Chapter 4

Foundations for inference

Statistical inference

We have encountered several distributions that depend on a parameter:

How do we find these parameters? This is the goal of statistical inference but not only.
Consider the problem of the insurance company (1,000 employees, probability of injury 0.02). Where do these parameters come from: past observations

Year	2006	2007	2008	2009	2010	Average
# of injuries	22	3	16	36	23	**20**

Statistical inference

We have encountered several distributions that depend on a parameter:

$$Bin(n, p)$$

How do we find these parameters? This is the goal of statistical inference but not only.
Consider the problem of the insurance company (1,000 employees, probability of injury 0.02). Where do these parameters come from: past observations

Year	2006	2007	2008	2009	2010	Average
# of injuries	22	3	16	36	23	**20**

Statistical inference

We have encountered several distributions that depend on a parameter:

$$Bin(n, p) \qquad\qquad N(\mu, \sigma^2)$$

How do we find these parameters? This is the goal of statistical inference but not only.
Consider the problem of the insurance company (1,000 employees, probability of injury 0.02). Where do these parameters come from: past observations

Year	2006	2007	2008	2009	2010	Average
# of injuries	22	3	16	36	23	**20**

Statistical inference

We have encountered several distributions that depend on a parameter:

$$Bin(n, p) \qquad N(\mu, \sigma^2) \qquad Poiss(\lambda)$$

How do we find these parameters? This is the goal of statistical inference but not only.

Consider the problem of the insurance company (1,000 employees, probability of injury 0.02). Where do these parameters come from: past observations

Year	2006	2007	2008	2009	2010	Average
# of injuries	22	3	16	36	23	**20**

Statistical inference

Year	2006	2007	2008	2009	2010	Average
# of injuries	22	3	16	36	23	**20**

Not every year is equal to 20. There are fluctuations!
If we see the number of injuries each year as the realization of random variables $X_{2006}, X_{2007}, X_{2008}, X_{2009}, X_{2010}$ the average

$$\bar{X} = \frac{X_{2006} + X_{2007} + X_{2008} + X_{2009} + X_{2010}}{5}$$

is also a random variable. Thus 20 is only an estimate of the true average number of injuries λ

Statistical inference

We may ask how close 20 is likely to be from the true value λ

Other questions we may ask:

1. Is $\lambda \geq 21$?
2. Is $19 \leq \lambda \leq 21$?
3. What is the smallest interval in which λ is likely to be?

Answering such questions is called

Statistical inference

We may ask how close 20 is likely to be from the true value λ

Other questions we may ask:

1. Is $\lambda \geq 21$?
2. Is $19 \leq \lambda \leq 21$?
3. What is the smallest interval in which λ is likely to be?

Answering such questions is called

Statistical inference

The 2009 cherry blossom run

The credit union Cherry Blossom Run takes place every year in D.C.
In 2009 there were 14974 participants in the 10 mile race.
Using the R command data(run10), you can have the following information on each participant:

The 2009 cherry blossom run

The credit union Cherry Blossom Run takes place every year in D.C.
In 2009 there were 14974 participants in the 10 mile race.
Using the R command data(run10), you can have the following information on each participant:

place

time

hometown

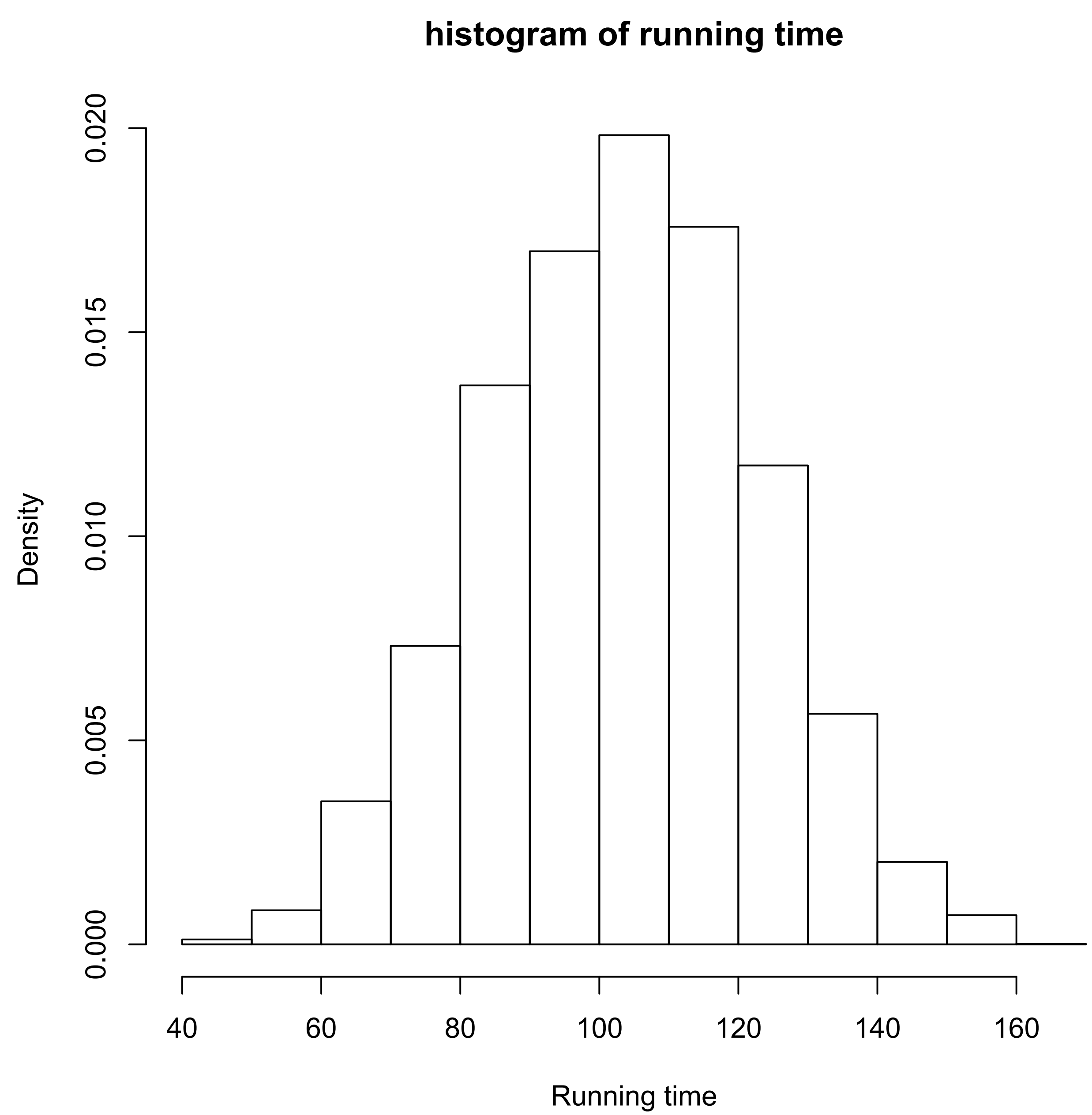
home country

age

gender

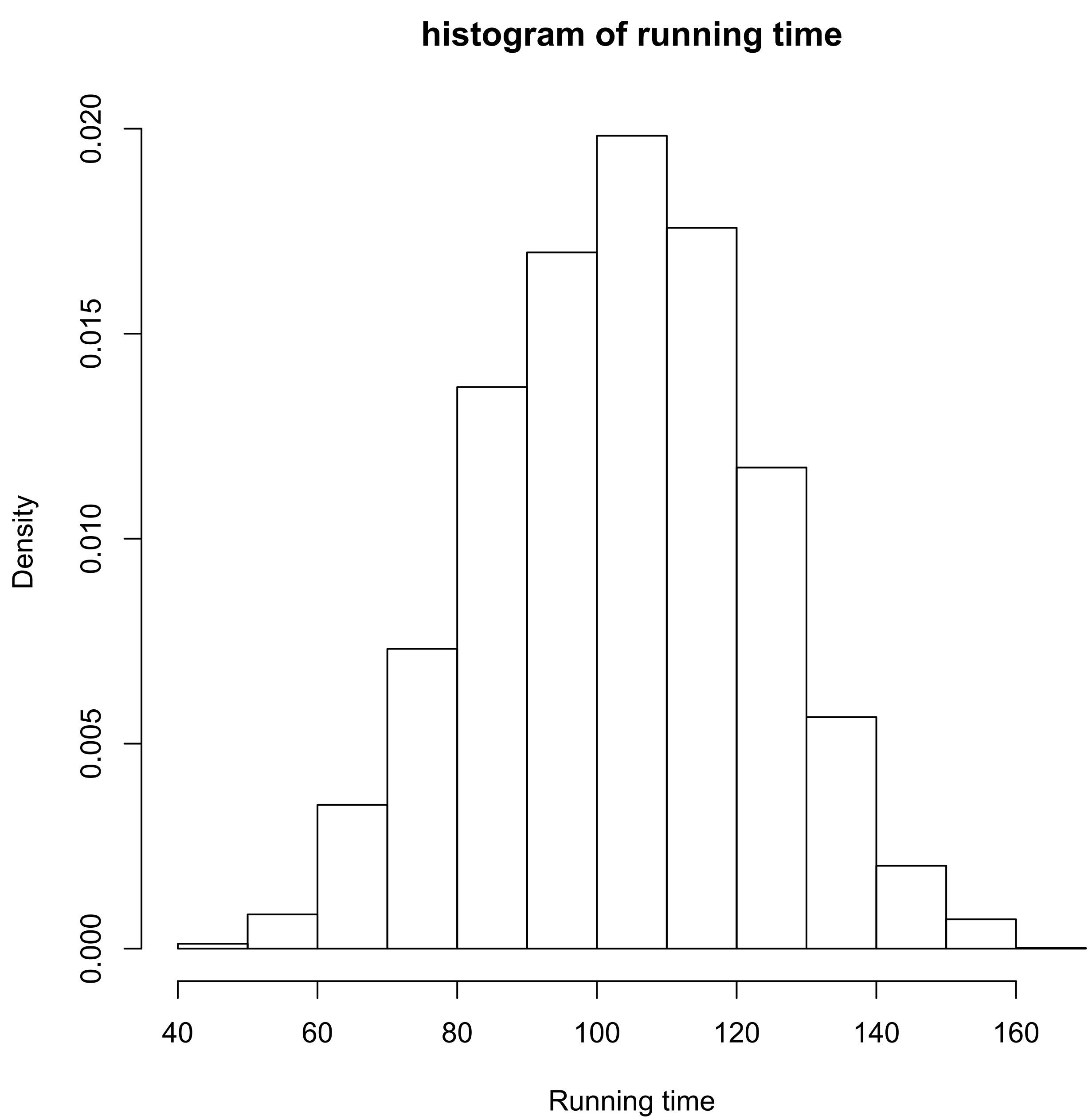
Running time

Let us focus on : the running time



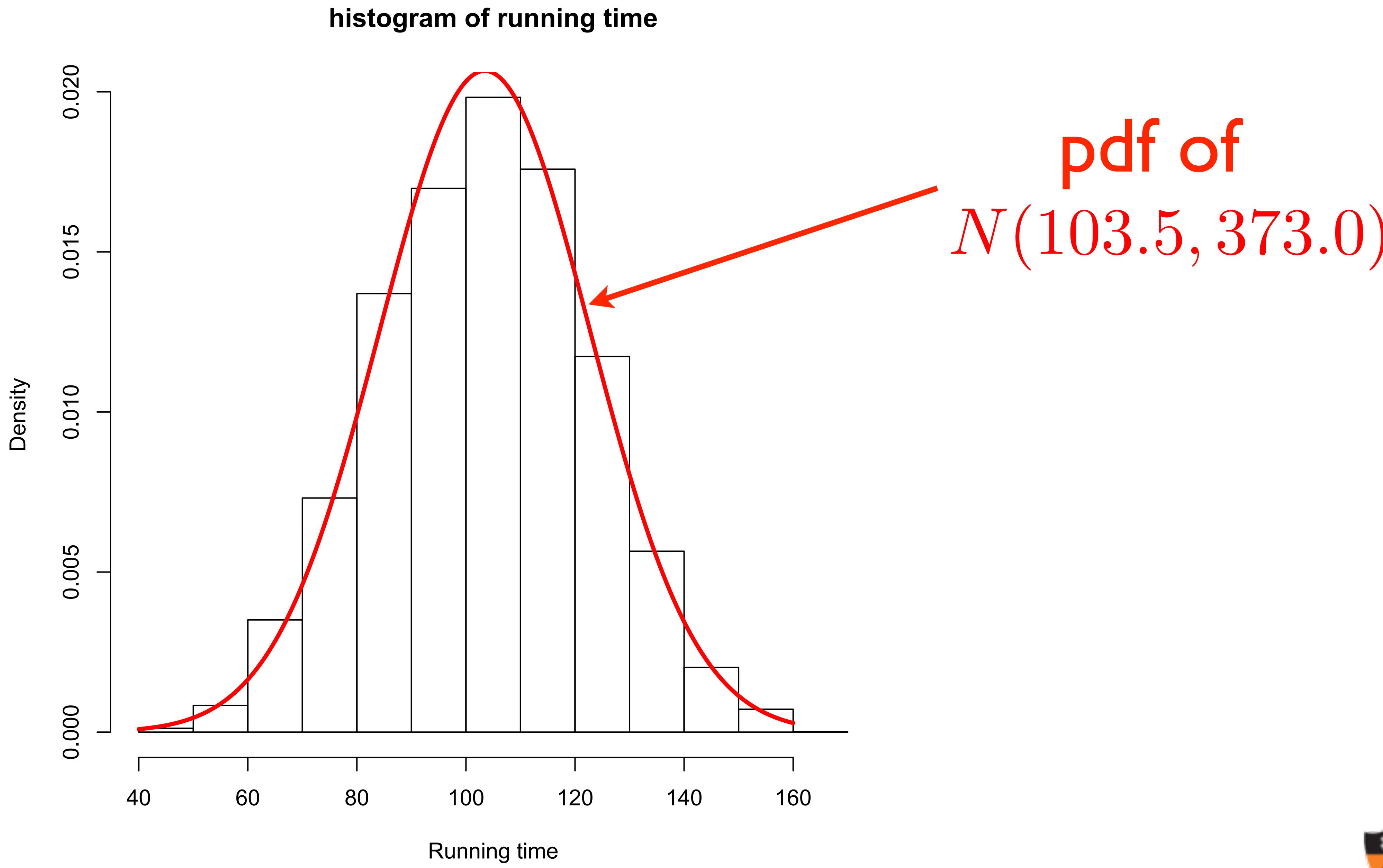
Running time

Let us focus on time : the running time



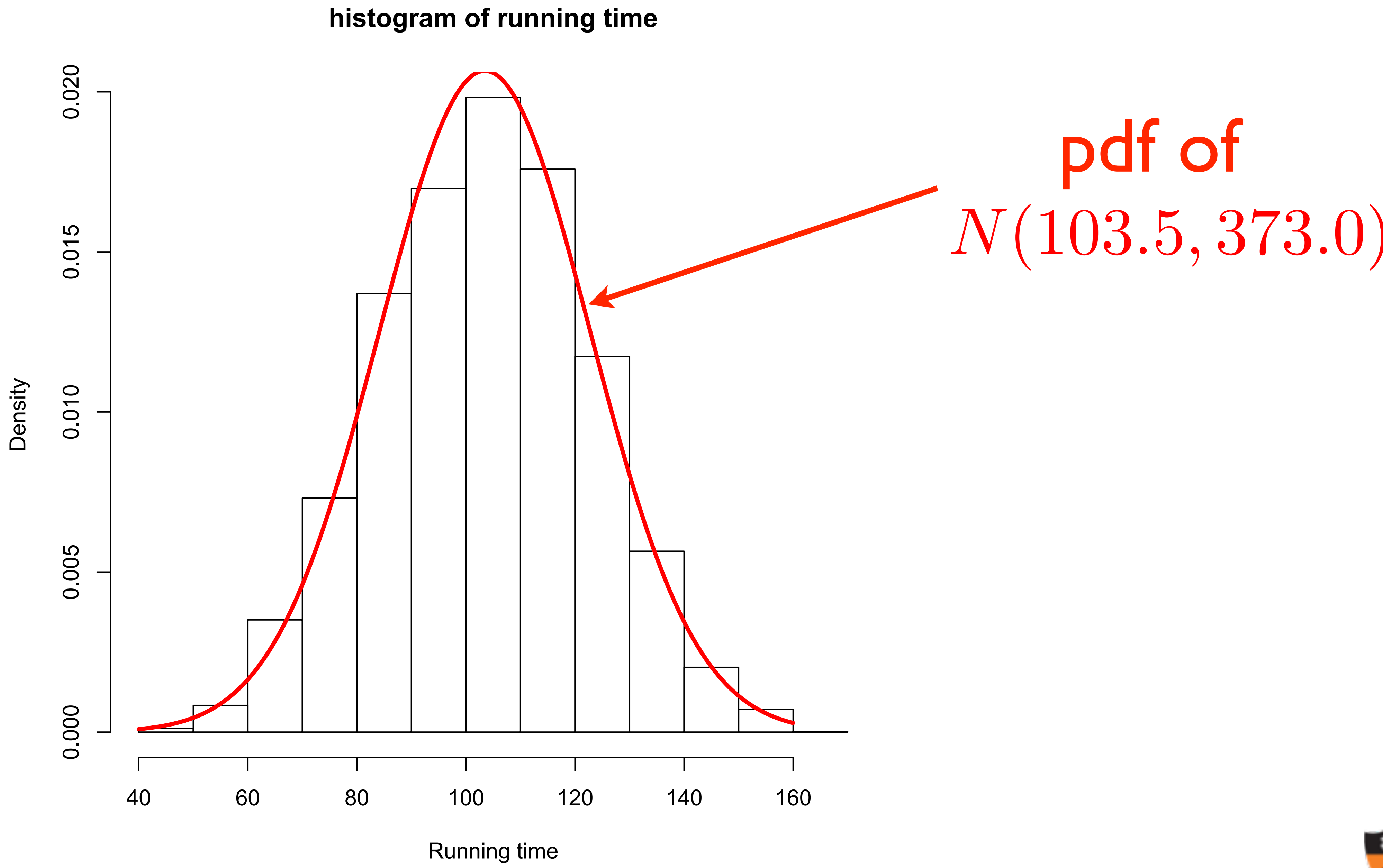
Running time

Let us focus on : the running time



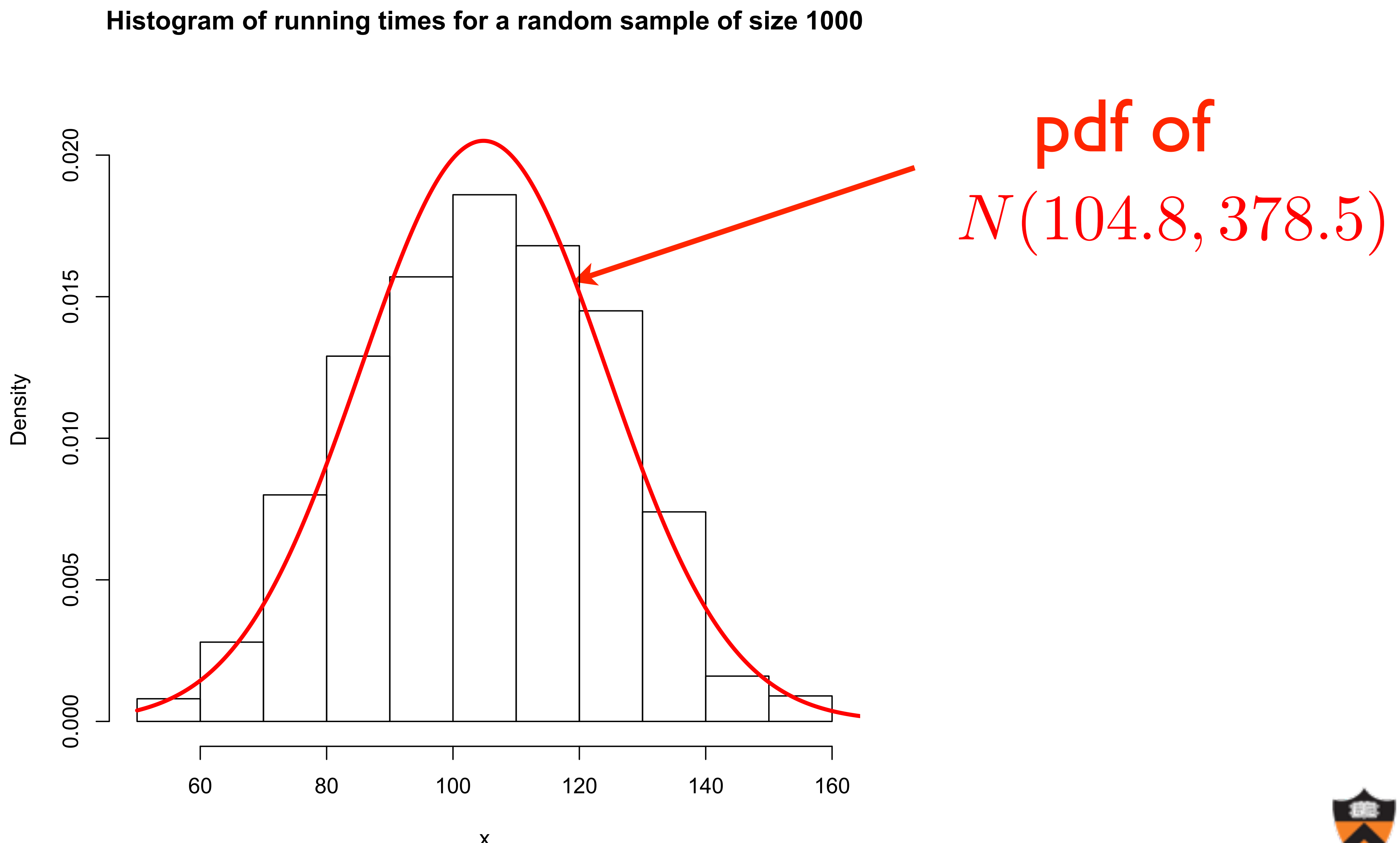
Running time

Let us focus on time : the running time



Sample

Assume now that we look only at 1000 randomly selected runners. We get



Estimates

Where do the numbers 104.8 and 378.5 come from?

They are estimates of true **unknown** parameters μ and σ^2

They are **numbers** (an estimate is a number!).

But would you bet 100$ that in 2009 the average running time was 104.8 overall?

If you said yes, you will most likely loose 100$ simply because of fluctuations

Estimates Vs estimators

We computed 104.8 by taking the average of all running times:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_{1000}}{1000} = 104.8$$

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{1000}}{1000}$$

Estimates Vs estimators

We computed 104.8 by taking the average of all running times:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_{1000}}{1000} = 104.8$$

We can see this number as the **realization** of a random variables which is the average of random variables

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{1000}}{1000}$$

Estimates Vs estimators

We computed 104.8 by taking the average of all running times:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_{1000}}{1000} = 104.8$$

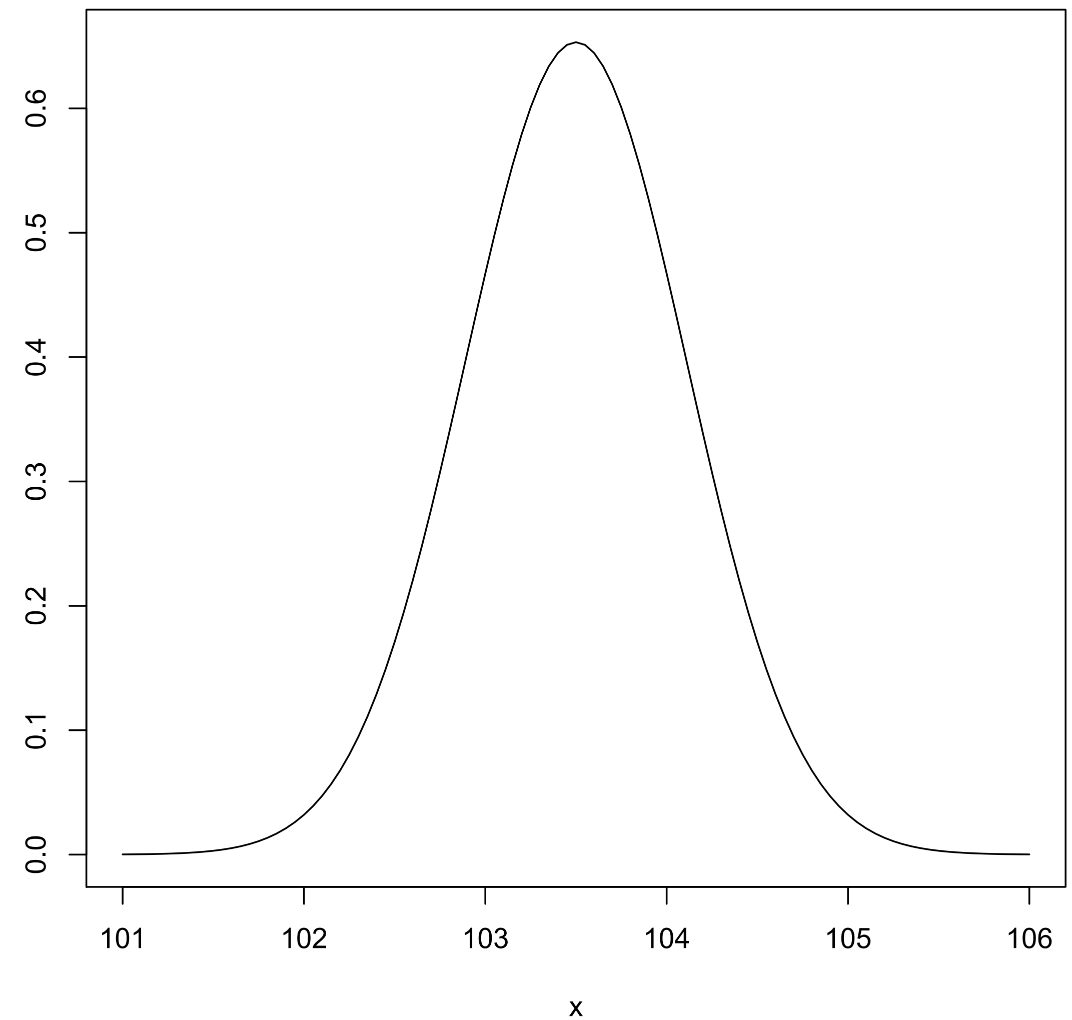
We can see this number as the **realization** of a random variables which is the average of random variables

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{1000}}{1000}$$

This is an **estimator** (it's not a number! It's a random variable)

Estimates Vs estimators

$$\bar{X} \sim N(103.5, \frac{373.0}{1000})$$



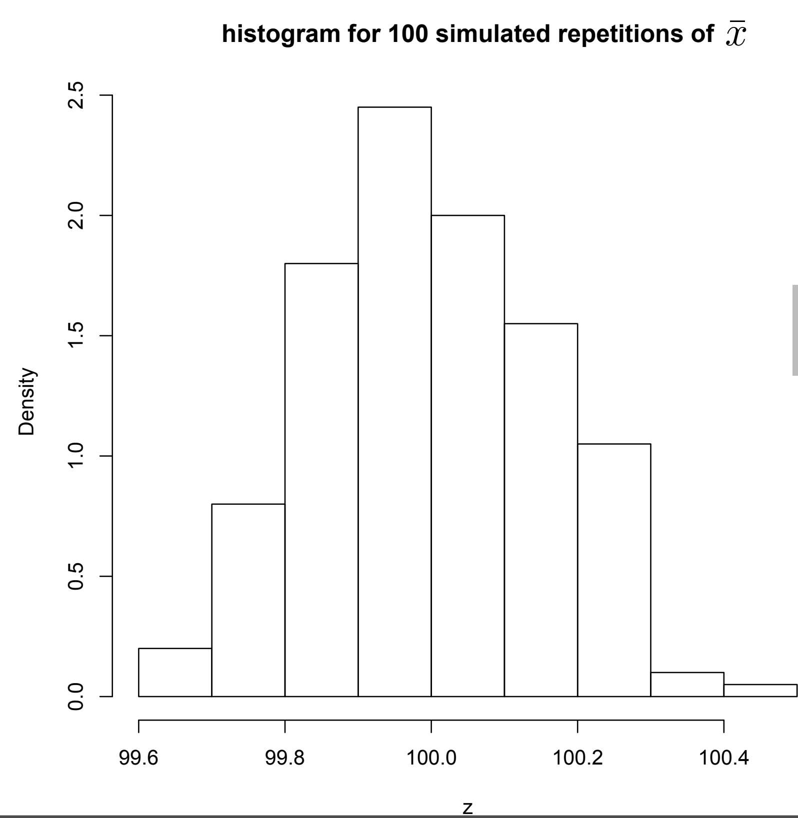
It is very likely that we can observe averages between 102 and 105

But we can already answer question from statistical inference: the true expected value is **not** equal to 100

Variability of the of the mean

We know that if the variance (or standard deviation) of the estimator $\bar{X}$ is small, then there will be small variability around its expected value $E(\bar{X}) = \mu$

What does it mean for the estimate $\bar{x}$?



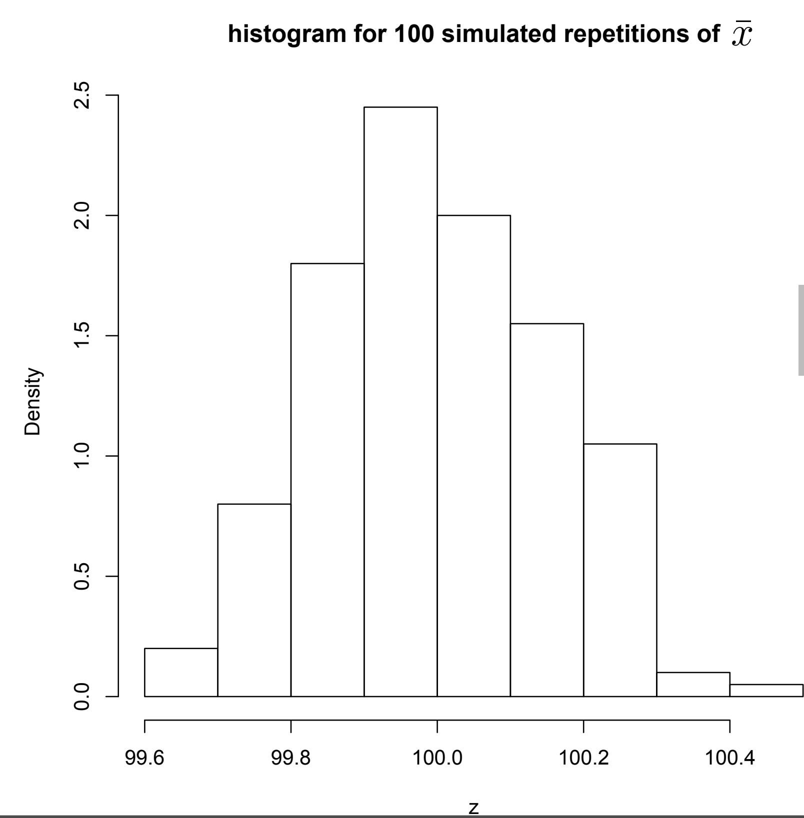

```
z=rnorm(200, 100, sqrt(350/14974))
hist(z, freq=F)
```

Variability of the of the mean

We know that if the variance (or standard deviation) of the estimator $\bar{X}$ is small, then there will be small variability around its expected value $E(\bar{X}) = \mu$

What does it mean for the estimate $\bar{x}$?

The estimate is a number: no variability!




```
z=rnorm(200, 100, sqrt(350/14974))
hist(z, freq=F)
```

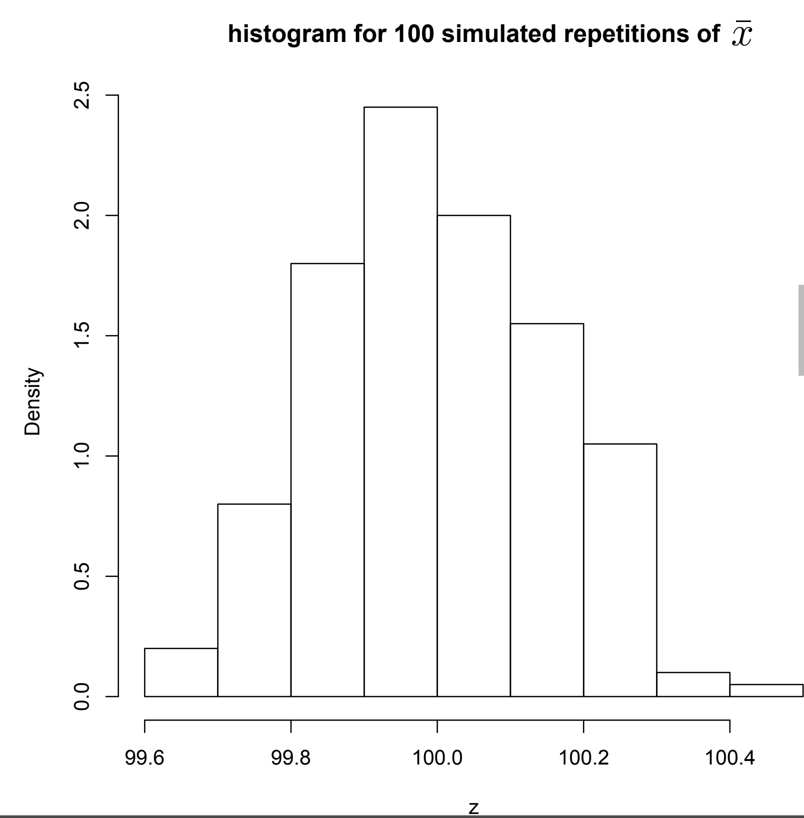
Variability of the of the mean

We know that if the variance (or standard deviation) of the estimator $\bar{X}$ is small, then there will be small variability around its expected value $E(\bar{X}) = \mu$

What does it mean for the estimate $\bar{x}$?

The estimate is a number: no variability!

But if we repeat the experiment, we will get different values for $\bar{x}$ and we can build a histogram




```
z=rnorm(200, 100, sqrt(350/14974))
hist(z, freq=F)
```

Standard error

The standard deviation of an **estimator** (it's a random variable; for example $\bar{X}$) is often called standard error

In the case of $\bar{x}$ and if the n observations are i.i.d, we have

$$SE(\bar{x}) = \sqrt{var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

where $\sigma^2 = var(X_1) = \cdots = var(X_n)$
(Note that we already know that this is true for normally distributed random variables)

We do not know σ^2 but we estimate it by s^2

Summary

- Estimates (for example $\bar{x}$) are numbers that give a good prediction of true unknown parameter.

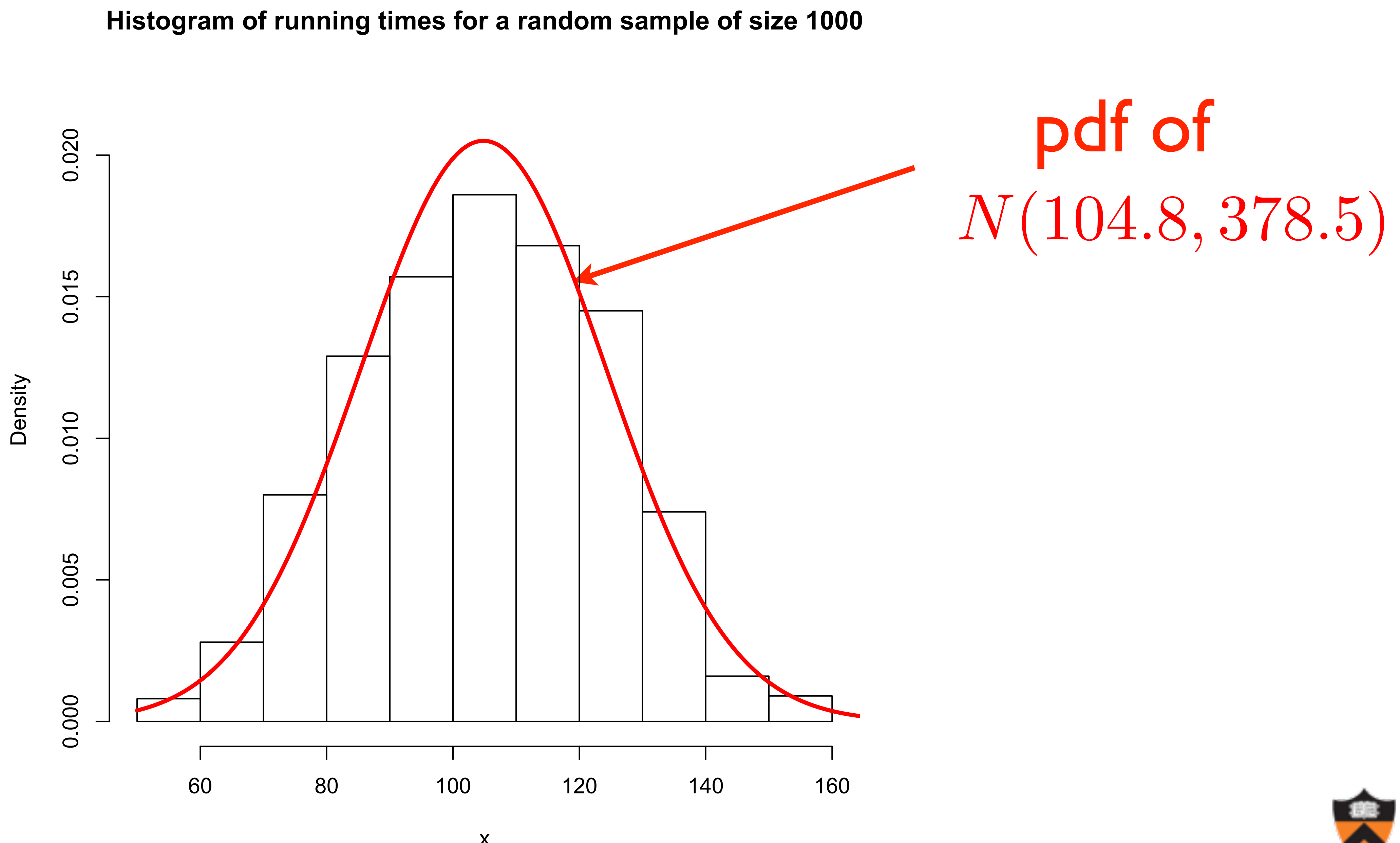
- Estimates are subject to variability: if we repeat the experiment, we may get another value

- Estimators (for example $\bar{X}$) are random variables that allow us to understand the variability of the estimate: we see the estimate as a realization of the estimator.

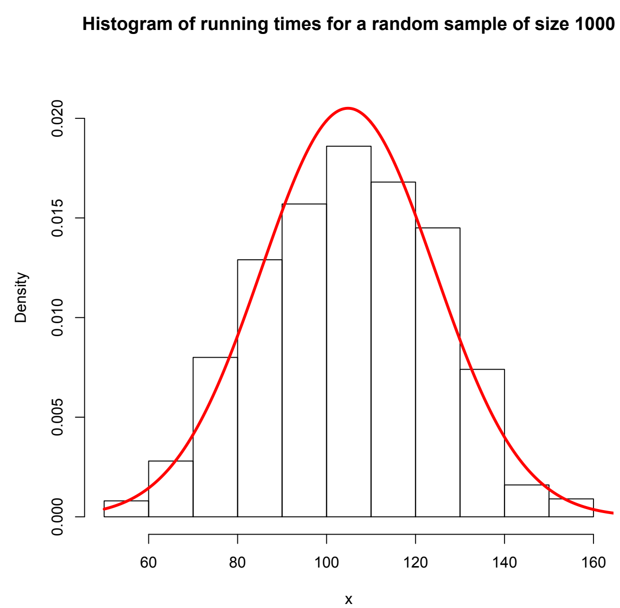
- The larger the sample size, the smaller the variability of the estimate.

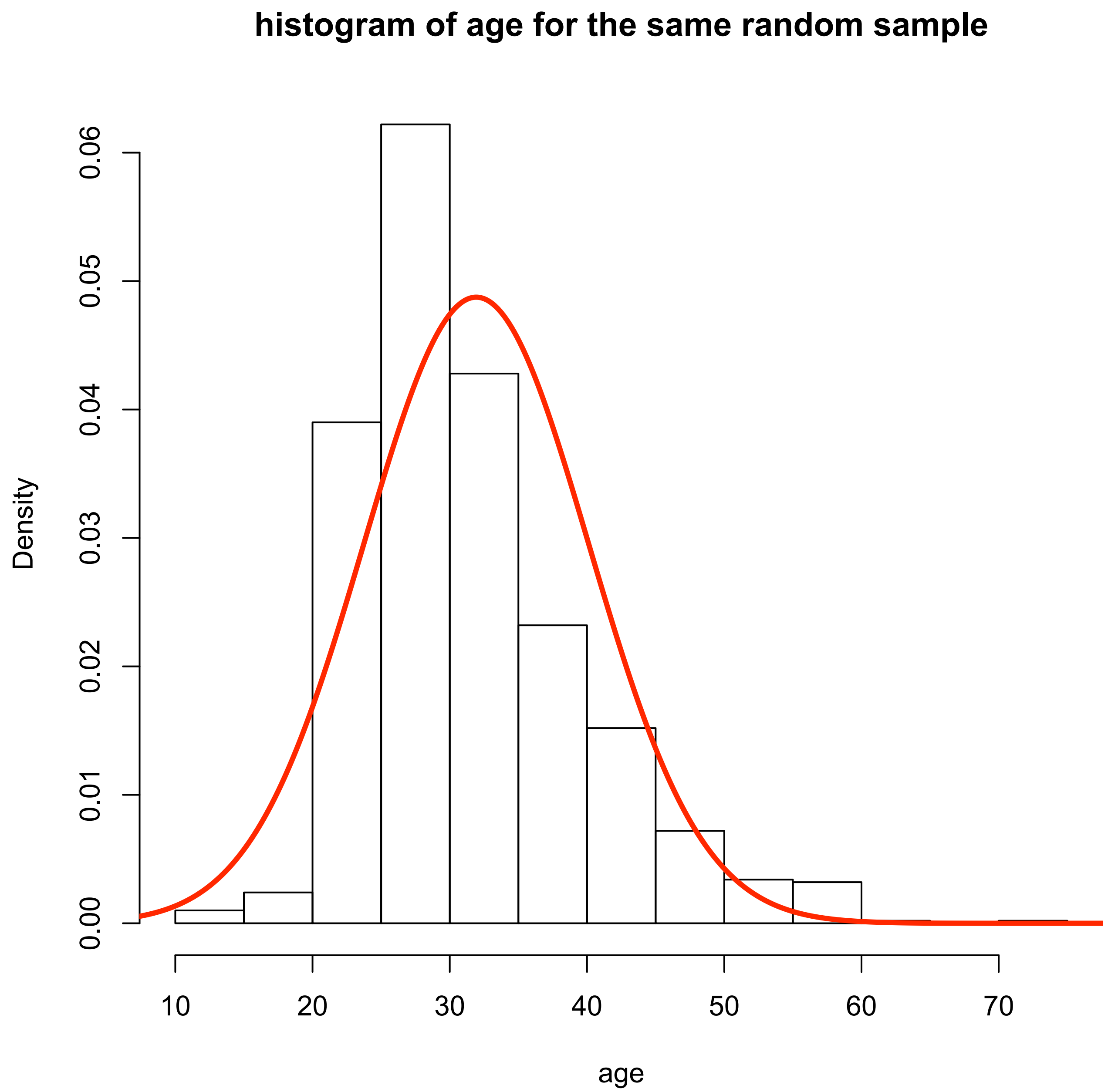
Sampling distribution

We had the following histogram for our sample of size 1000 → Almost normal distribution!



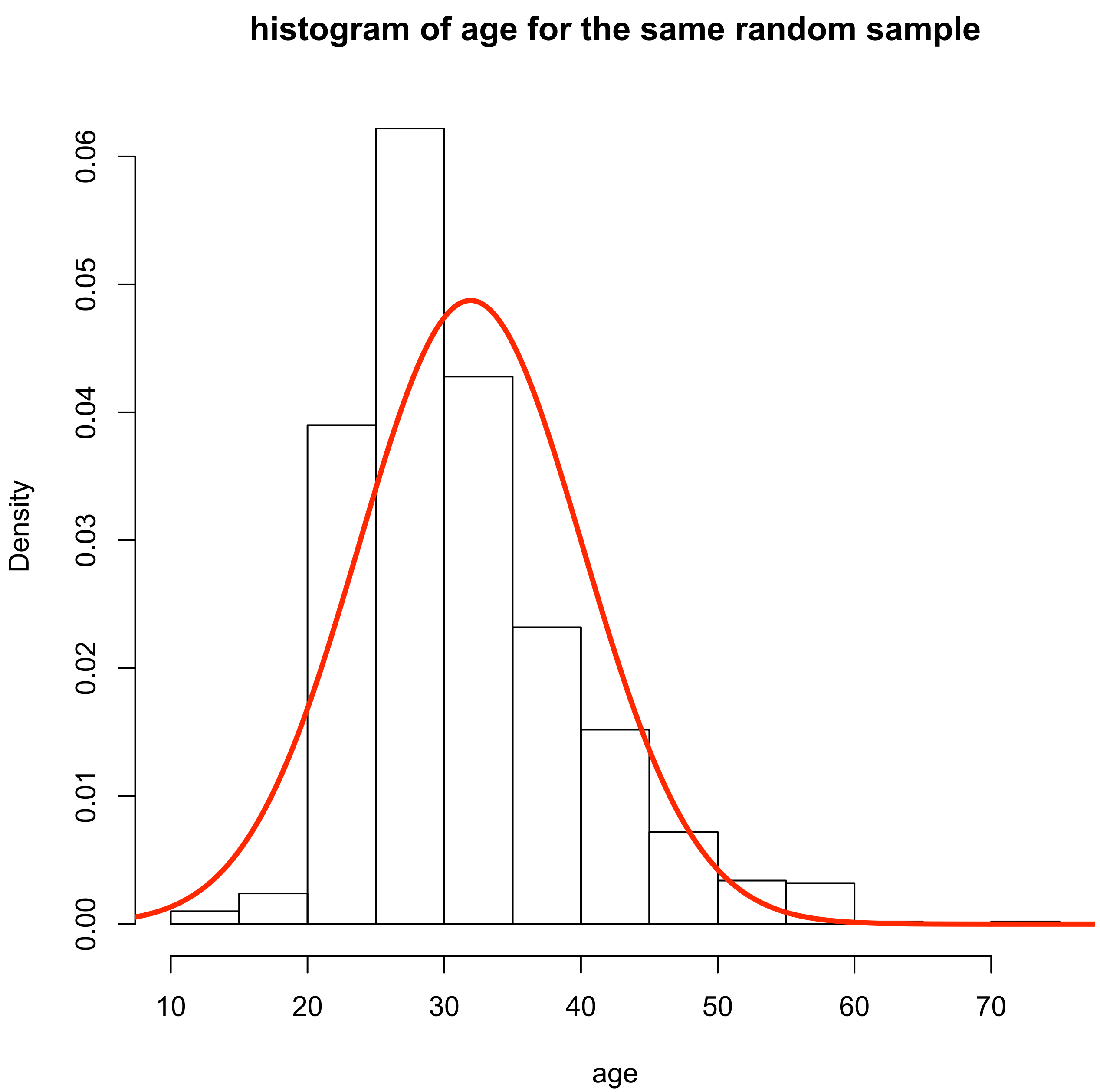
Sampling distribution

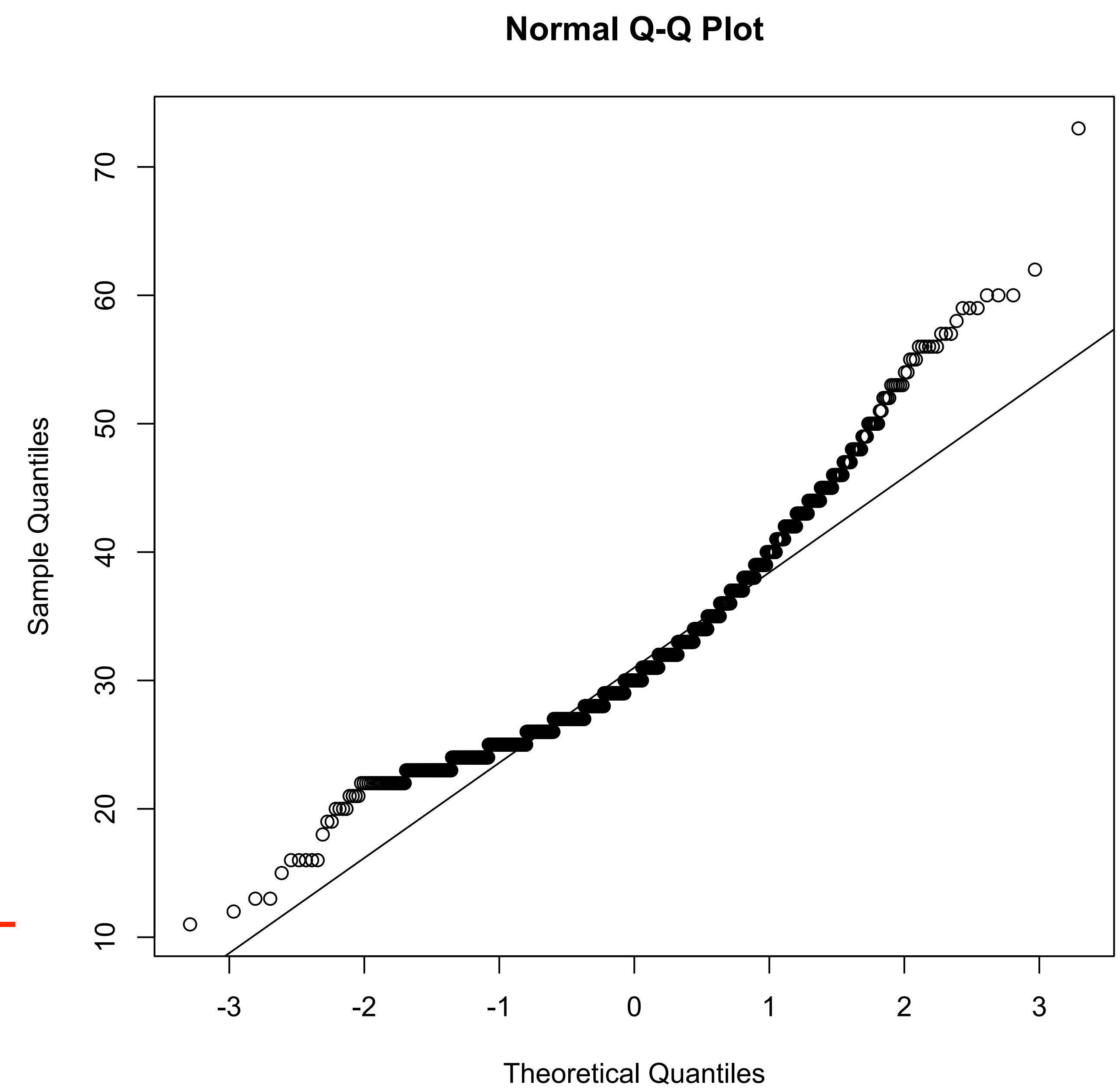




Does it look like a normal distribution?

Sampling distribution



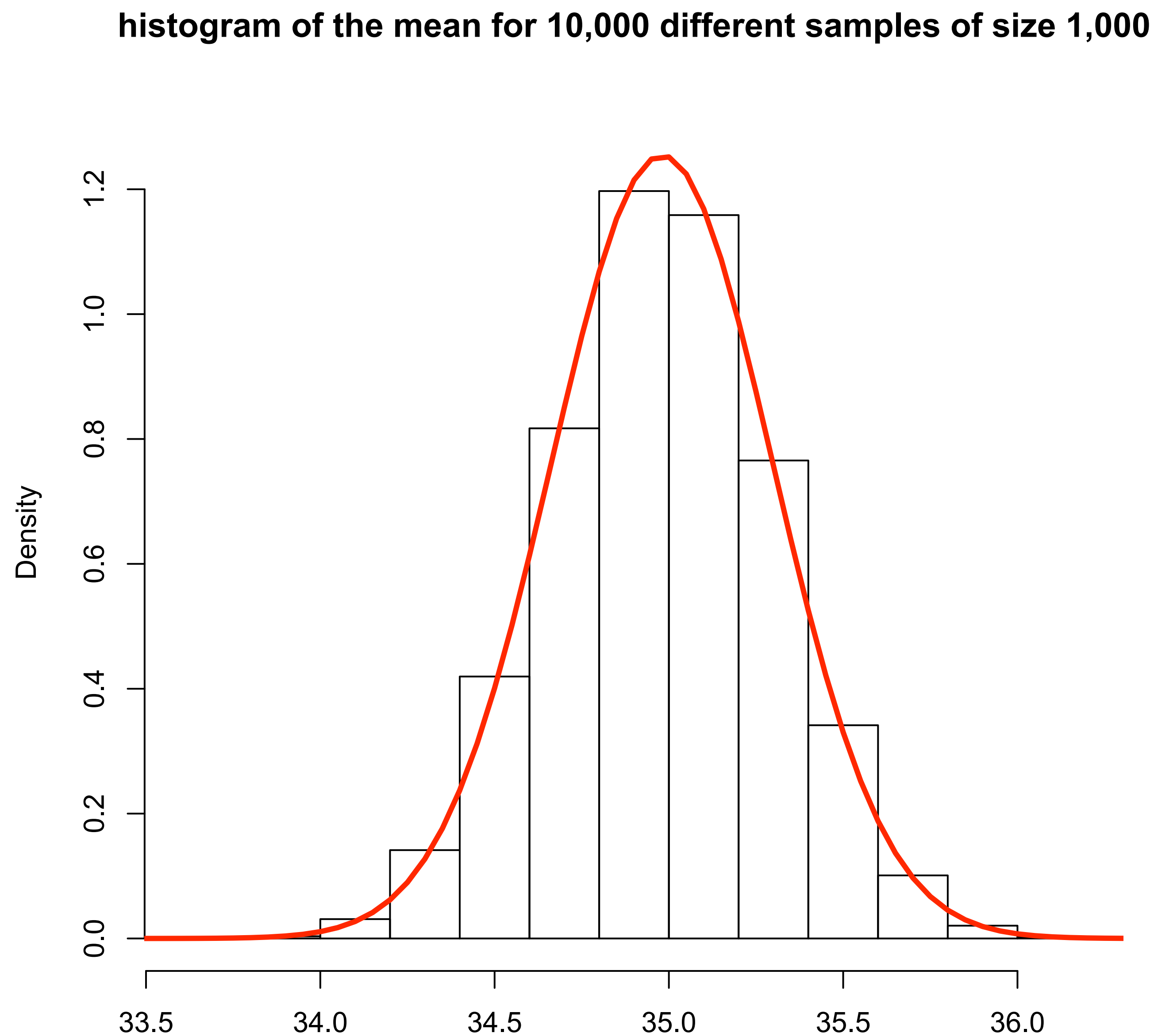


normal Q-Q plot

Sampling distribution

The random sample **does not** have a normal distribution.

The mean $\bar{X}$ does (approximately)



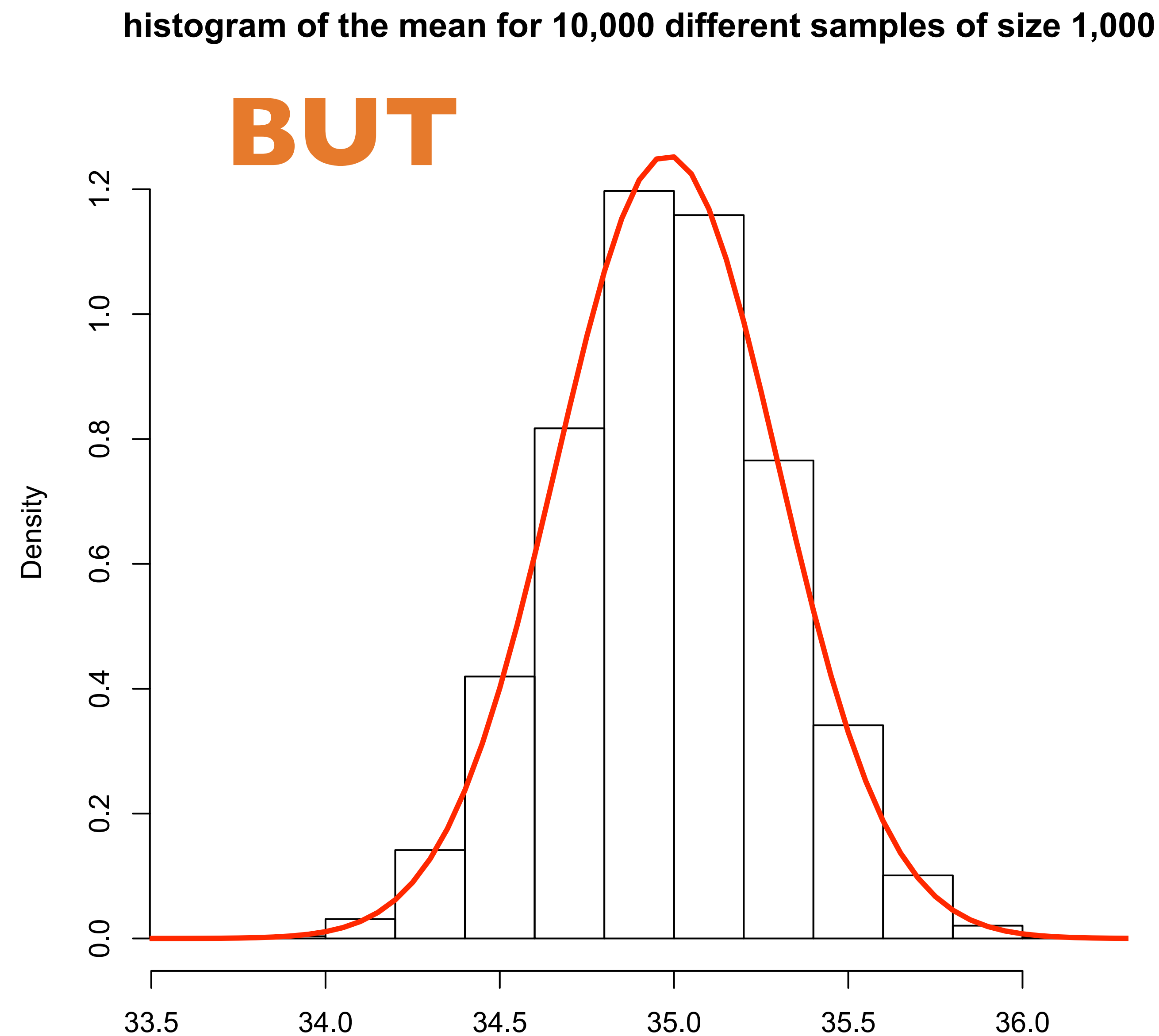

```
m=c()
for (i in 1:10000){m=c(m,mean(sample(run10$age, 1000, replace=TRUE)))}
hist(m, main="histogram of the mean for 10,000 different samples of size 1,000", xlab="mean", freq=F, ylim=c(0,1.3))
x=seq(33, 37, by=0.05)
lines(x, dnorm(x, mean=mean(m), s=sd(m)), col=2, lwd=3)
```

Sampling distribution

The random sample **does not** have a normal distribution.

BUT

The mean $\bar{X}$ does (approximately)




```
m=c()
for (i in 1:10000){m=c(m,mean(sample(run10$age, 1000, replace=TRUE)))}
hist(m, main="histogram of the mean for 10,000 different samples of size 1,000", xlab="mean", freq=F, ylim=c(0,1.3))
x=seq(33, 37, by=0.05)
lines(x, dnorm(x, mean=mean(m), s=sd(m)), col=2, lwd=3)
```

Central limit theorem

We rarely (never!) have the opportunity to draw 10,000 samples but a **theorem** tells us that this is always true as long as the sample size is large enough

If the sample consists of at least

50 independent

observations then

$$\bar{X} \sim N(\mu, \sigma^2)$$

Other estimates/estimators

Note that the sample mean is not the only possible estimate but it's a good candidate to estimate the expected value

The central limit theorem (CLT) is valid only for the mean.

Confidence intervals

We know that $\bar{x} = 104.8$ is not the true value of the expected running time. But can we provide an interval (a range of plausible values) for the true expected running time?
A confidence interval is a plausible range of values for the true unknown parameter.

From the book:
"Using only a point estimate is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net. We can throw a spear where we saw a fish but we will probably miss. On the other hand if we toss a net in that area, we have a good chance of catching the fish."

Confidence level

We can use a very large interval (0-200 minutes) but it is not very informative.
We can also use a vary narrow interval (104.75-104.91) but it is almost as using a single value.

$$\bar{x} \pm \text{ something}$$

Confidence level

We can use a very large interval (0-200 minutes) but it is not very informative.
We can also use a vary narrow interval (104.75-104.91) but it is almost as using a single value.

There is a **tradeoff** between confidence and accuracy

$$\bar{x} \pm \text{ something}$$

Confidence level

We can use a very large interval (0-200 minutes) but it is not very informative.
We can also use a vary narrow interval (104.75-104.91) but it is almost as using a single value.

There is a **tradeoff** between confidence and accuracy

We usually proceed as follows: given a pre-specified confidence level (typically 95% but also 90% or 99%) we try to construct the narrowest possible confidence interval.
We also favor confidence intervals that are symmetric about the point estimate of the unknown parameter such as

$$\bar{x} \pm \text{ something}$$

Approximate 95% confidence interval

As a rule of the thumb many practitioners use the following approximate 95% confidence interval

$$\text{point estimate} \pm\ 2{*}\text{standard error}$$

In the case of $\bar{x}$ this gives the following confidence interval for μ

$$\bar{x} \pm 2\frac{s}{\sqrt{n}}$$

In our example: $\bar{x} = 104.8,\ s = \sqrt{378.5} = 19.5,\ n = 1000$

so the 95% confidence interval is $[103.56, 106.03]$

Approximate 95% confidence interval

As a rule of the thumb many practitioners use the following approximate 95% confidence interval

$$\text{point estimate} \pm\ 2{*}\text{standard error}$$

In the case of $\bar{x}$ this gives the following confidence interval for μ

$$\bar{x} \pm 2\frac{s}{\sqrt{n}}$$

In our example: $\bar{x} = 104.8,\ s = \sqrt{378.5} = 19.5,\ n = 1000$

so the 95% confidence interval is $[103.56, 106.03]$

But what does 95% confidence mean?

Approximate 95% confidence interval

Indeed, this is either 0 (false) or 1 (true). There is nothing random here. Remember that $\mu = 103.5$ is a number.

The 95 confidence level means that if we repeat the experiment 100 times (100 different samples of 1000 runners) the true μ will be in 95 of the constructed confidence intervals. (Here it does not, we were unlucky!)

Approximate 95% confidence interval

But what does 95% confidence mean?

Indeed, this is either 0 (false) or 1 (true). There is nothing random here. Remember that $\mu = 103.5$ is a number.

The 95 confidence level means that if we repeat the experiment 100 times (100 different samples of 1000 runners) the true μ will be in 95 of the constructed confidence intervals. (Here it does not, we were unlucky!)

Approximate 95% confidence interval

But what does 95% confidence mean?

It does not mean that $P(103.56 \leq \mu \leq 106.03)$

Indeed, this is either 0 (false) or 1 (true). There is nothing random here. Remember that $\mu = 103.5$ is a number.

The 95 confidence level means that if we repeat the experiment 100 times (100 different samples of 1000 runners) the true μ will be in 95 of the constructed confidence intervals. (Here it does not, we were unlucky!)

Approximate 95% confidence interval

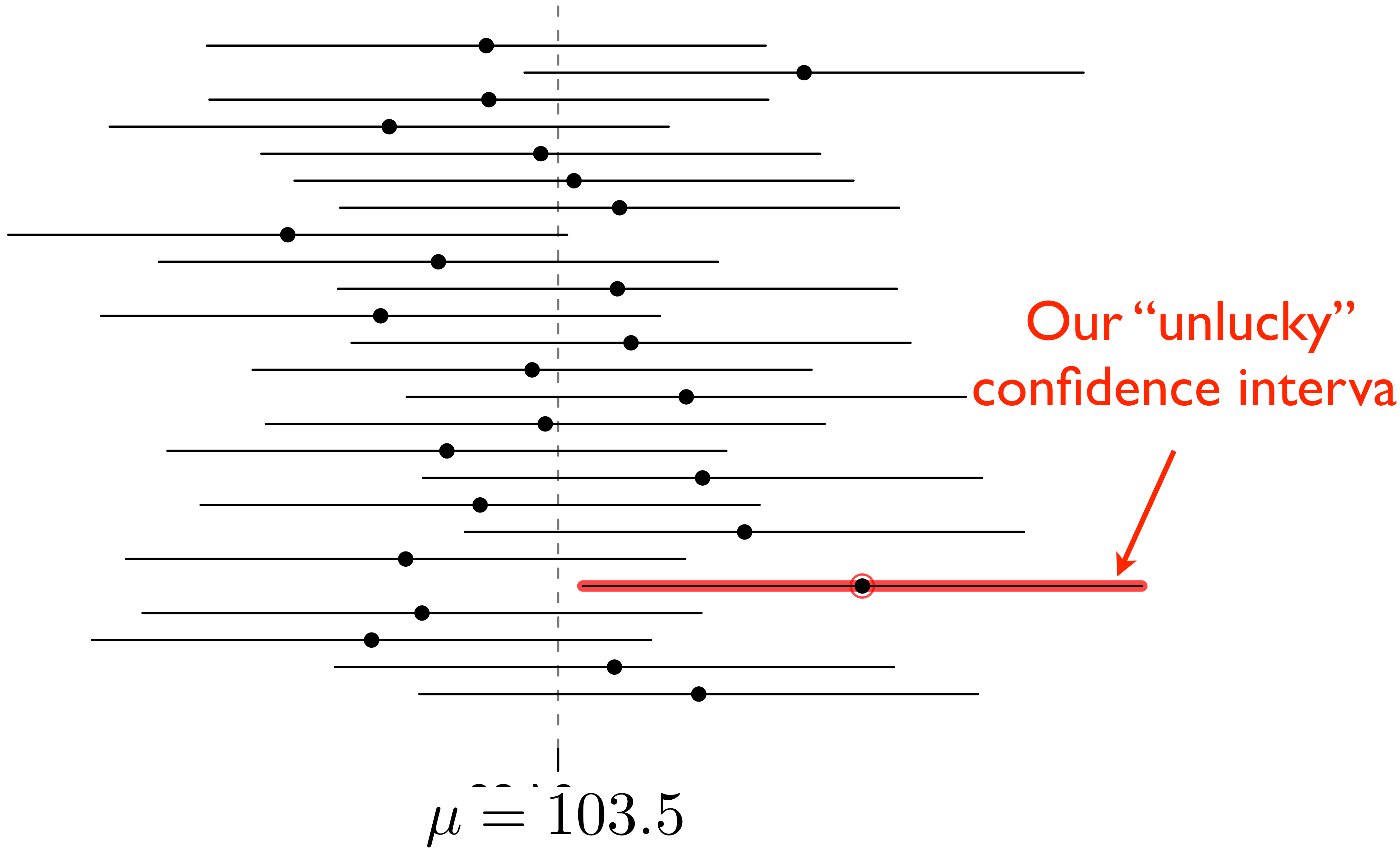
But what does 95% confidence mean?

It does not mean that $P(103.56 \leq \mu \leq 106.03)$

Indeed, this is either 0 (false) or 1 (true). There is nothing random here. Remember that $\mu = 103.5$ is a number.

The 95 confidence level means that if we repeat the experiment 100 times (100 different samples of 1000 runners) the true μ will be in 95 of the constructed confidence intervals. (Here it does not, we were unlucky!)

Approximate 95% confidence interval



Out of 25 confidence intervals, one did not contain the true value of μ (.95*25=23.75)

Why 2?

Recall the (approximate) 95% confidence interval

$$\bar{x} \pm 2\frac{s}{\sqrt{n}}$$

The number 2 is actually an **approximation** for 1.96
Indeed, we want to show that

$$P(\bar{X} - 2\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + 2\frac{s}{\sqrt{n}}) \geq .95$$

We can write this because $\bar{X}$ is a random variable (not $\bar{x}$)
this is the mathematical **definition** of the confidence interval

Z-score of $\bar{X}$

$$P(\bar{X} - 2\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + 2\frac{s}{\sqrt{n}})$$

$$= P(-2 \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq 2)$$

$$\simeq P(-2 \leq Z \leq 2)$$

Z-score of $\bar{X}$

$$P(\bar{X} - 2\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + 2\frac{s}{\sqrt{n}})$$

$$= P(-2 \le \frac{\bar{X} - \mu}{s/\sqrt{n}} \le 2)$$

$$\simeq P(-2 \le Z \le 2)$$

Z-score of $\bar{X}$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Z-score of $\bar{X}$

$$P(\bar{X} - 2\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + 2\frac{s}{\sqrt{n}})$$

$$= P(-2 \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq 2)$$

$$\simeq P(-2 \leq Z \leq 2)$$

Z-score of $\bar{X}$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

From the central limit theorem, we know that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ so its Z-score $Z \sim N(0,1)$

It simply remains to check that $P(-2 \leq Z \leq 2) \geq .95$

Normal probability 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
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
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

ORF 245 -
Prof. Rigollet
Spring 2011

Normal probability table

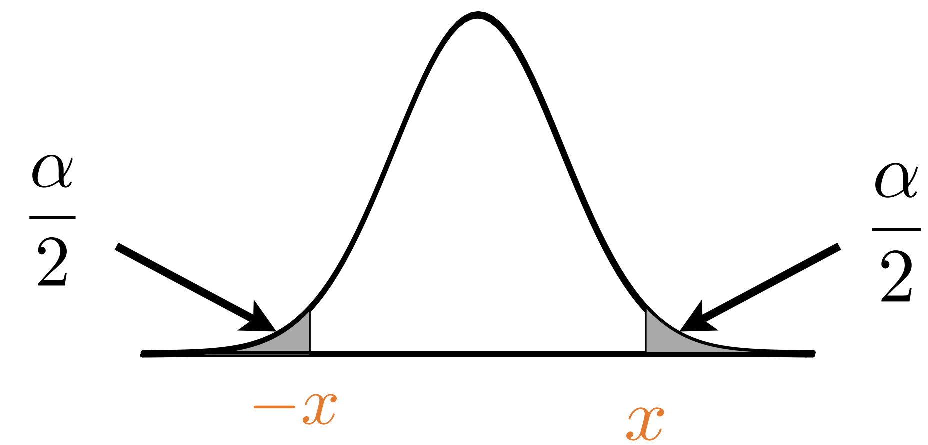
Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	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
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
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

Why 2?

The number 2 is actually an **approximation** for 1.96

We read from the table that $P(Z \leq 1.96) = 0.975$
Or equivalently that $P(Z \geq 1.96) = 0.025$

From Chapter 3 (symmetry) we had:



Here $x = 1.96,\ \dfrac{\alpha}{2} = 0.025$ so

$$P(-1.96 \leq Z \leq 1.96) = 1 - 0.025 - 0.025 = 0.95$$

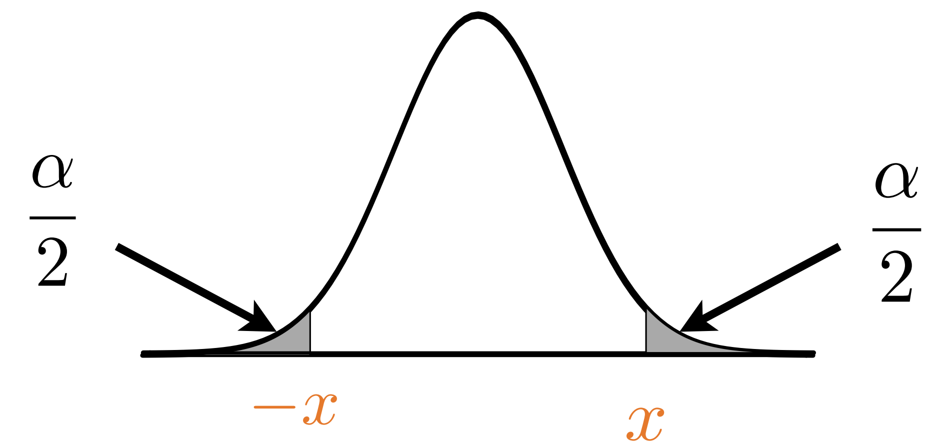
Changing the confidence level

How can we build a 99% confidence interval?
We need to find x such that

$$P(-x \leq Z \leq x) = .99$$

Therefore we need to take

$$\frac{\alpha}{2} = .005$$



It yields

$$P(Z \leq x) = 1 - \frac{\alpha}{2} = 1 - 0.005 = .995$$

Normal probability table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
	Second decimal place of Z									
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
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995

Normal probability table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
	Second decimal place of Z									
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
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995

99% confidence interval

So x is between 2.57 and 2.58. We take

$$x = \frac{2.57 + 2.58}{2} = 2.575$$

and the 99% confidence interval is

$$\bar{x} \pm 2.575 \frac{s}{\sqrt{n}}$$

Going back to the running time example, we had

$$\bar{x} = 104.8, \;\; s = \sqrt{378.5} = 19.5, \;\; n = 1000$$

which gives the 99% confidence interval $[103.21, 106.39]$

This time is contains the true expected value $\mu = 103.5$

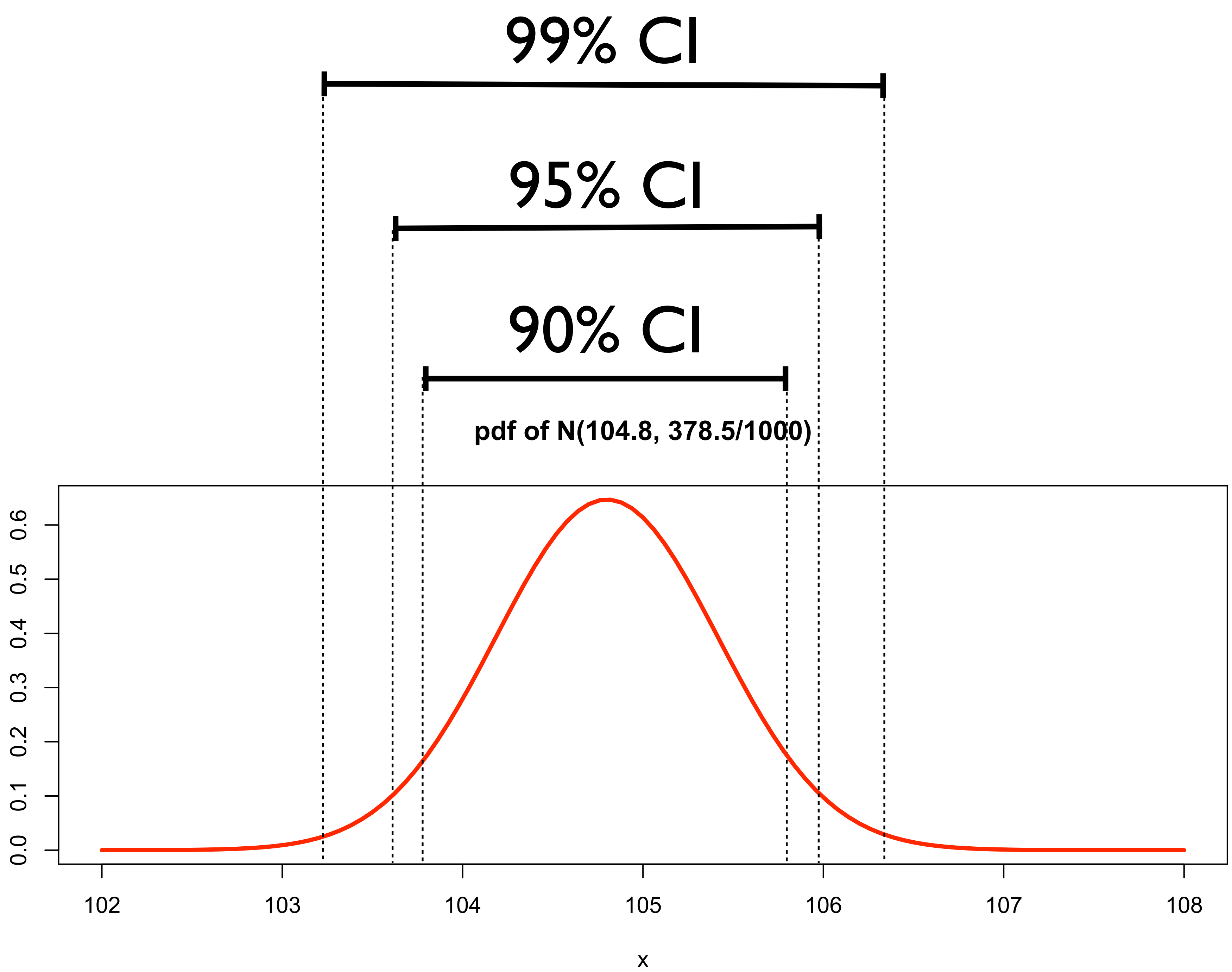
90% confidence interval

Using the same method, we can find the 90% confidence interval

$$\bar{x} \pm 1.645 \frac{s}{\sqrt{n}} \qquad [103.79, 105.81]$$

Since 1.645 < 1.96 < 2.575, the confidence intervals become **wider** when the confidence level **increases**.

Width Vs confidence level



Hypothesis testing

We already know that the true average running time for the Cherry Blossom run is 103.5 minutes. I am telling you that but you do not have to trust me and you would like to check it using a sample of size 1,000. This is a **hypothesis testing** problem.
The goal is to decide between two competing **hypotheses**:

H_0: The average running time is indeed 103.5 minutes
H_A: The average running time is *different* from 103.5 minutes

We call H_0 the **null hypothesis**
We call H_A the **alternative hypothesis**

Null Vs alternative

While it seems that the null and alternative hypotheses can be interchanged, they play a very different role

H_0 (null) represents the "status quo" (a perspective of no difference) or a skeptical position.

H_A (alternative) represents the claim under consideration, a discovery, a novelty.

In our example, the current position is that the expected value is 103.5 and we want to discover if this may be a false statement. Why would you think that I am lying?

Evidence in the data

The asymmetric role of the null and the alternative hypotheses lies in the fact that we want to find evidence **in the data** to prove that the null hypothesis is wrong in favor of the alternative.

An example to keep in mind: *"innocent until proven guilty"*
A person is always assumed to be innocent by default (null). It is the role of the prosecutor to bring significant evidence against innocence.

Concluding to the null hypothesis does not mean that it is true. It only means that we could not find enough evidence in the data. *(remember that the sentence of a jury is "not guilty" and not "innocent")*

Drug testing

Pharmaceutical companies use hypothesis testing **all the time** to test if a new drug is efficient.
To do so, they administer a drug to a group of patient (test group) and a placebo to another group (control group).

Assume that the drug is a cough syrup.
Let $\mu_{control}$ denote the expected number of expectorations per hour after a patient has used the placebo.
Let μ_{drug} denote the expected number of expectorations per hour after a patient has used the syrup.

$$H_0 : \mu_{drug} \geq \mu_{control} \qquad H_0 : \mu_{drug} \leq \mu_{control}$$

$$H_A : \mu_{drug} < \mu_{control} \quad \text{or} \quad H_A : \mu_{drug} > \mu_{control}$$

Drug testing

$$H_0 : \mu_{drug} \geq \mu_{control}$$
$$H_A : \mu_{drug} < \mu_{control}$$

The pharmaceutical company needs to **bring evidence** that the syrup is working (better that the placebo). A drug that is better than the pacebo will show a smaller number of expected expectorations.

If we could not conclude to H_A it does not mean that the drug is worse that a placebo. We say that

"we failed to reject H_0"

"H_0 is not implausible"

Hypothesis testing with confidence intervals

Let us go back to our test for the Cherry blossom run

H_0: The average running time is indeed 103.5 minutes
H_A: The average running time is *different* from 103.5 minutes

Mathematically, it reads:

$$\boxed{\begin{aligned} H_0 &: \mu = 103.5 \\ H_A &: \mu \neq 103.5 \end{aligned}}$$

Recall that our 95% confidence interval for μ was

$$[103.56, 106.03]$$

It means that a range of plausible value is $[103.56, 106.03]$

It means that 103.5 is **not** a plausible value: we reject H_0

Hypothesis testing with confidence intervals

What if we used the 99% confidence interval:

$$[103.21, 106.39]$$

Then, 103.5 becomes a plausible value: we fail to reject H_0

What has changed?

Hypothesis testing with confidence intervals

What if we used the 99% confidence interval:

$$[103.21, 106.39]$$

Then, 103.5 becomes a plausible value: we fail to reject H_0

What has changed? The confidence level

Hypothesis testing with confidence intervals

What if we used the 99% confidence interval:

$$[103.21, 106.39]$$

Then, 103.5 becomes a plausible value: we fail to reject H_0

What has changed? The confidence level

Unless specified otherwise, we use the 95% confidence interval.
In this case we would make an error (because we know that the true value is actually 103.5 and therefore that H_0 is true).

Let us look into more details at how likely it is to make an error.

Thalidomide

In 1957, a new medicine appeared on the market. Thalidomide was an effective sedative, but it was also promising as a treatment for pregnant women because it quelled nausea and vomiting. And scientists had great confidence in thalidomide's safety. It had been tested extensively [...]

But thalidomide was withdrawn from the market after only a few years. [...] thalidomide caused human limbs to stop growing prematurely in utero, resulting in the birth of babies with malformed arms and legs.

Source: WIRED

Decision errors

Hypothesis tests are subject to errors.
(In the example of thalidomide, statistics are not the only one to blame though)

The four possible scenarios when making a test:

		Decision	
		Reject H_0	Fail to reject H_0
Reality	H_0 true	Type 1 error	Correct decision
	H_A true	Correct decision	Type 2 error

The court example

In a US court the defendant is either innocent or guilty.

When does the jury make a type 1 error?

When does the jury make a type 2 error?

How could the jury make sure to make no type 1 error?
How would this effect the type 2 error?

How could the jury make sure to make no type 2 error?
How would this effect the type 1 error?

Conflicting errors

We see that if we try to reduce the error of one type, we generally make more error of the other type.

Which error should we favor? This is where the asymmetry in the hypotheses H_0 and H_A enters the game.

We said that the null hypothesis H_0 is the conservative choice and that data should bring significant evidence against it to reject it.

To quantify "*significant evidence*" we build a test that will not erroneously reject H_0 more than 5% of the time.

Conflicting errors

We see that if we try to reduce the error of one type, we generally make more error of the other type.

Which error should we favor? This is where the asymmetry in the hypotheses H_0 and H_A enters the game.

We said that the null hypothesis H_0 is the conservative choice and that data should bring significant evidence against it to reject it.

To quantify "*significant evidence*" we build a test that will not erroneously reject H_0 more than 5% of the time.

type I error

Conflicting errors

We see that if we try to reduce the error of one type, we generally make more error of the other type.

Which error should we favor? This is where the asymmetry in the hypotheses H_0 and H_A enters the game.

We said that the null hypothesis H_0 is the conservative choice and that data should bring significant evidence against it to reject it.

To quantify "*significant evidence*" we build a test that will not erroneously reject H_0 more than 5% of the time.

type I error

Why? What does it mean?

Significance level

5% is called the significance level (or simply "level") of the test (we talk about a test with significant level 5%). We can take other values (just like for confidence intervals) such as 1% or 10%.

We control the type 1 error but what about the type 2 error?

Significance level

5% is called the significance level (or simply "level") of the test (we talk about a test with significant level 5%). We can take other values (just like for confidence intervals) such as 1% or 10%.

We control the type 1 error but what about the type 2 error?

We don't

Significance level

5% is called the significance level (or simply "level") of the test (we talk about a test with significant level 5%). We can take other values (just like for confidence intervals) such as 1% or 10%.

We control the type 1 error but what about the type 2 error?

We don't

All that we can do is build a sensible test and hope that it will have small type 2 error.

For example the test that consists in never rejecting H_0 certainly has level 5% (we make 0<5% error of type 1) but has bad type 2 error.

When we use confidence intervals that are the narrowest possible we use a sensible test.

Why confidence intervals work

To manipulate probabilities for tests we have to go back to the random variables:

$$\bar{x} \rightsquigarrow \bar{X}$$

The 95% confidence interval is obtained by looking at the realization of

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

and was constructed such that

$$P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Why confidence intervals work

To manipulate probabilities for tests we have to go back to the random variables:

$$\bar{x} \rightsquigarrow \bar{X}$$

The 95% confidence interval is obtained by looking at the realization of

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

and was constructed such that

$$P(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95$$

Why does this lead to a probability of type I error of at most 5%?

Why confidence intervals work

Why does this lead to a probability of type I error of at most 5%?

By the complement rule:

$$\begin{aligned} P(\text{reject } H_0) &= 1 - P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) \\ &= 1 - 0.95 \\ &= 0.05 \end{aligned}$$

Example

Perform a test of $\boxed{\begin{array}{l} H_0 : \ \mu = 103.5 \\ H_A : \ \mu \neq 103.5 \end{array}}$ at significance level 1%.

Example

Perform a test of $\boxed{\begin{array}{l} H_0: \ \mu = 103.5 \\ H_A: \ \mu \neq 103.5 \end{array}}$ at significance level 1%.

We look at the 99% confidence interval (1-0.99=0.01) and reject if 103.5 is not in this interval. The interval is

$$[103.21, 106.39]$$

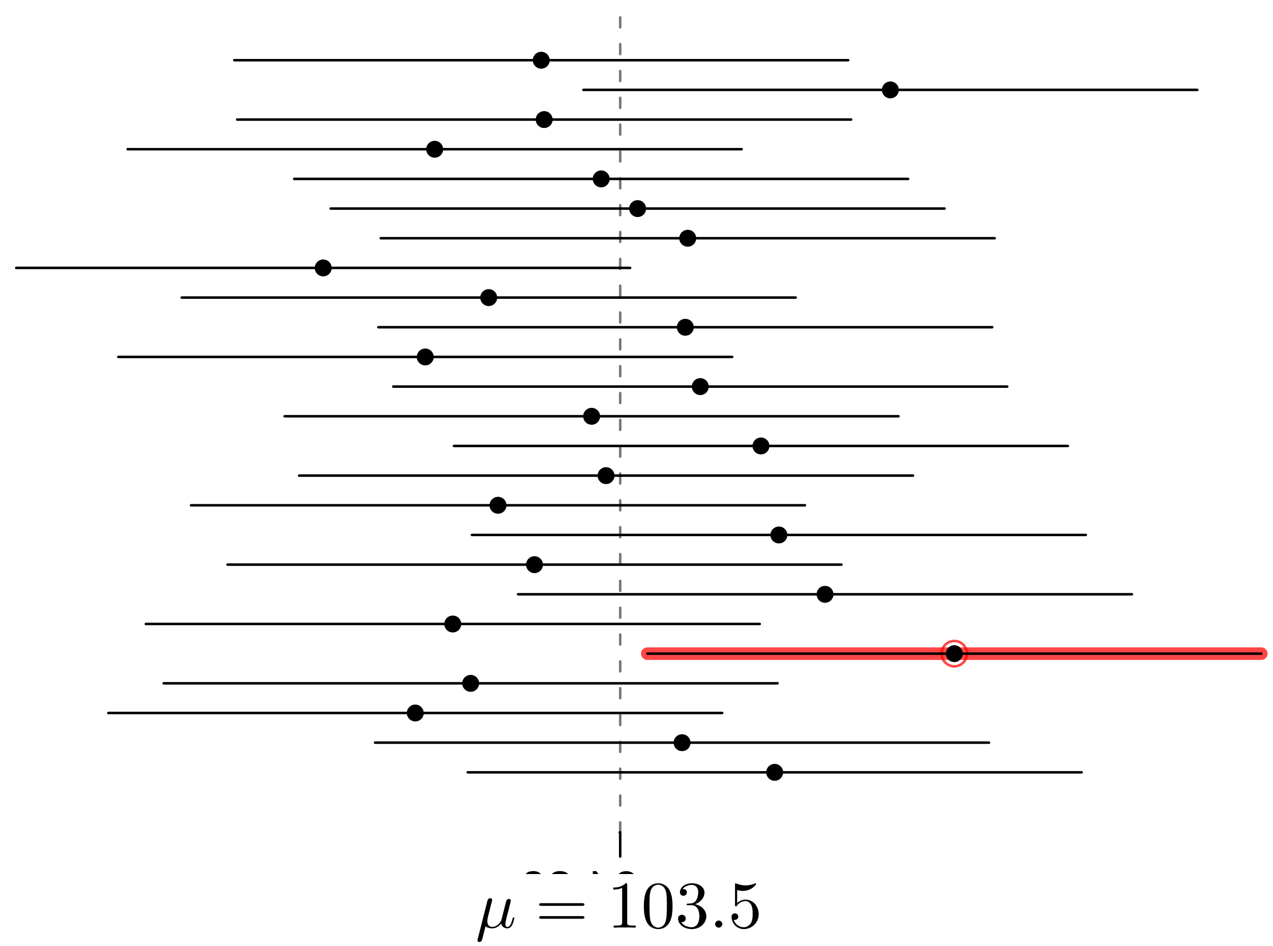
and we fail to reject because 103.5 is in this interval.

Note that the conclusion is different than for the test at significance level 5%. Indeed, we have forced the probability of type 1 error to be smaller, which makes us reject less often.

Interpretation of the significance level

In our example, we have rejected the test at level 5% but not the test at level 1%. What do those 5% and 1% tests mean?

Just like for confidence interval, it has meaning only if the experiment can be repeated:
If we draw 100 samples of size 1000, the 1% test will make in average one error of type 1, whereas the test with level 5% will make around five errors of type 1.



Choice of the null hypothesis

In our example the null hypothesis was the truth (but we were not supposed to know that) so it's obvious that the type I error is the worst.
In practical applications, we do not know what the truth is so we choose H_0 in such a way that the most important error to control is indeed the type I error (this is another guideline that agrees with the previous one. Check why).

As a result, we should choose H_0 when the error made by rejecting is the most delicate.

Choice of the null hypothesis

In our example the null hypothesis was the truth (but we were not supposed to know that) so it's obvious that the type I error is the worst.
In practical applications, we do not know what the truth is so we choose H_0 in such a way that the most important error to control is indeed the type I error (this is another guideline that agrees with the previous one. Check why).

As a result, we should choose H_0 when the error made by rejecting is the most delicate.

Examples:

Choice of the null hypothesis

In our example the null hypothesis was the truth (but we were not supposed to know that) so it's obvious that the type I error is the worst.
In practical applications, we do not know what the truth is so we choose H_0 in such a way that the most important error to control is indeed the type I error (this is another guideline that agrees with the previous one. Check why).

As a result, we should choose H_0 when the error made by rejecting is the most delicate.

Examples:

H_0: innocent
H_A: guilty

Choice of the null hypothesis

In our example the null hypothesis was the truth (but we were not supposed to know that) so it's obvious that the type I error is the worst.
In practical applications, we do not know what the truth is so we choose H_0 in such a way that the most important error to control is indeed the type I error (this is another guideline that agrees with the previous one. Check why).

As a result, we should choose H_0 when the error made by rejecting is the most delicate.

Examples:

H_0: innocent
H_A: guilty

H_0: drug inefficient
H_A: drug effective

Choice of the null hypothesis

In our example the null hypothesis was the truth (but we were not supposed to know that) so it's obvious that the type I error is the worst.
In practical applications, we do not know what the truth is so we choose H_0 in such a way that the most important error to control is indeed the type I error (this is another guideline that agrees with the previous one. Check why).

As a result, we should choose H_0 when the error made by rejecting is the most delicate.

Examples:

H_0: innocent	H_0: drug inefficient	H_0: patient sick
H_A: guilty	H_A: drug effective	H_A: patient healthy

P-values

The testing procedure based on confidence interval is as follows

$\bar{x}\ (= 104.8) \longrightarrow$ Test

$\alpha\ (= 5\%) \longrightarrow$ Test

Test $\longrightarrow$ Decision ("reject")

The decision can only be "reject" or "fail to reject" but we don't know how close we were from taking the other decision For example, in our example, if we perform a test at 5% we reject and if we perform a test at 1%, we fail to reject. This is valuable information.

P-values

The p-value of a test is a number between 0 and 1 such that

p-value < α	reject test at level α
p-value > α	FAIL TO reject test at level α

Gives the answer to tests with any significance level.

$\bar{x}\ (= 104.8)$ → Test → p-value (=3.16%)

P-values

p-value (=3.16%)

If we want to perform a test at 1%, the decision is

If we want to perform a test at 5%, the decision is

If we want to perform a test at 10%, the decision is

Instead of communicating the decision of the test, it is much more informative to record the p-value.

P-values with R

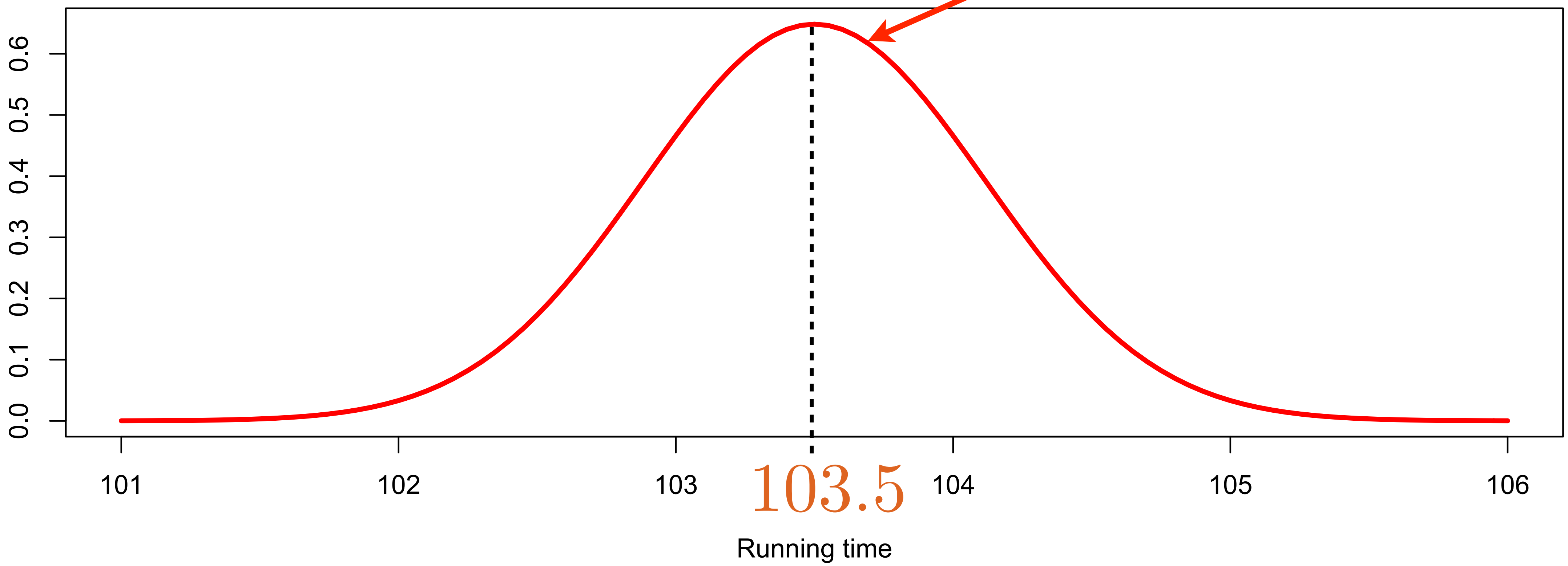
```
x=run10$time[1:1000]
t.test(x, mu=103.5)
```

```
        One Sample t-test

data:  x
t = 2.1528, df = 999, p-value = 0.03157
alternative hypothesis: true mean is not equal to 103.5
95 percent confidence interval:
 103.6172 106.0318
sample estimates:
mean of x
 104.8245
```

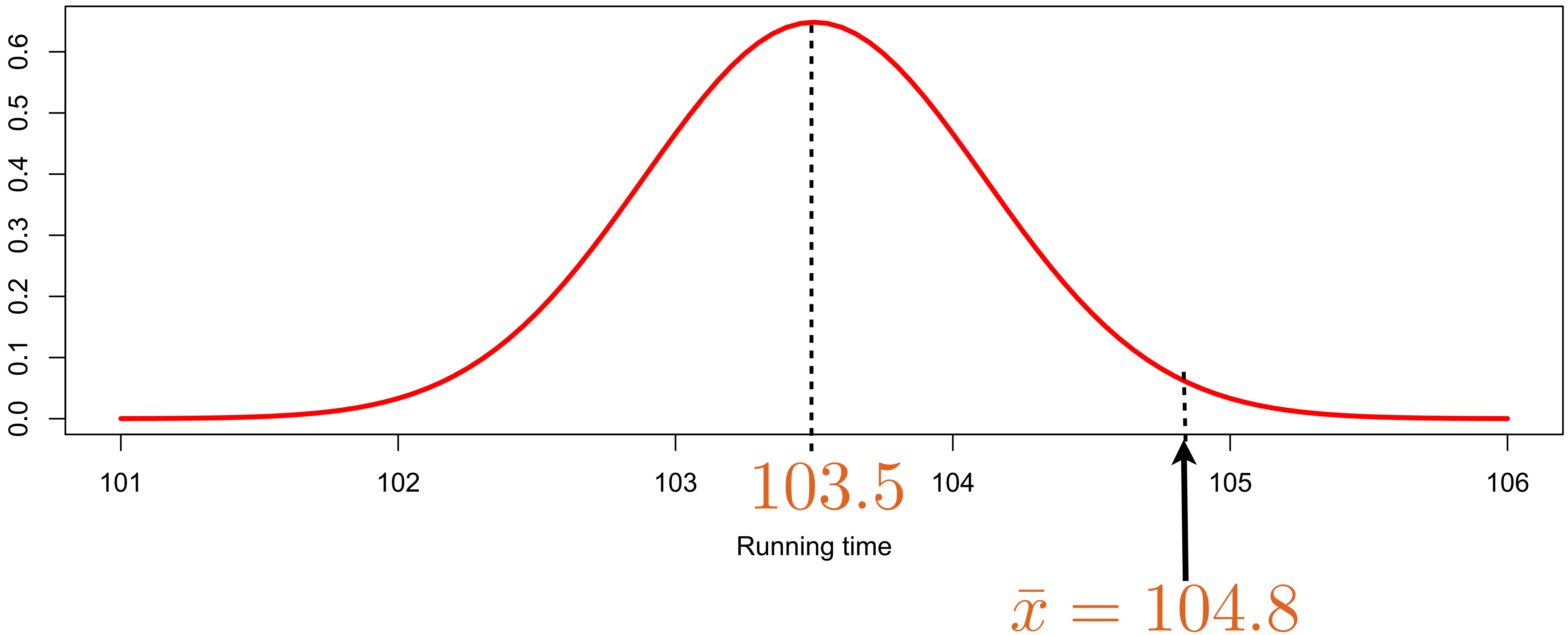
What are P-values?

pdf of
$N(103.5, \frac{373}{1000})$



Recall that $\bar{X} \sim N(\mu, \frac{373}{1000})$ whatever μ is (it is the true, *unknown*, expected value).
If H_0 ($\mu = 103.5$) is true, then $\bar{X} \sim N(103.5, \frac{373.0}{1000})$ and the above function allows us to measure if our **observed value** $\bar{x}$ is likely to happen (if H_0 is true)

What are P-values?



The p-value is the probability of observing data ($\bar{X}$) at least as

favorable to the alternative

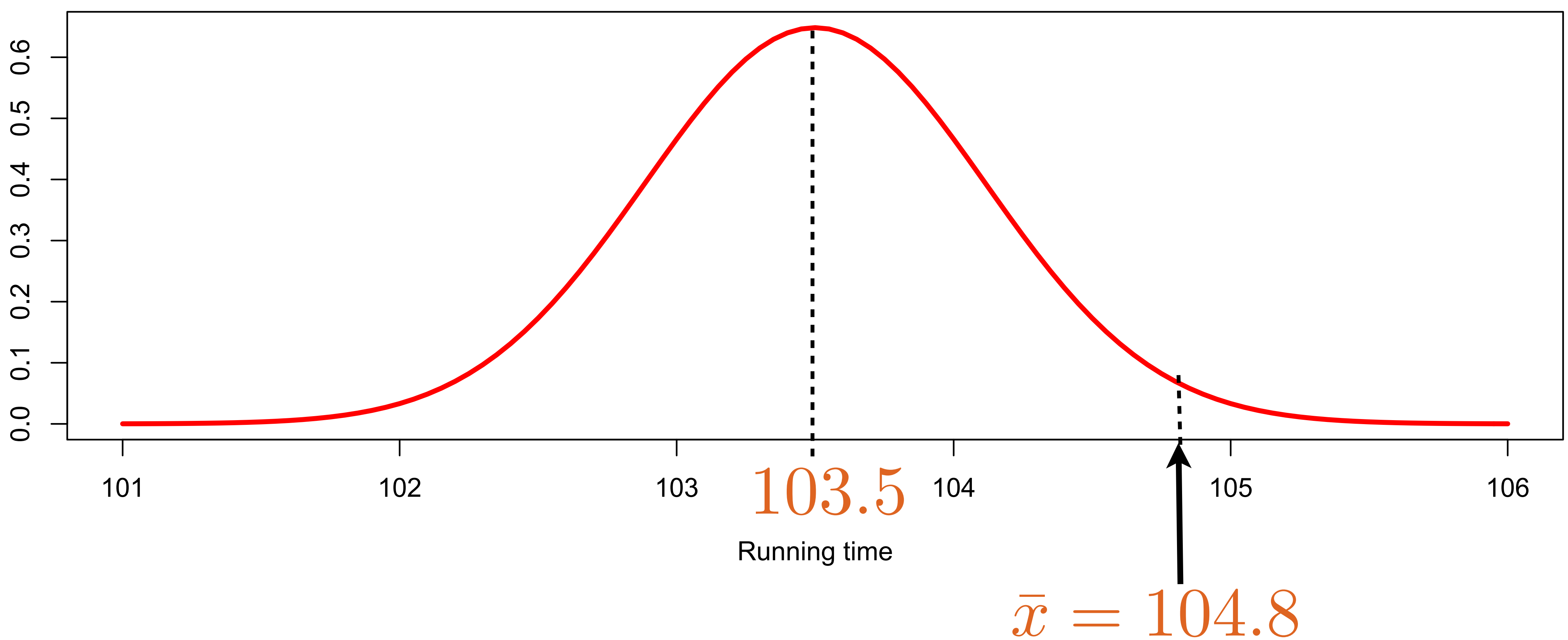
as our current data set ($\bar{x}$) <u>if H_0 was true</u>

What are P-values?

favorable to the alternative

We reject if $|\bar{X} - 103.5|$ is too large (far from 103.5 in any direction). So

$$p - value = P(|\bar{X} - 103.5| > |\bar{x} - 103.5|) = P(|\bar{X} - 103.5| > 1.3)$$



What are P-values?

favorable to the alternative

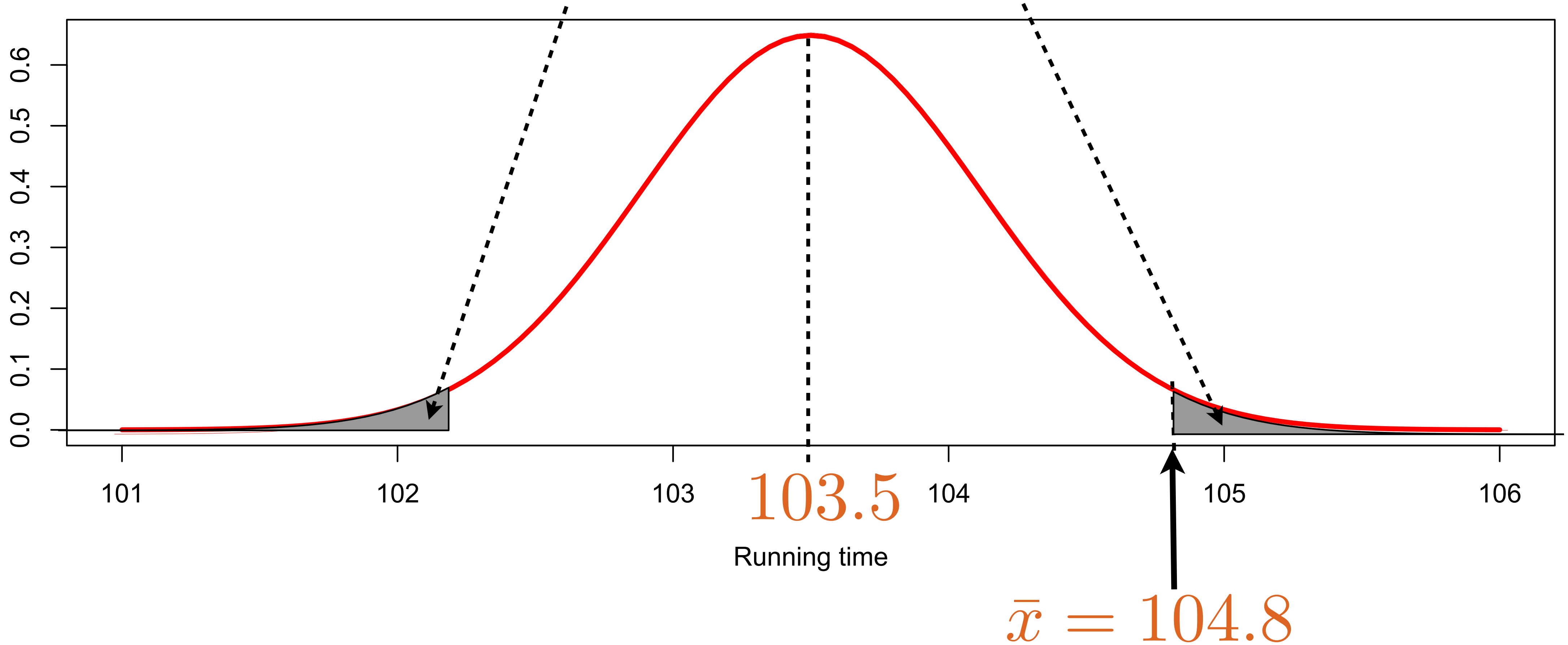
We reject if $|\bar{X} - 103.5|$ is too large (far from 103.5 in any direction). So

$$p - value = P(|\bar{X} - 103.5| > |\bar{x} - 103.5|) = P(|\bar{X} - 103.5| > 1.3)$$

What are P-values?

The p-value is the sum of these two areas

$$p - value = P(|\bar{X} - 103.5| > |\bar{x} - 103.5|) = P(|\bar{X} - 103.5| > 1.3)$$



Computing P-values?

$$p - value = P(|\bar{X} - 103.5| > |\bar{x} - 103.5|) = P(|\bar{X} - 103.5| > 1.3)$$

Using the Z-score, compute the p-value:

$$\begin{aligned} p - value &= P(|\bar{X} - 103.5| > 1.3) \\ &= P(\frac{|\bar{X} - 103.5|}{\sqrt{373.0/1000}} > \frac{1.3}{\sqrt{373/1000}}) \\ &= P(|Z| > \frac{1.3}{\sqrt{373/1000}}) \\ &= 2P(Z < -2.13) \\ &\simeq 2 * 0.0166 = 0.0332 \end{aligned}$$

Corresponds to the R output up to rounding errors

Summary

We have two ways of testing:
1. Using confidence interval
2. Using p-values

When using confidence intervals, we use a fixed level test and the answer is binary: either "reject" or "fail to reject".

When using the p-value, we obtain a number between 0 and 1
The smaller the p-value the less likely is H_0 to be true.

The p-value, is the probability of observing data at least as favorable to the alternative as the current data set. (if small, it means that the current data is already in favor of H_A)

One sided tests

The test that we have considered so far is

$$\boxed{\begin{aligned} H_0 &: \ \mu = 103.5 \\ H_A &: \ \mu \neq 103.5 \end{aligned}}$$

What if we are only interested in discovering whether the true average running time is less than 103.5:

$$\boxed{\begin{aligned} H_0 &: \ \mu \leq 103.5 \\ H_A &: \ \mu > 103.5 \end{aligned}}$$

This is our "toy" example but this happens a lot in reality. Consider the cough syrup example. If the average number of expectorations per hour of sick patient is 10, we want to test

$$\boxed{\begin{aligned} H_0 &: \ \mu \geq 10 \\ H_A &: \ \mu < 10 \end{aligned}}$$

One sided tests

$$\boxed{\begin{aligned} H_0 &: \mu \leq 103.5 \\ H_A &: \mu > 103.5 \end{aligned}}$$

This is called a one-sided test because the alternative is only one side of 103.5.

$$\boxed{\begin{aligned} H_0 &: \mu = 103.5 \\ H_A &: \mu \neq 103.5 \end{aligned}}$$

This is called a two-sided test because the alternative is both sides of 103.5.

$$\boxed{\begin{aligned} H_0 &: \mu \geq 103.5 \\ H_A &: \mu < 103.5 \end{aligned}}$$

This is also a one-sided test

Rule

the strict inequality sign is
always
in the alternative.

P-value for one sided tests

Consider the testing problem

$$\boxed{\begin{aligned} H_0 &: \mu \leq 103.5 \\ H_A &: \mu > 103.5 \end{aligned}}$$

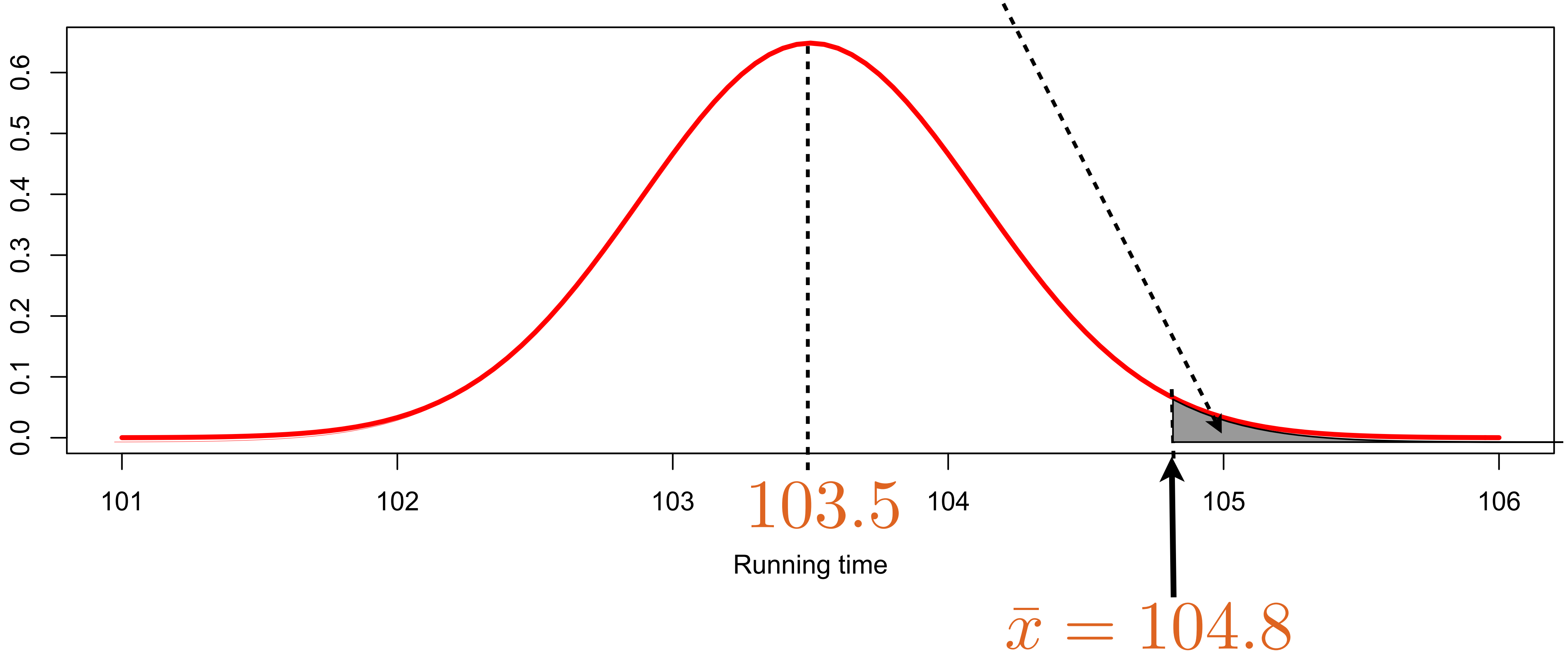
We will reject if $\bar{x} >$ something.

This is enough to compute the p-value:

The p-value, is the probability of observing data at least as favorable to the alternative as the current data set.

P-value for one sided tests

The p-value is this area



P-value for one sided tests

Mathematically, the p-value can be computed as follows:

$$
\begin{aligned}
p - value &= P(\bar{X} > \bar{x}) \\
&= P(\bar{X} > 104.8) \\
&= P\left(\frac{\bar{X} - 103.5}{\sqrt{373.0/1000}} > \frac{104.8 - 103.5}{\sqrt{373.0/1000}}\right) \\
&= P(Z > 2.13) \\
&= 1 - P(Z < 2.13) \\
&= 1 - 0.9834 = 0.0166
\end{aligned}
$$

One sided tests `with R`

```
x=run10$time[1:1000]
t.test(x, mu=103.5, alternative="greater")
```

```
    One Sample t-test

data:  x
t = 2.1528, df = 999, p-value = 0.01579
alternative hypothesis: true mean is greater than 103.5
95 percent confidence interval:
 103.8116      Inf
sample estimates:
mean of x
 104.8245
```