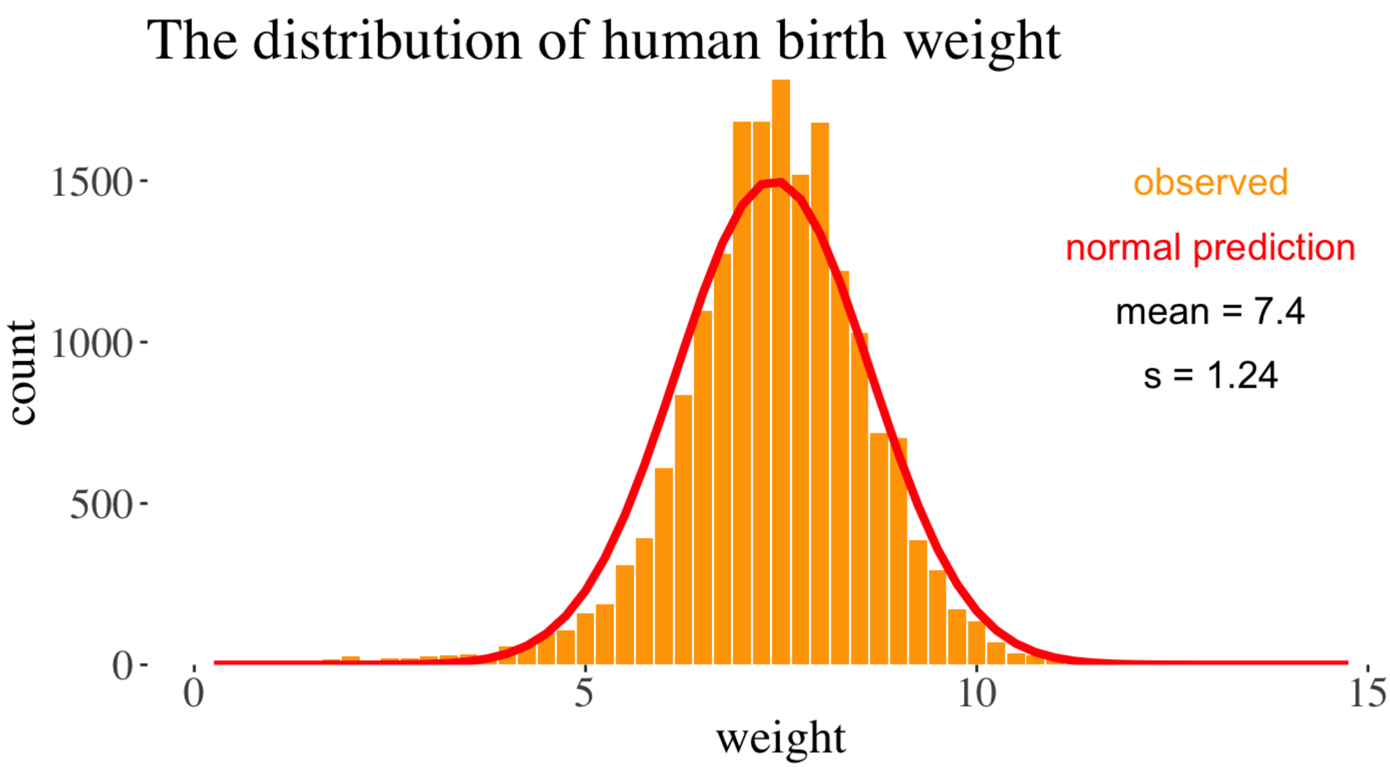
For example:

- Human height is realized as the addition of lot of genetic effects and a lot of environmental factors.
- The distance a seed moves is the sum of a lot of wind currents.

Consequently, Many Biological Variables Are (Approximately) Normally Distributed

Human Birth Weight

Birth weight is (roughly) normally distributed

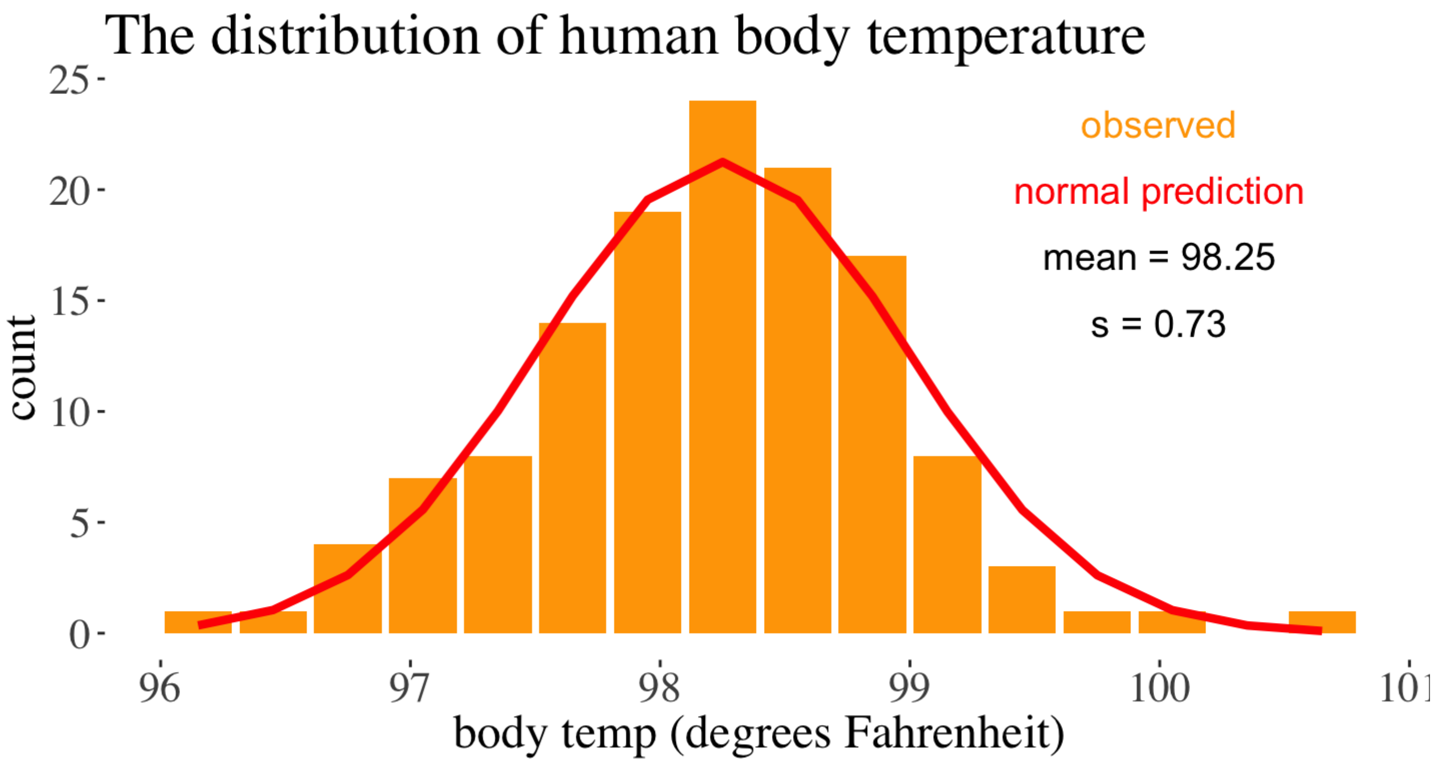


Data from Pethybridge, Ashford, and Fryer 1974

6/45

Human Body Temp

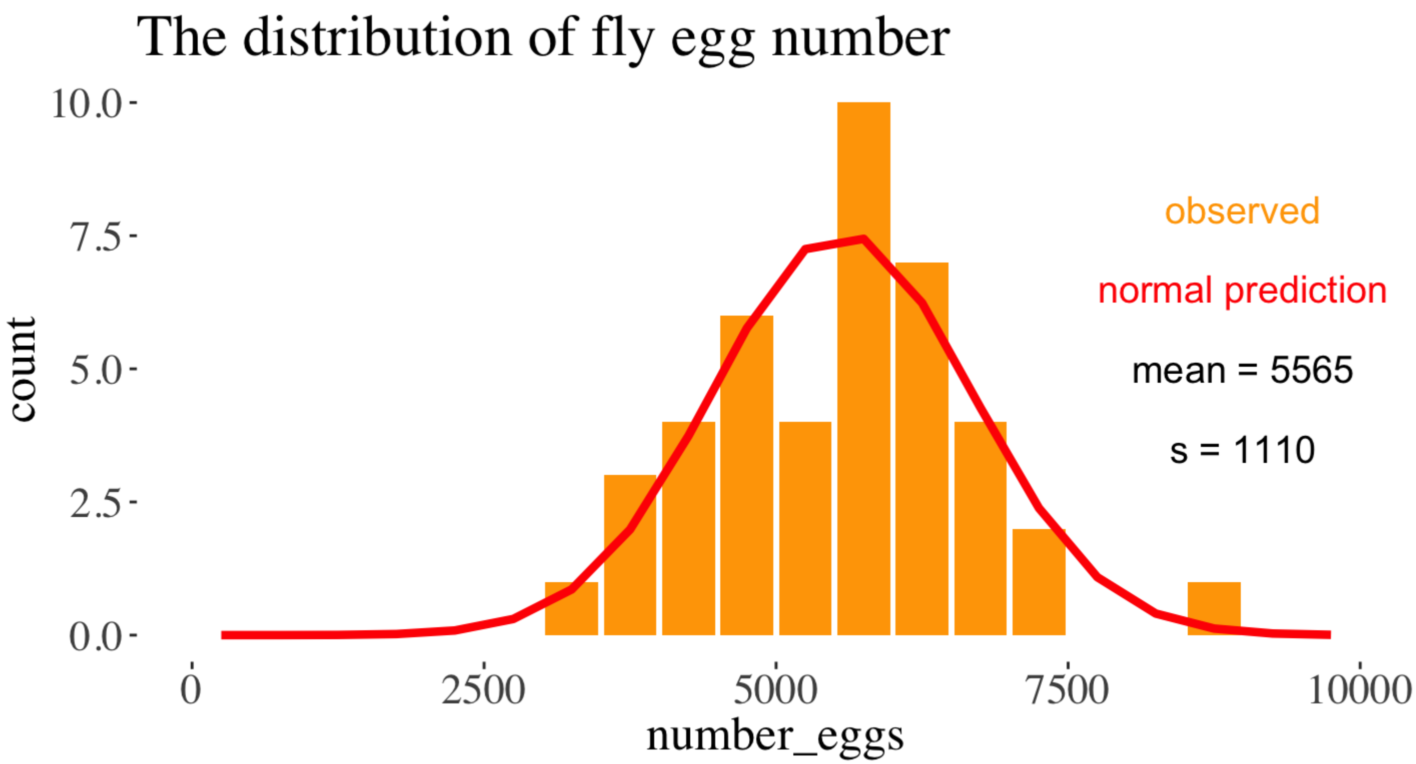
Body temp is (roughly) normally distributed



Data from Shoemaker 1996

Egg Number

egg number (roughly) normally distributed

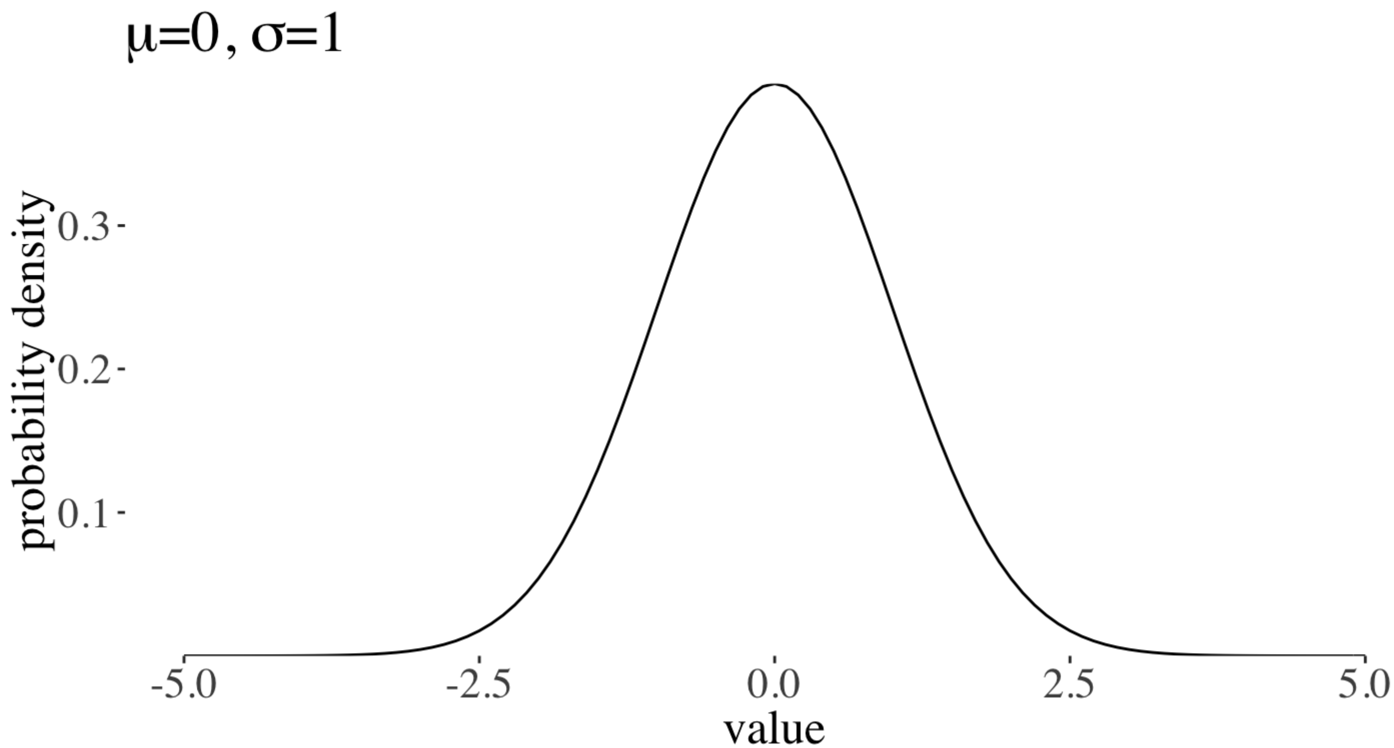


Data from Paaby, Bergland, Behrman and Schmidt 2014, data link

8/45

The Normal Distribution: Definitions and Properties

Probability Density of A Normal Dist.

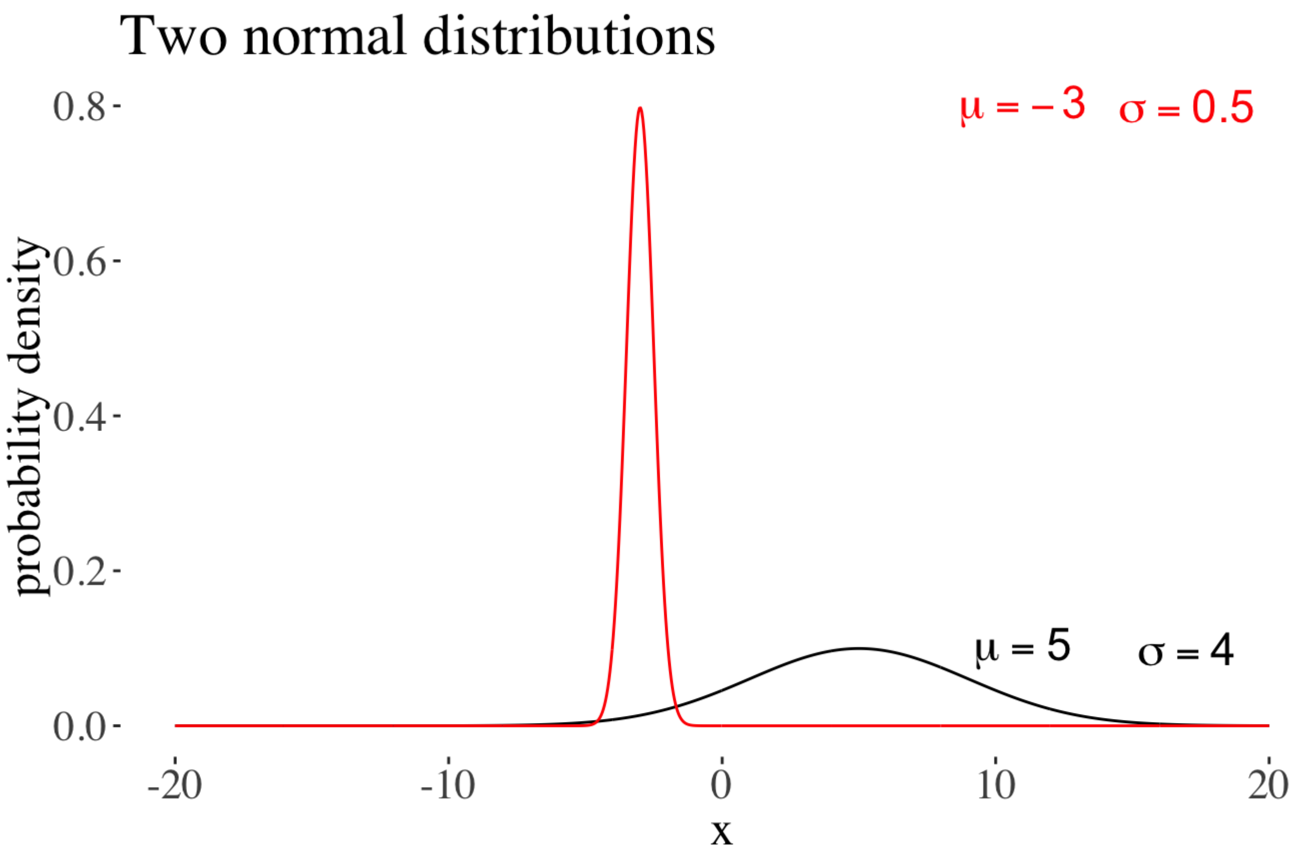


$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

10/45

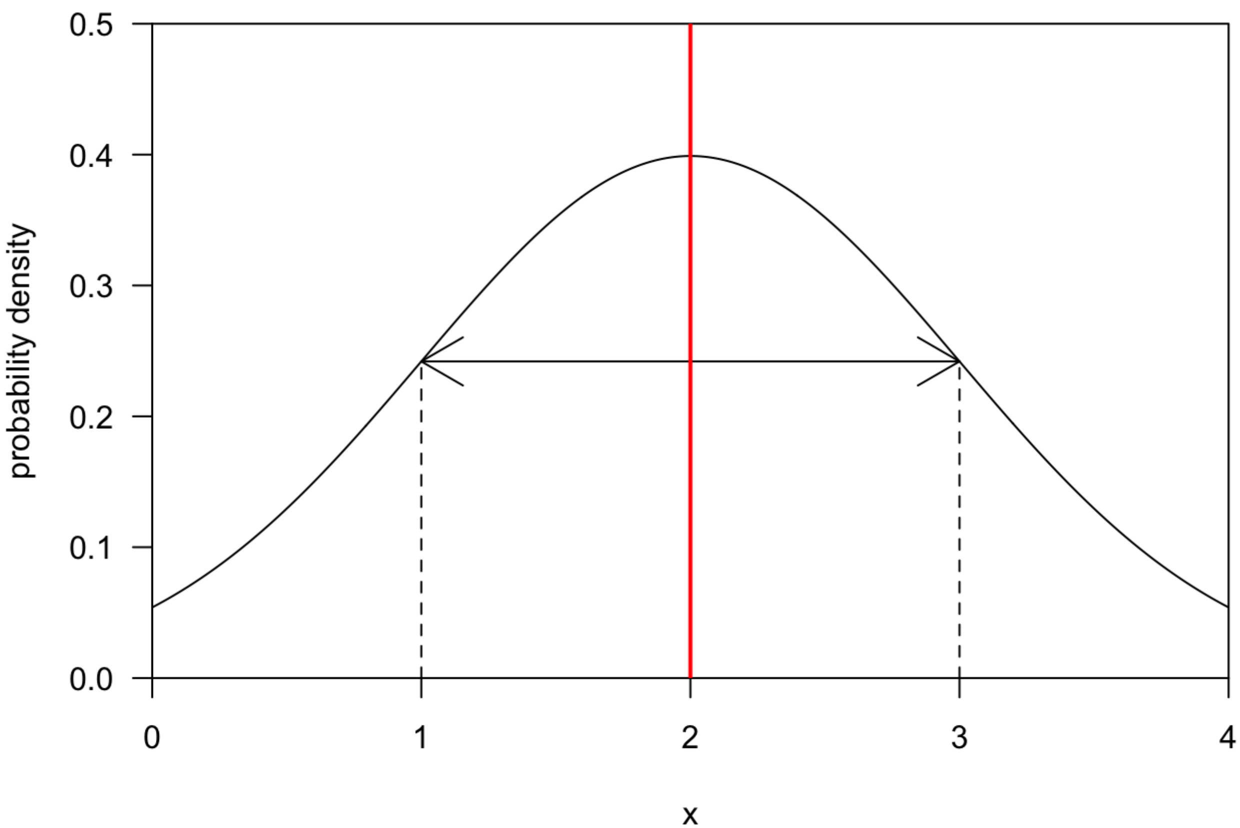
A Normal Dist. Has Two Params: μ & σ

$\mathcal{N}(\mu, \sigma)$ – These parameters fully specify a normal distribution



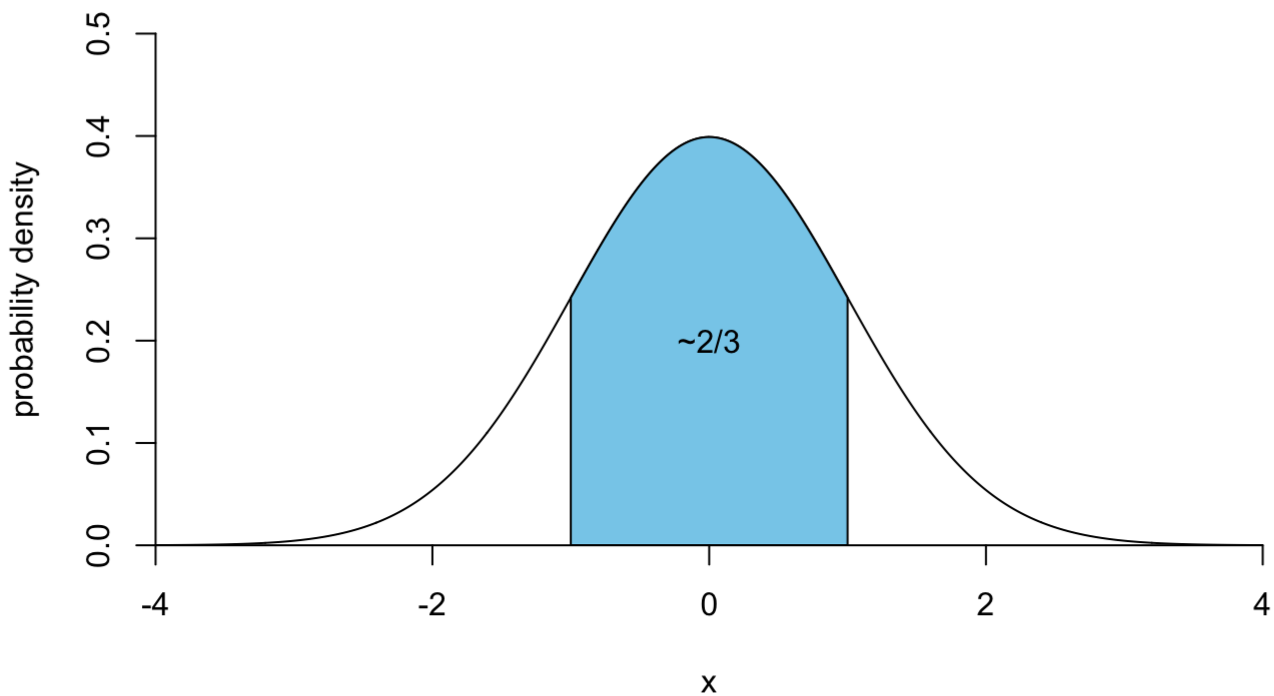
A Normal Distribution is Symmetric

A normal distribution is symmetric & centered around its mean.



$\approx 66\%$ of a Normal is Within $\mu \pm \sigma$

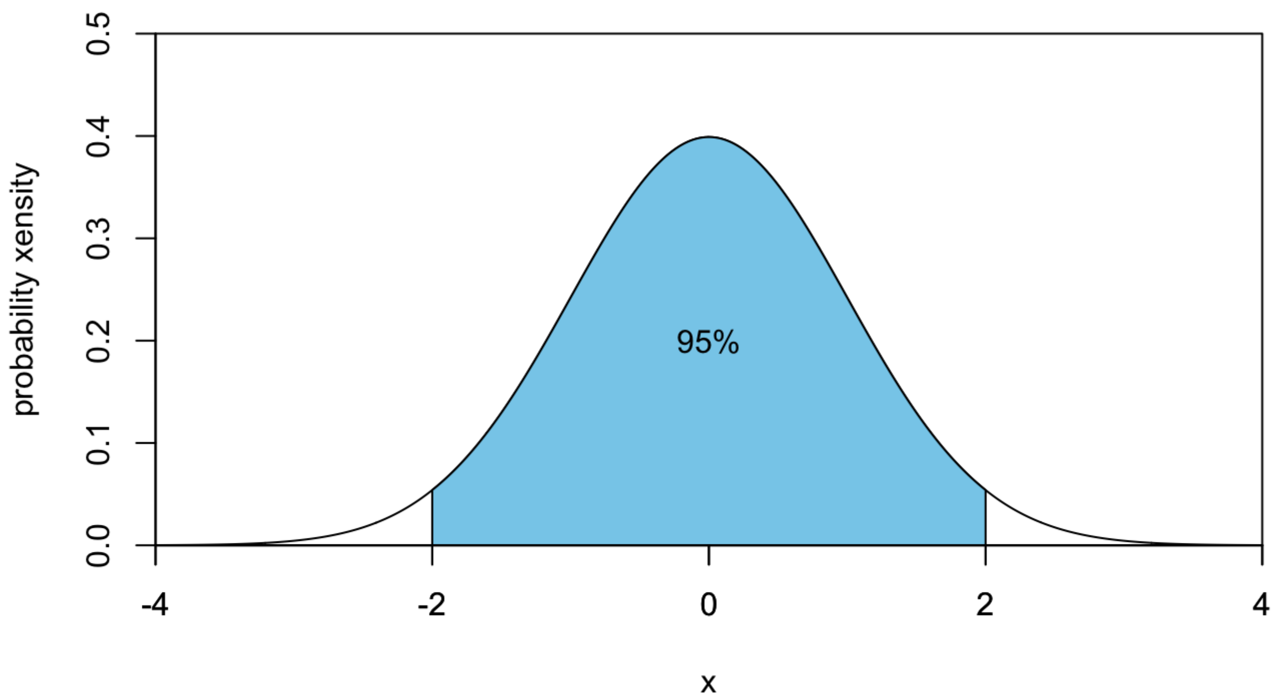
$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \approx 2/3$$



13/45

$\approx 95\%$ of a Normal is Within $\mu \pm 2\sigma$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \approx 0.95$$



R Exercises: Pick μ & σ

Simulate a normal with `rnorm()`, and convince yourself that

- A normal distribution is symmetric around its mean.
- About 66% of draws are between $\mu - \sigma$ and $\mu + \sigma$.
- About 95% of draws are between $\mu - 2\sigma$ and $\mu + 2\sigma$.

Show that `dnorm()` returns $(1/\sqrt{2\pi\sigma^2})e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Use `pnorm()` to find the proportion of your normal $< \mu - 2\sigma$

Use `pnorm()` to find the proportion of your normal $> \mu + 2\sigma$

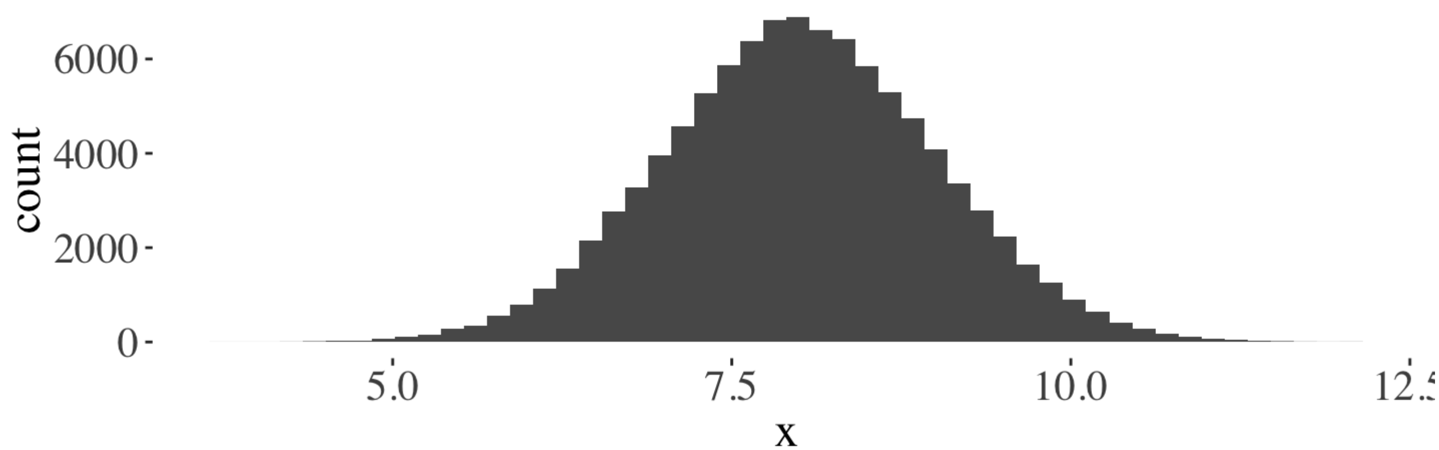
Use `qnorm()` to find the cutoff for the lower 2.5% tail

Explain the difference between & usage of `q` `p` `r` and `d` `norm()`

R Exercises: 1. Symmetry

A normal distribution is symmetric around its mean.

```
mu   <- 8; sigma <- 1; X <- 7
sim  <- tibble(x = rnorm(n = 100000, mean = mu, sd = sigma))
```




```
sim %>% summarise(mean(x>mu)) %>% pull() # How much is greater than mu
```

```
## [1] 0.49934
```

R Exercises: 2. Within one or two σs by simulation

```
sim %>% summarise(
  within.one  = mean(x > mu - 1 * sigma &  x < mu + 1 * sigma),
  within.two = mean(x > mu - 2 * sigma &  x < mu + 2 * sigma))
```

```
## # A tibble: 1 x 2
##   within.one within.two
##        <dbl>      <dbl>
## 1      0.680      0.955
```

R Exercises: 3. **`dnorm()`** returns

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

```
(1/sqrt(2 * pi  * sigma^2)) * exp(-(X-mu)^2/(2*sigma^2)) # math
```

```
## [1] 0.2419707
```

```
dnorm(x = X, mean = mu, sd = sigma)                        # dnorm()
```

```
## [1] 0.2419707
```

R Exercises: 4. Within one or two σs by pnorm()

Use `pnorm()` to find the proportion of your normal $< \mu - \sigma$

```
pnorm(q = mu -  sigma, mean = mu, sd = sigma) # A bit less than 0.33/2
```

```
## [1] 0.1586553
```

Use `pnorm()` to find the proportion of your normal $> \mu - 2\sigma$

```
pnorm(q = mu - 2 * sigma, mean = mu, sd = sigma) # A bit less than 0.025
```

```
## [1] 0.02275013
```

R Exercises: 5. Critical value qnorm()

Use `qnorm()` to find the cutoff for the lower 2.5% tail

```
critical.val <- qnorm(p = .025,mean = mu, sd = sigma);  critical.val
```

```
## [1] 6.040036
```

```
(critical.val - mu) / sigma # In units of sigma from mean
```

```
## [1] -1.959964
```

R Exercises: 6. $\sigma_{\bar{x}} = \sigma/\sqrt(n)$?

Demonstrate that the standard deviation of the sampling distribution of your normal is roughly $\sigma/\sqrt{n}$ with `rnorm()`

```
sample.sizes <- rep(c(5,10,20,50,100,500), each = 1000)
tibble(x = rnorm(n = sum(sample.sizes), mean = mu, sd = sigma),
       trial = rep(seq_along(sample.sizes), times = sample.sizes)) %>%
  group_by(trial) %>%       summarise(estimate = mean(x), n = n()) %>%
  group_by(n)     %>%       summarise(sd(estimate))                %>%
  mutate(prediction = sigma / sqrt(n))
```

n	sd(estimate)	prediction
5	0.44631	0.44721
10	0.31117	0.31623
20	0.22566	0.22361
50	0.14419	0.14142
100	0.09663	0.10000
500	0.04543	0.04472

Our sims and math match!

21/45

Central limit theorem

Central limit theorem

The sum or mean of a large number of measurements randomly sampled from **ANY** population is approximately normally distributed.

Button Pushing Example [1/3]

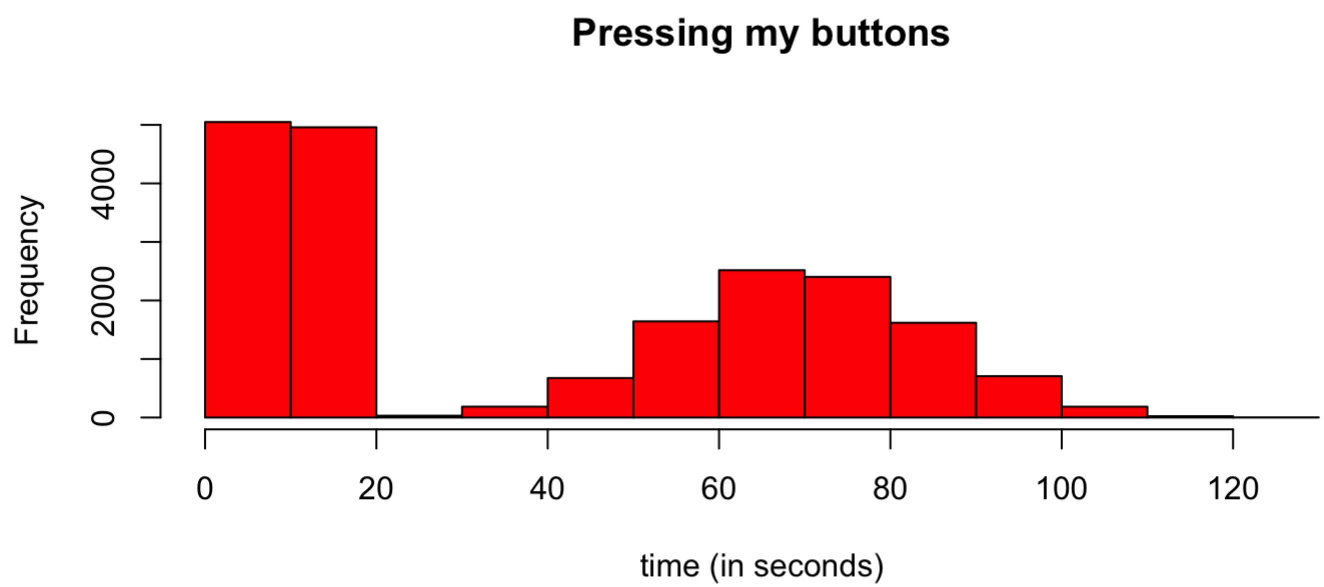
Imagine that 20000 people are asked to press a button.

Button Pushing Example [2/3]

Imagine that 20000 people are asked to press a button.

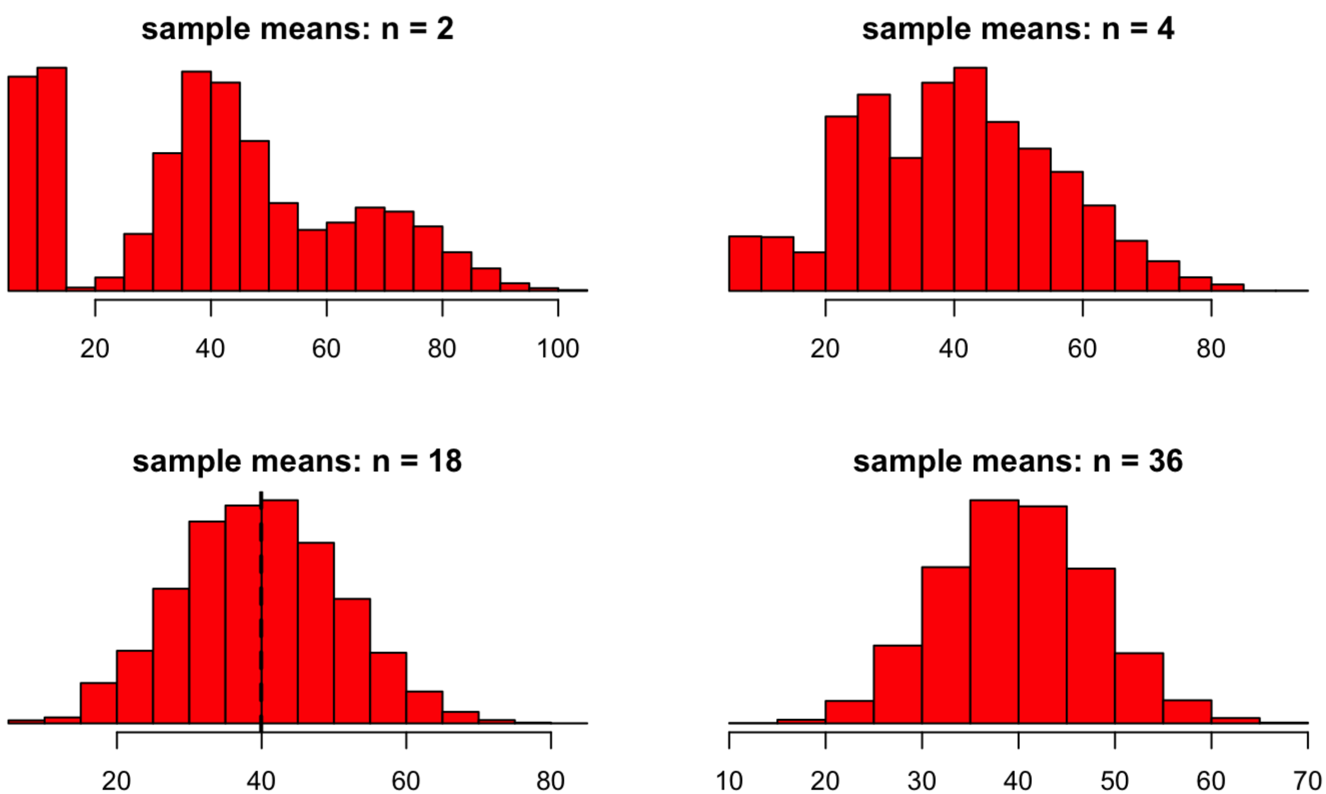
- Half are anxious & do this quickly $\mathcal{N}(10, 1)$
- Half are not & do this slowly $\mathcal{N}(70, 15)$

This is not normally distributed



Button Pushing Example [3/3]

The sampling distribution becomes normal as n gets large.



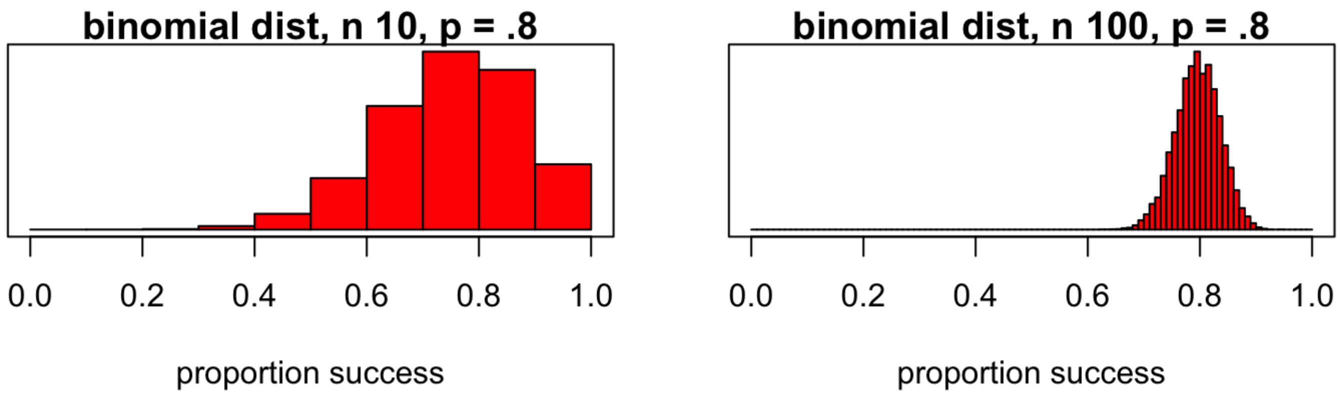
Normal Approximation to the Binomial

The central limit theorem provides a simple way to estiamte binomial probabilities.

When number of trials (n) is large and probability of success (p) is not too close to 0 or 1

We can approximate the binomial by a normal distribution with$\mu = np$ and $\sigma = \sqrt{np(1-p)}$

The binomial dist. approaches a normal dist. as n gets larger



This is an example of the Central Limit Theorem in action

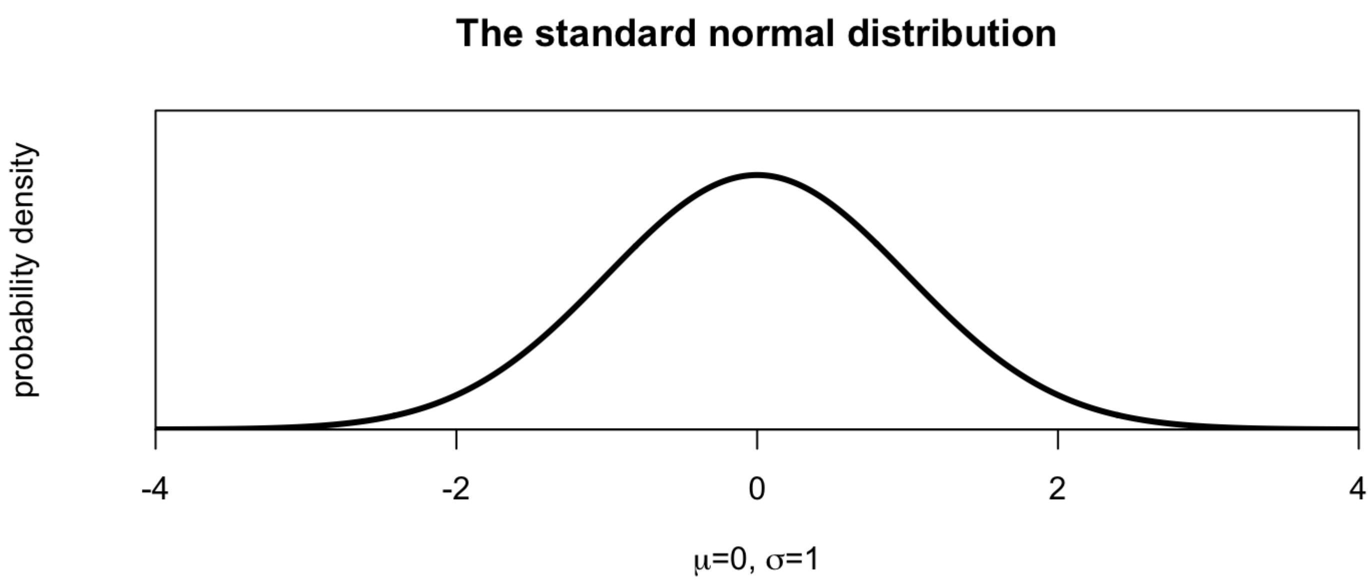
Normal approximation to the binomial distribution

Pr[number of successes $\geq X$]$= Pr[Z > \frac{X-np}{\sqrt{np(1-p)}}]$

The Standard Normal Distribution

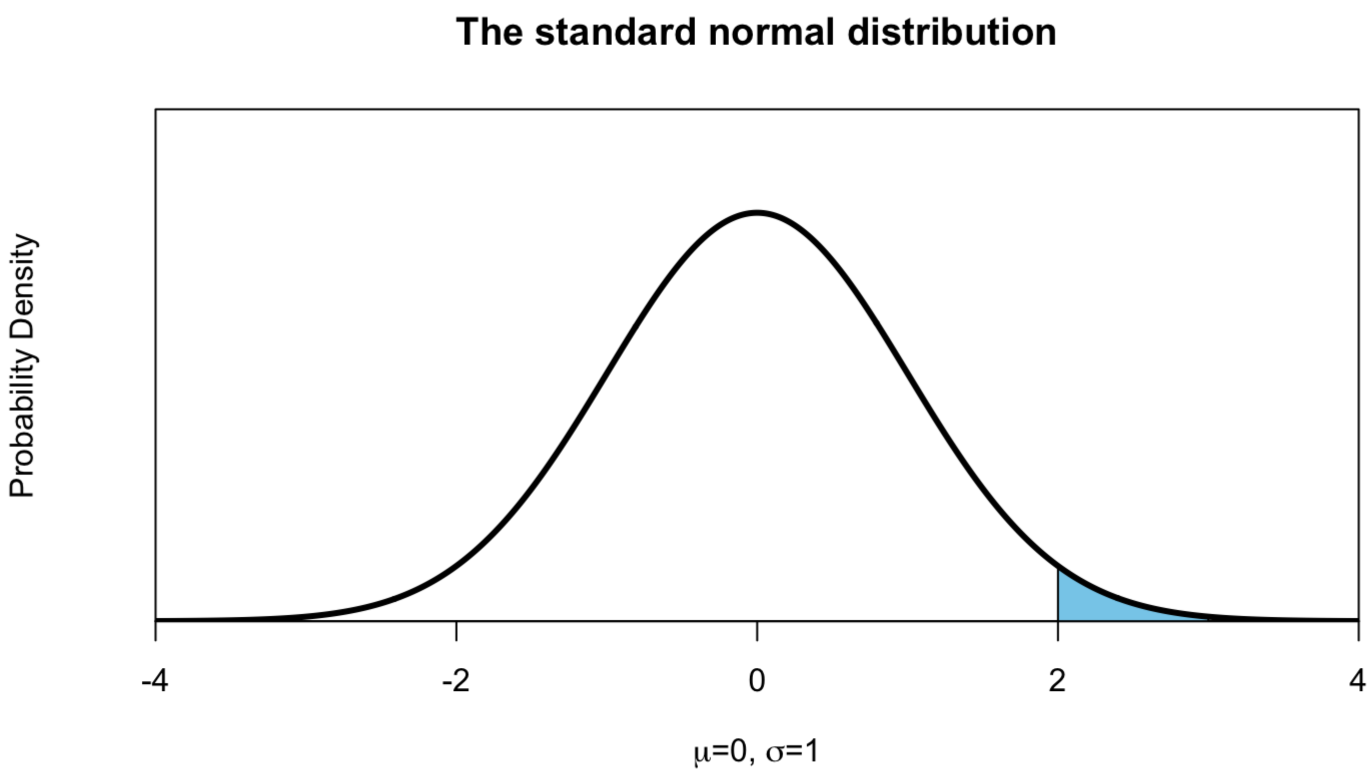
One Normal Distribution To Rule Them

- Of the infinite normal distributinos, the ,
$\mathcal{N}(\mu = 0, \sigma = 1)$ is particularly useful.
- We can easily tansform any normal distribution into the standard normal distribution.



The Standard Normal Table [1/2]

- Gives the probability of getting a random draw from a standard normal distribution greater than a given value



The Standard Normal Table [2/2]

This is available in the back of the text.

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0.5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0.7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.215
0.8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0.9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161

Standard normal is symmetric, so…

- $Pr[Z > x] = Pr[Z < -x]$
- $Pr[Z < x] = 1 - Pr[Z > x]$

34/45

Using the Standard Normal Table

Finding the critical Z value for a two-sided test with $\alpha = 0.05$

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0.5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0.7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.215
0.8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0.9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161

Other Normal Distributions

What About Other Normals?

- Normal distributions can have distint values of μ and σ but must have the same shape.
- Any normal distribution can be converted to a standard normal distribution, by a

$$Z = \frac{Y - \mu}{\sigma}$$

Z is called a "standard normal deviate"

Z = Distance Between Y & μ (in σ units)

$$Z = \frac{Y - \mu}{\sigma}$$

The probability of getting a value greater than Y is the same as the probability of getting a value greater than Z from a standard normal distribution.

Solve This Example: British Spies

MI5 says males spies mut be < 180.3 cm tall.

Mean height of British men is $\mathcal{N}(177.0 \text{ cm}, 7.1 \text{ cm})$

What proportion of British men are excluded from a career as a spy by this height criteria?

Bond heights

39/45

British Spies Solution

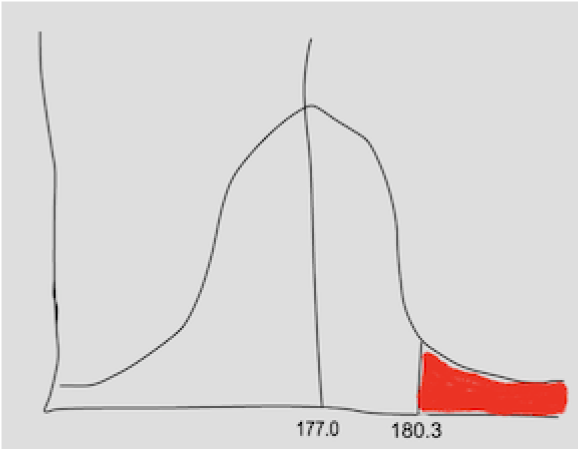
```
mu <- 177; X <- 180.3; sigma <- 7.1; Z  <- (X - mu) / sigma
```

With these data directly, the proportion of men too tall to be spies is

`pnorm(q = 177, mean = 180.3, sd = 7.1, lower.tail = FALSE)` = 0.321

With a Z-transform of these data, the proportion of men too tall to be spies is

`pnorm(q =`0.465`, mean = 0, sd = 1, lower.tail = FALSE)` = 0.321



Ths Sampling Distribution of Samples from a Normal Distribution

Sample means are normally distributed

- (If the variable itself is normally distributed)
- The mean of the sample means is μ
- The standard deviation of the sample means is $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$

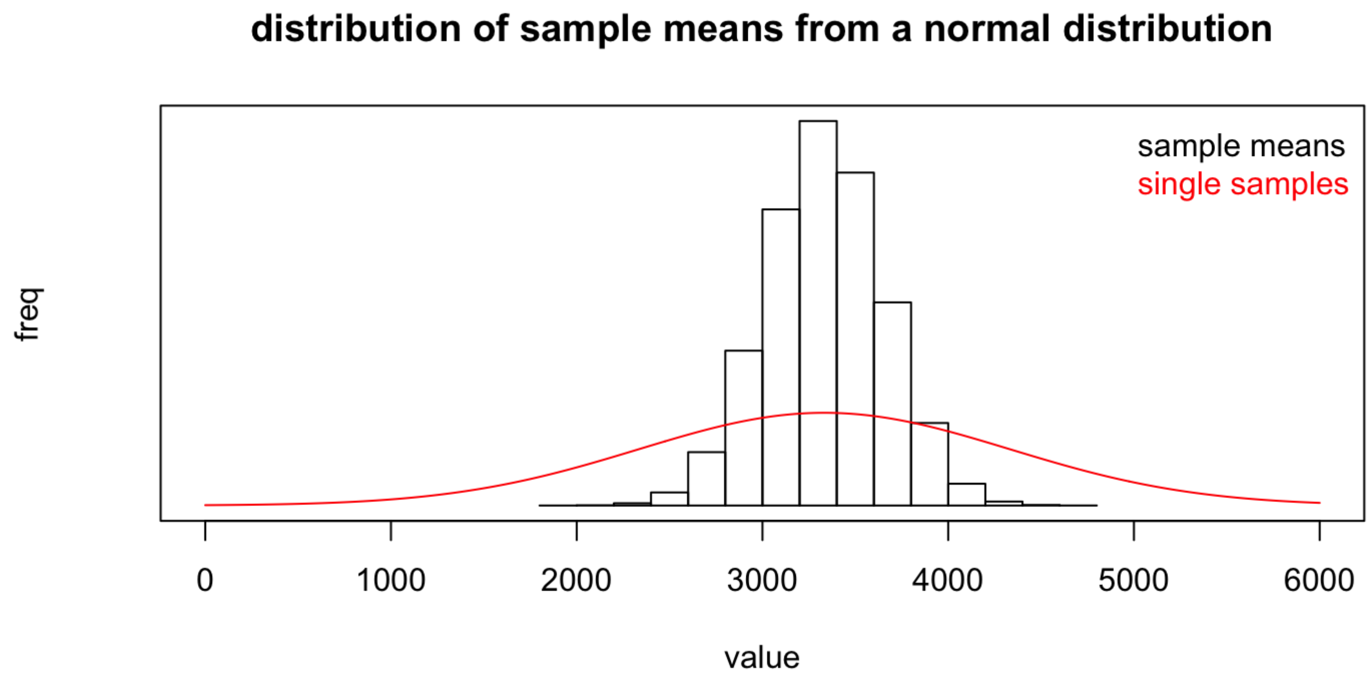
Standard error

The standard error of an estimate of a mean is the standard deviation of the distribution of sample means.

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

We can approximate this by $SE_{\overline{Y}} = \frac{s}{\sqrt{n}}$

Distribution of Sample Means (n = 10)



Law of Large Numbers

Larger samples make for tighter distributions & smaller standard errors