Chapter 3

Distributions of random variables

A dictionary of distributions

In many instances, the same “usual” distributions are encountered in nature.

This may come from the design of the random experiment

A dictionary of distributions

In many instances, the same "usual" distributions are encountered in nature.

This may come from the design of the random experiment

X=number of sick patients among n patients
X=number of ORFE majors among n students
X=number of heads in n coin tosses
X=number of spades in a hand of n cards
.....

All those have the **same distribution** only the parameter change from one to another.

A dictionary of distributions

In many instances, the same "usual" distributions are encountered in nature.

This may come from the design of the random experiment

X=number of sick patients among n patients
X=number of ORFE majors among n students
X=number of heads in n coin tosses
X=number of spades in a hand of n cards
.....

→ All those have the **same distribution** only the parameter change from one to another.

For continuous distributions, the ubiquitous distribution is

A dictionary of distributions

In many instances, the same "usual" distributions are encountered in nature.

This may come from the design of the random experiment

X=number of sick patients among n patients
X=number of ORFE majors among n students
X=number of heads in n coin tosses
X=number of spades in a hand of n cards
.....

→ All those have the **same distribution** only the parameter change from one to another.

For continuous distributions, the ubiquitous distribution is

The normal distribution

Bernoulli distribution

Is a drug effective on a randomly selected patient?
Does a randomly selected tire satisfy quality requirements?
Is a randomly selected student a senior?
Is a randomly selected census tract on the Charles river?
Is a randomly selected employee a woman?
Is the outcome of a coin toss "heads"?

These random experiment are very simple since they have two outcomes. We can code numerically them by 0 (for failure="no") and 1(for success="yes") to obtain a random variable X.
In this case we say that "X has Bernoulli distribution with parameter p" and write $X \sim Bernoulli(p)$ where p is the probability of success:

$$P(X = 1) = p$$

Bernoulli distribution

Since there are only two possible outcomes, we have:

$$P(X=0) = 1 - p \;= q$$

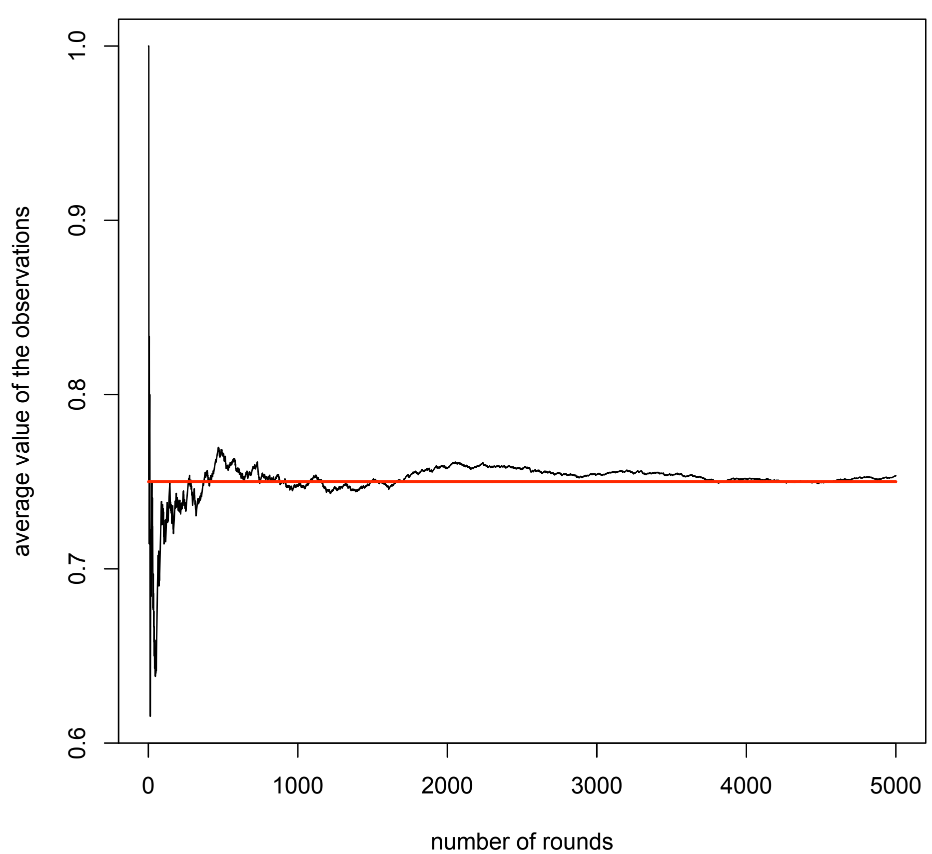
Interesting quantities are

$$E(X) = p \qquad\qquad var(X) = p(1-p) = pq$$

Bernoulli distribution

What is the advantage of using the numerical codes 0 and 1 over 3 for success and 62 for failure for example?

The average of observations is the proportion of success which tends to p as the number of repetition goes to infinity:



$$\frac{1+1+1+1+0+1+0+1}{8} = \frac{6}{6} = 0.75$$

number of success $= \sum_{i=1}^{n} X_i$

Binomial distribution

number of success = $\sum_{i=1}^{n} X_i$

Binomial distribution

number of success = $\sum_{i=1}^{n} X_i$ This is also a random variable

Binomial distribution

number of success = $\sum_{i=1}^{n} X_i$ This is also a random variable

Takes vales 0, 1, 2, ..., n and called binomial distribution with parameters n and p

$$X \sim Bin(n,p)$$

The probability distribution is given by

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 1, 2, \ldots, n$$

$$E(X) = np \qquad var(X) = np(1-p)$$

Binomial distribution

The binomial distribution is often encountered in practice:

X=number of heads in 10 coin tosses

X=number of patients carrying H1N1 in a cohort of 100 randomly selected patients (prevalence is 6%)

X=numbers of computers infected by a virus with prevalence 56% in a company with 157 workstations

Binomial distribution

The binomial distribution is often encountered in practice:

X=number of heads in 10 coin tosses

$$X \sim Bin(10, .5)$$

X=number of patients carrying H1N1 in a cohort of 100 randomly selected patients (prevalence is 6%)

X=numbers of computers infected by a virus with prevalence 56% in a company with 157 workstations

Binomial distribution

The binomial distribution is often encountered in practice:

X=number of heads in 10 coin tosses

$$X \sim Bin(10, .5)$$

X=number of patients carrying H1N1 in a cohort of 100 randomly selected patients (prevalence is 6%)

$$X \sim Bin(100, .06)$$

X=numbers of computers infected by a virus with prevalence 56% in a company with 157 workstations

Binomial distribution

The binomial distribution is often encountered in practice:

X=number of heads in 10 coin tosses

$$X \sim Bin(10, .5)$$

X=number of patients carrying H1N1 in a cohort of 100 randomly selected patients (prevalence is 6%)

$$X \sim Bin(100, .06)$$

X=numbers of computers infected by a virus with prevalence 56% in a company with 157 workstations

$$X \sim Bin(157, .56)$$

Binomial distribution

Is it binomial? (read more in Section 3.4 of OpenIntro)

1. The trials are independent
2. The number of trials n is fixed
3. Each trial outcome is either success or failure
4. The probability of success p is **the same** for each trial

Poisson distribution

- A random variable X that has Poisson distribution can take any nonnegative integer value 0,1,2,....
- It can be seen as a generalization of the binomial value when the number of trials is very large $(n \to \infty)$
- At the same time the probability of success has to become very small $(p \to 0)$ otherwise, the expected value of X would be infinite.

- A random variable X has Poisson distribution $X \sim Poiss(\lambda)$ with parameter $\lambda > 0$ if $X \sim Bin(n, p)$ with

$$n \to \infty, \ p \to 0, \ np \to \lambda > 0$$

Poisson distribution

From the results of the Binomial distribution we can find the expected value and variance of $X \sim Poiss(\lambda)$
Indeed,

$$np \to \lambda \qquad\qquad np(1-p) \to \lambda(1-0) = \lambda$$

So if $X \sim Poiss(\lambda)$ then

$$E(X) = \lambda \qquad\qquad var(X) = \lambda$$

Poisson distribution

Why not use the binomial distribution directly?
If $n = 10^6, p = 10^{-6}$ it is difficult to compute P(X=x).

Every day, insurance companies use the poisson distribution.

Consider a company that subscribes to an insurance policy for its 1,000 employees. The insurance company's investigation reveal that the probability that a an employee is injured in a given year is 0.02. If we assume that injuries are independent of each other then the total number of injuries X per year is $X \sim Bin(1000, 0.02)$ which can be approximated by $X \sim Poiss(20)$. In average: 20 injuries, and we also know the standard deviation $\sqrt{20} \simeq 4.47$

Uniform distribution

The uniform distribution is the simplest continuous distribution.

Given an interval [a, b] the distribution puts the same weight on each region of the interval with a constant probability density function:

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

$$E(X) = \frac{a+b}{2}$$

Exercise: check that $\int_{-\infty}^{\infty} p(x)dx = 1$

$$var(X) = \frac{(b-a)^2}{12}$$

Uniform distribution

In this case, we say that the random variable "X has uniform distribution on the interval [a,b]" and we write

$$X \sim U([a,b])$$

Exercise: check that
$$\int_{-\infty}^{\infty} p(x)dx = 1$$

This is a calculus exercise.

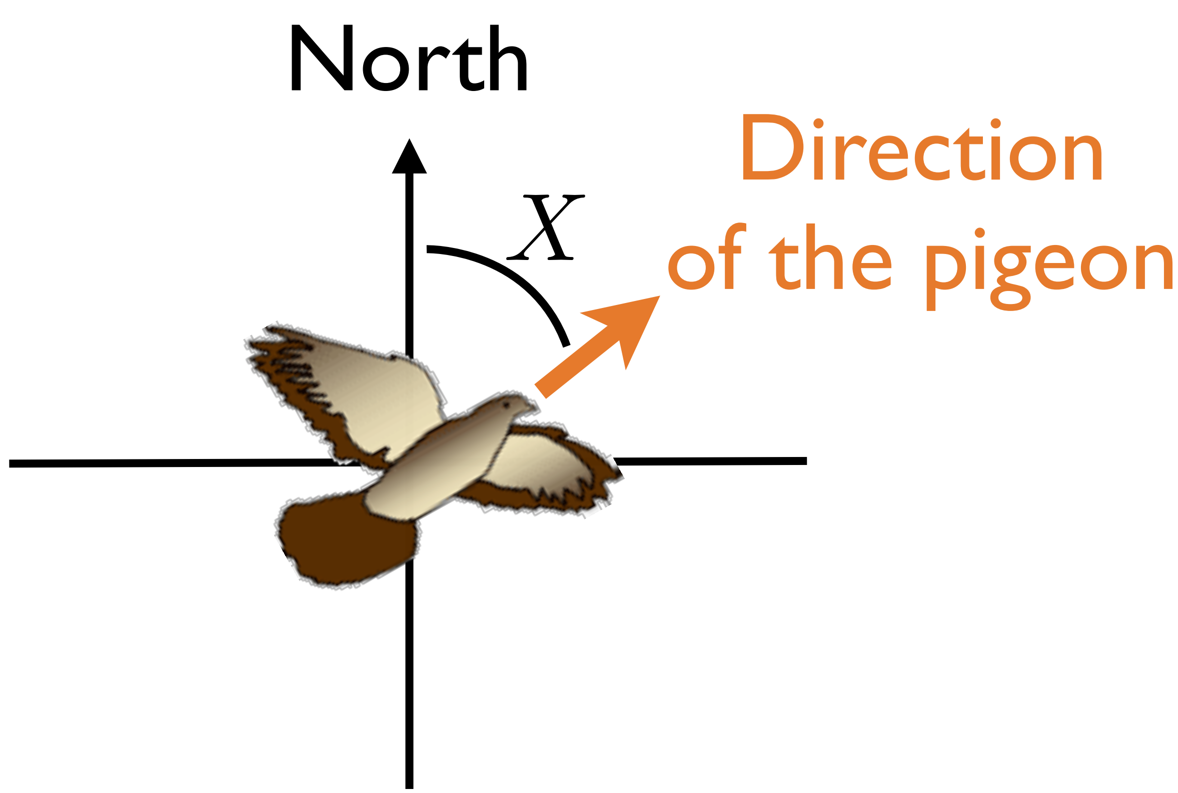
$$E(X) = \frac{a+b}{2}$$

$$var(X) = \frac{(b-a)^2}{12}$$

Uniform distribution

An example is the following:

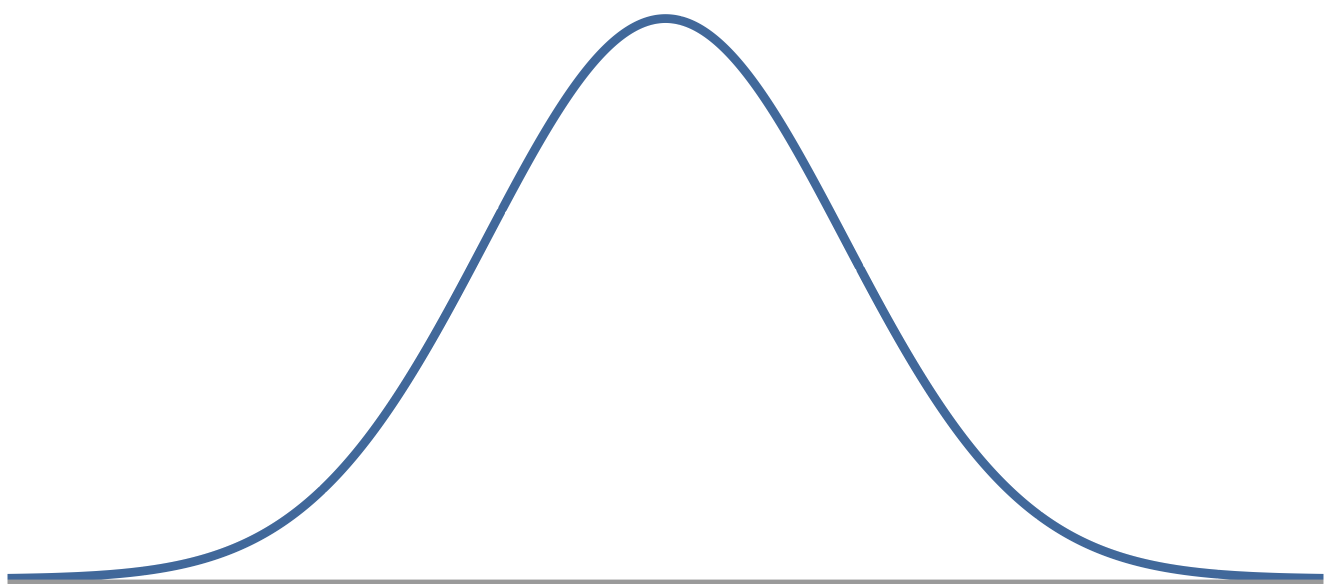
A pigeon is released from a rooftop and it can take any direction randomly. Let X denote the azimuth (in degrees) of the direction take by the pigeon:



Then $X \sim U([0, 360])$

Normal distribution

You probably know its **bell-shaped** curve



It is symmetric, unimodal, smooth, going to 0 fast in the tails

The curve is obtained by plotting the probability density function (pdf) of a "**normal random variable**" (a random variable that has normal distribution).

a.k.a Gaussian distribution

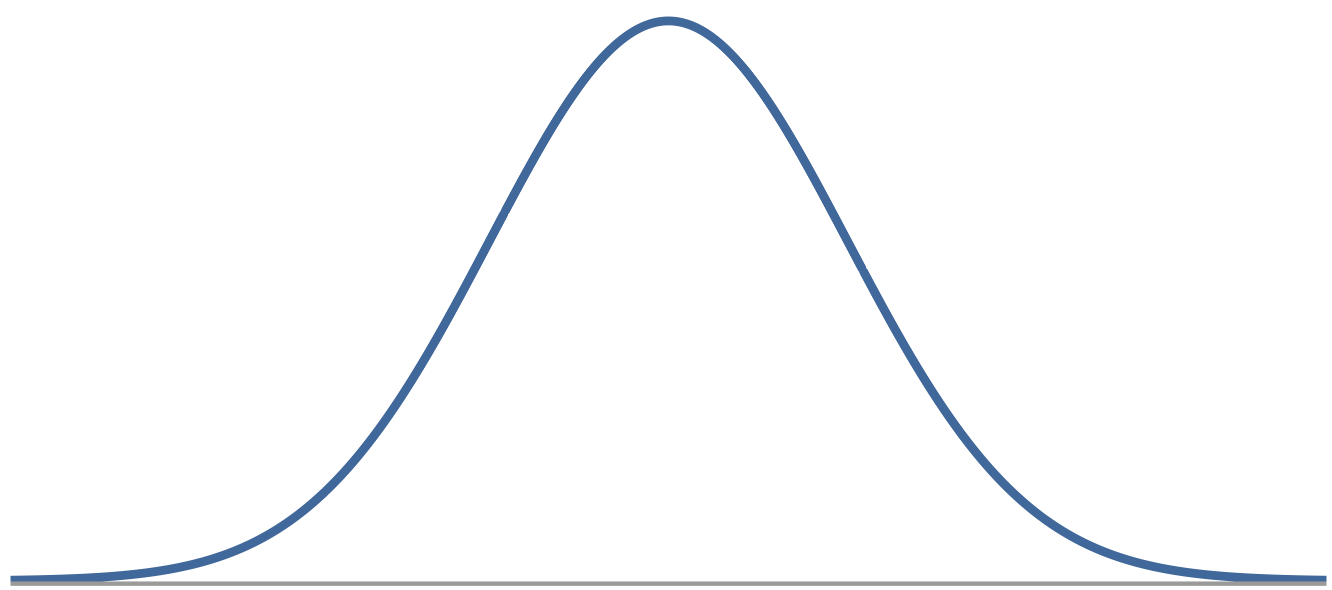
It is also commonly called the Gaussian distribution after German mathematician Carl Friedrich Gauss who demonstrated that the method of "least-squares" (see chapter 7) was valid when the observations were drawn from this distribution.

Since then, it has been shown that this distribution is much more than a simple tool. It is **everywhere**! Social sciences, chemistry, physics, electronics, medicine, ...

Normal distribution model

The bell-shaped curve is the plot of the pdf of a random variable. The pdf is the following function:

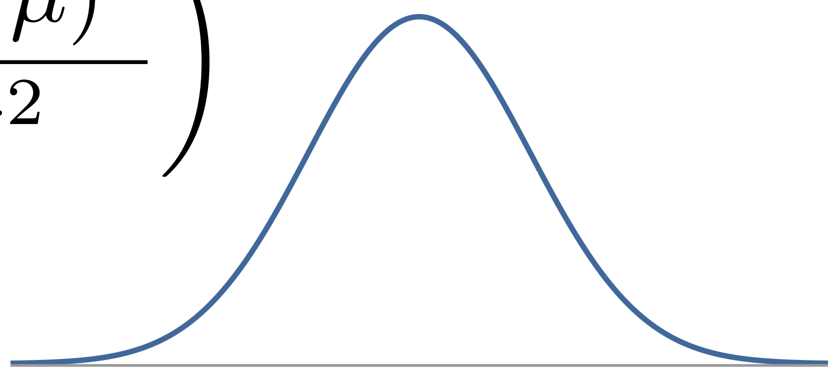
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Normal distribution model

The bell-shaped curve is the plot of the pdf of a random variable. The pdf is the following function:

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



Depends on two parameters: μ and σ^2

Calculus shows that if X has pdf given as above then

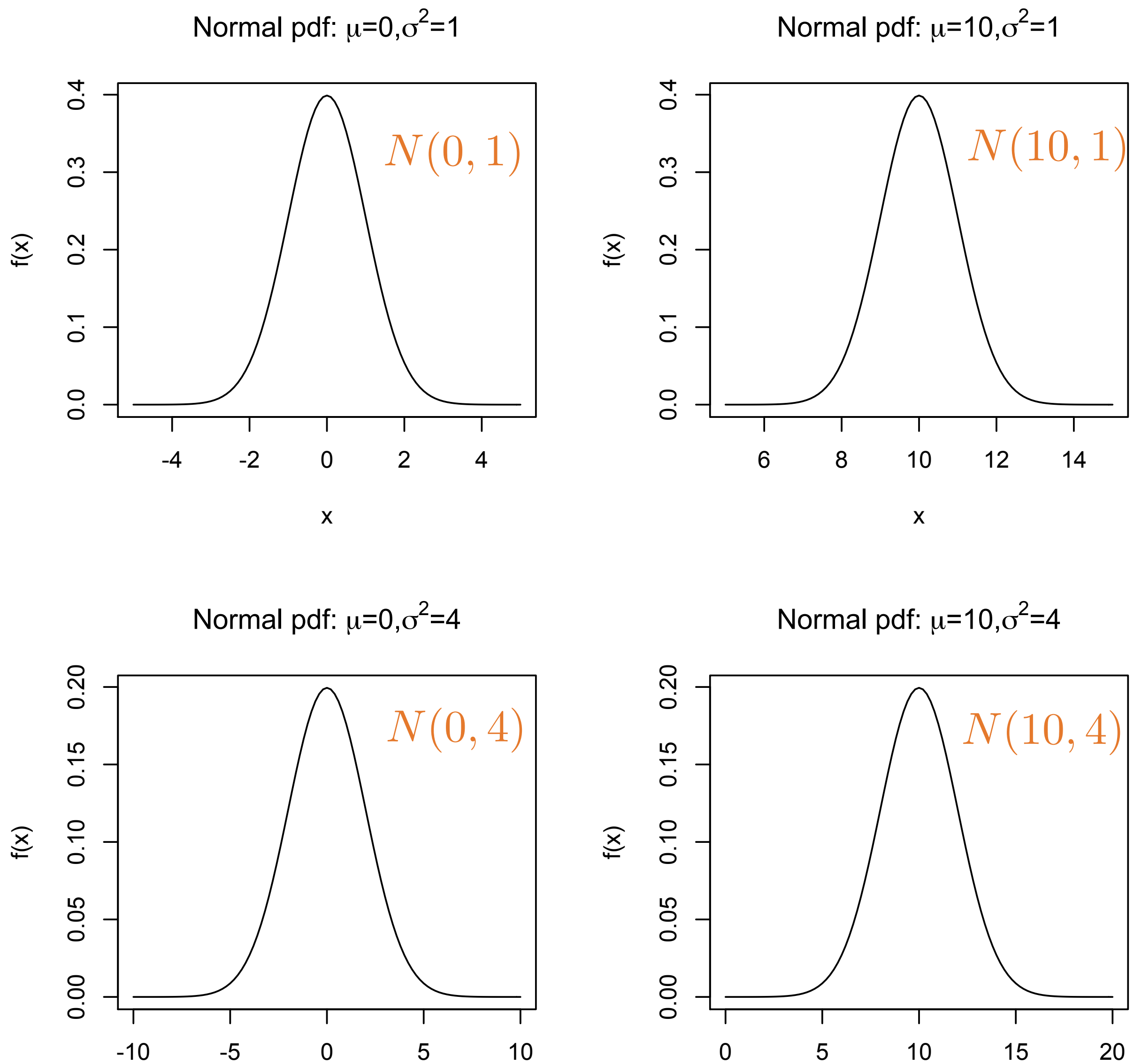
$$E(X) = \mu \qquad var(X) = \sigma^2$$

If X is a normal random variable with parameters μ and σ^2 we write $X \sim N(\mu, \sigma^2)$

or we say that "the distribution of X is $N(\mu, \sigma^2)$"

Effect of μ and σ^2

We know that the expected value measures the average location and the variance measures the average variability around this location.

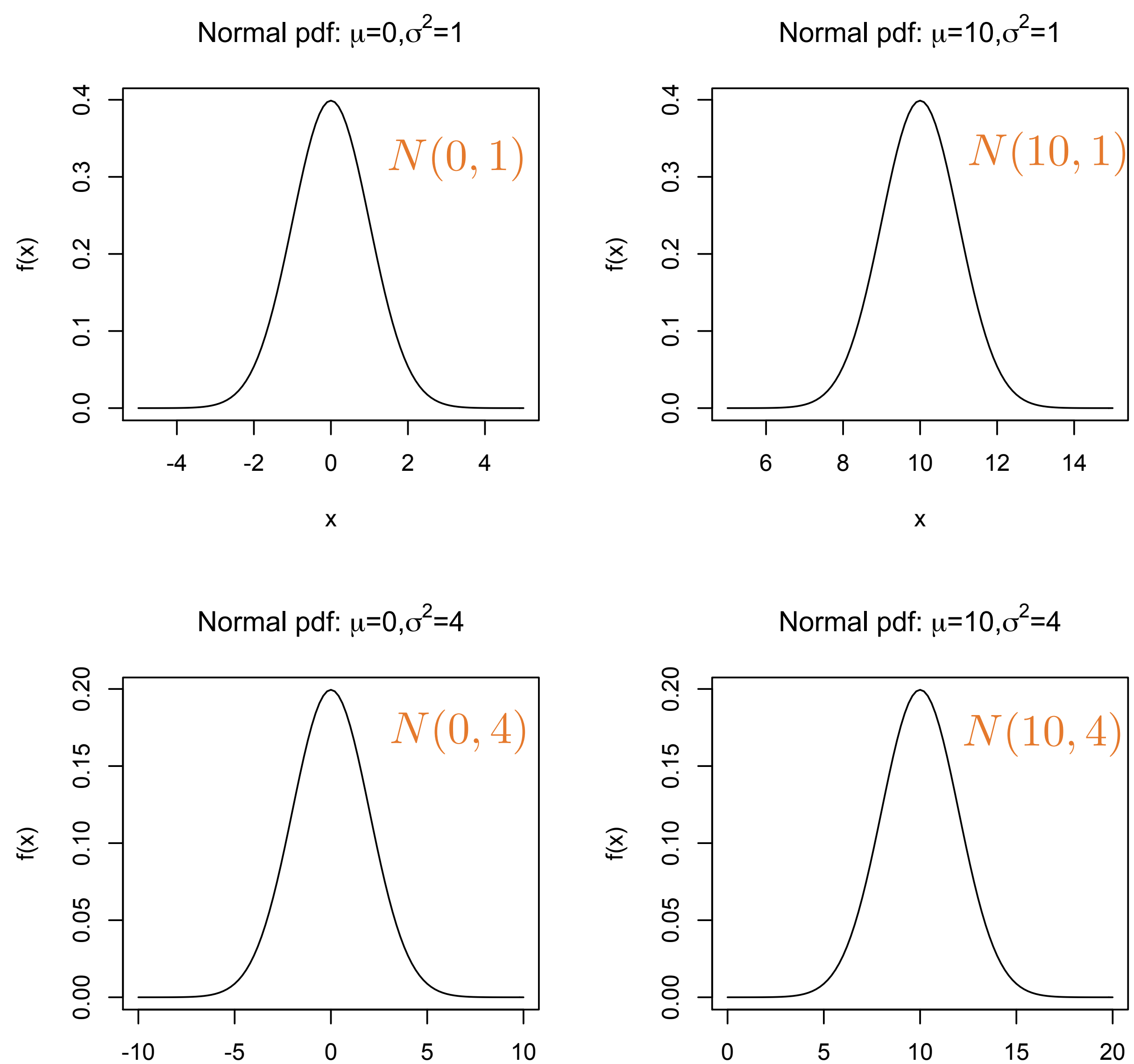



```
par(mfrow=c(2,2))
plot(function(x) dnorm(x, mean=0, sd=1), -5, 5, main = expression(paste("Normal pdf: ",symbol("m"),"=",0, ",",sigma^2,"=",1)),ylab="f(x)")
plot(function(x) dnorm(x, mean=10, sd=1), -5, 5, main = expression(paste("Normal pdf: ",symbol("m"),"=",10, ",",sigma^2,"=",1)),ylab="f(x)")
plot(function(x) dnorm(x, mean=0, sd=2), -5, 5, main = expression(paste("Normal pdf: ",symbol("m"),"=",0, ",",sigma^2,"=",4)),ylab="f(x)")
plot(function(x) dnorm(x, mean=10, sd=2), -5, 5, main = expression(paste("Normal pdf: ",symbol("m"),"=",10, ",",sigma^2,"=",4)),ylab="f(x)")
```

Effect of μ and σ^2

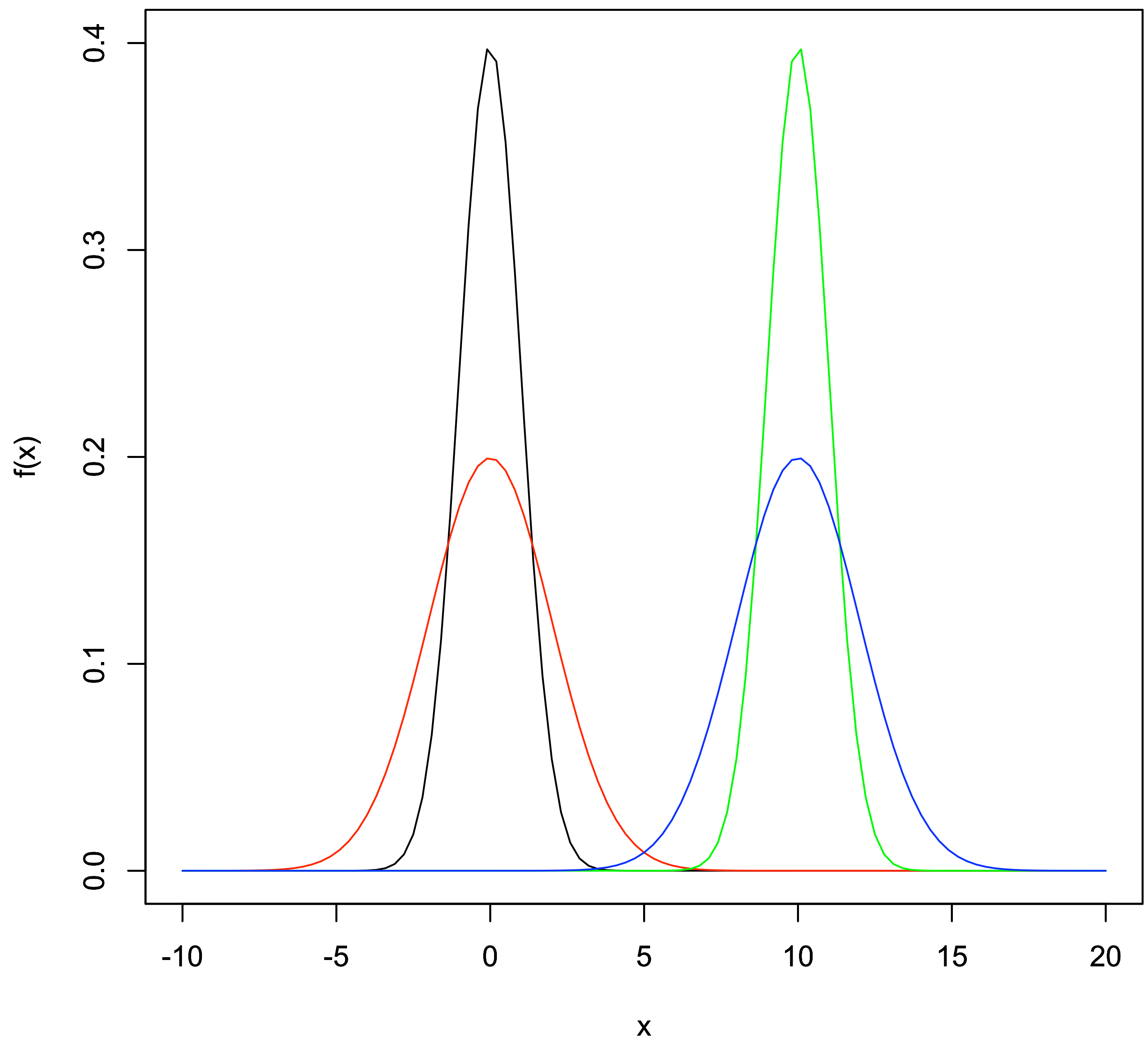
We know that the expected value measures the average location and the variance measures the average variability around this location.

Notice that the scale of the x-axis changes




```
par(mfrow=c(2,2))
plot(function(x) dnorm(x, mean=0, sd=1), -5, 5, main = expression(paste("Normal pdf: ",symbol("m"),"=",0, ",",sigma^2,"=",1)),ylab="f(x)")
plot(function(x) dnorm(x, mean=10, sd=1), -5, 5, main = expression(paste("Normal pdf: ",symbol("m"),"=",10, ",",sigma^2,"=",1)),ylab="f(x)")
plot(function(x) dnorm(x, mean=0, sd=2), -5, 5, main = expression(paste("Normal pdf: ",symbol("m"),"=",0, ",",sigma^2,"=",4)),ylab="f(x)")
plot(function(x) dnorm(x, mean=10, sd=2), -5, 5, main = expression(paste("Normal pdf: ",symbol("m"),"=",10, ",",sigma^2,"=",4)),ylab="f(x)")
```

Effect of μ and σ^2




```
plot(function(x) dnorm(x, mean=0, sd=1), -10, 20, ylim=c(0,.4), ylab="f(x)")
par(new=TRUE)
plot(function(x) dnorm(x, mean=0, sd=2), -10, 20, ylim=c(0,.4), ylab="f(x)", col='red')
par(new=TRUE)
plot(function(x) dnorm(x, mean=10, sd=1), -10, 20, ylim=c(0,.4), ylab="f(x)", col='green')
par(new=TRUE)
plot(function(x) dnorm(x, mean=10, sd=2), -10, 20, ylim=c(0,.4), ylab="f(x)", col='blue')
```

Standardization

There is a **canonical** choice for the parameters μ and σ^2

$$\begin{cases} \mu = 0 \\ \sigma^2 = 1 \end{cases}$$

The distribution $N(0,1)$ is called

standard normal distribution

Starting from a random variable $X \sim N(\mu, \sigma^2)$ we can always transform it into a **standard** normal random variable:

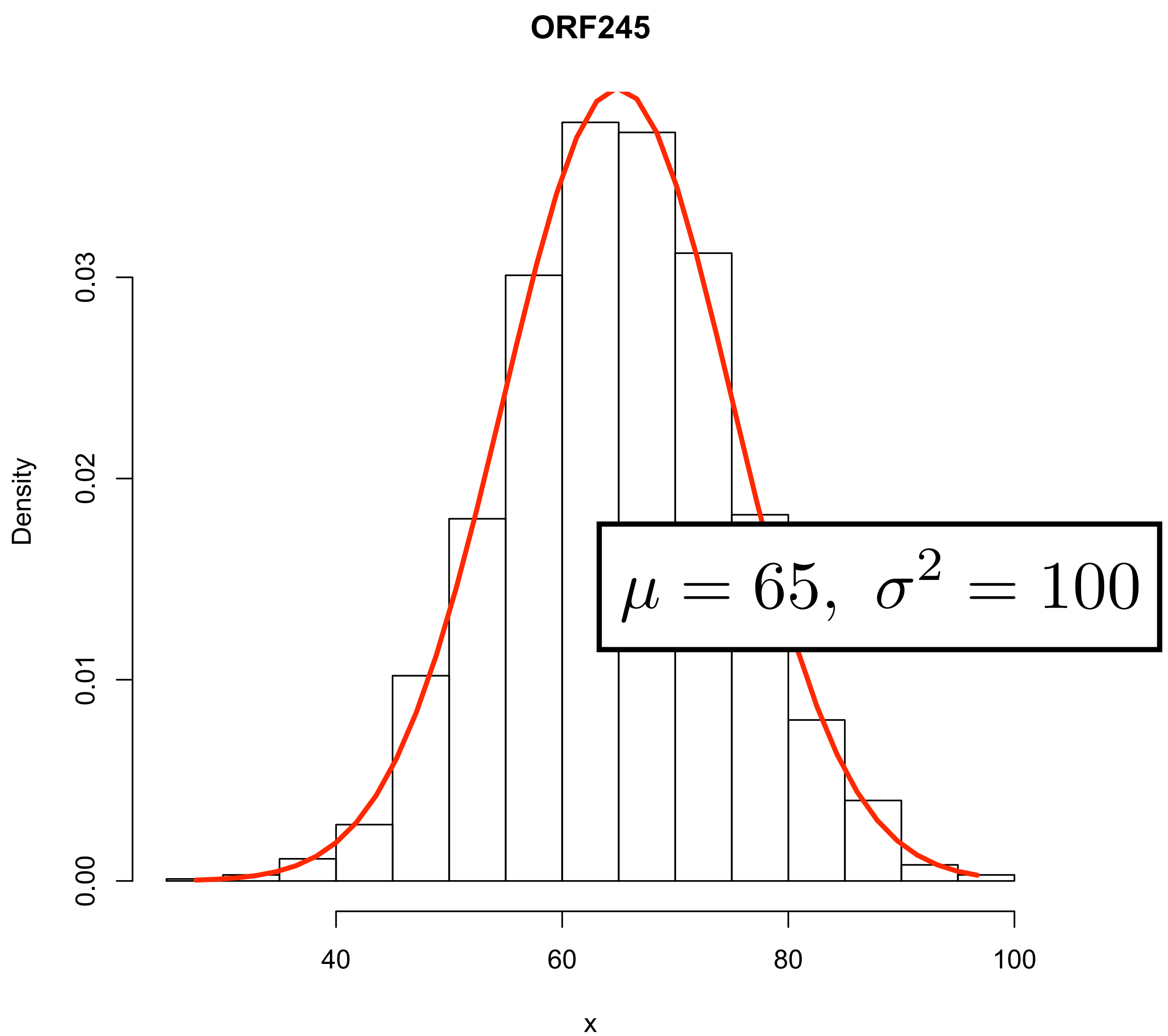
$$Z = \frac{X - \mu}{\sigma}$$

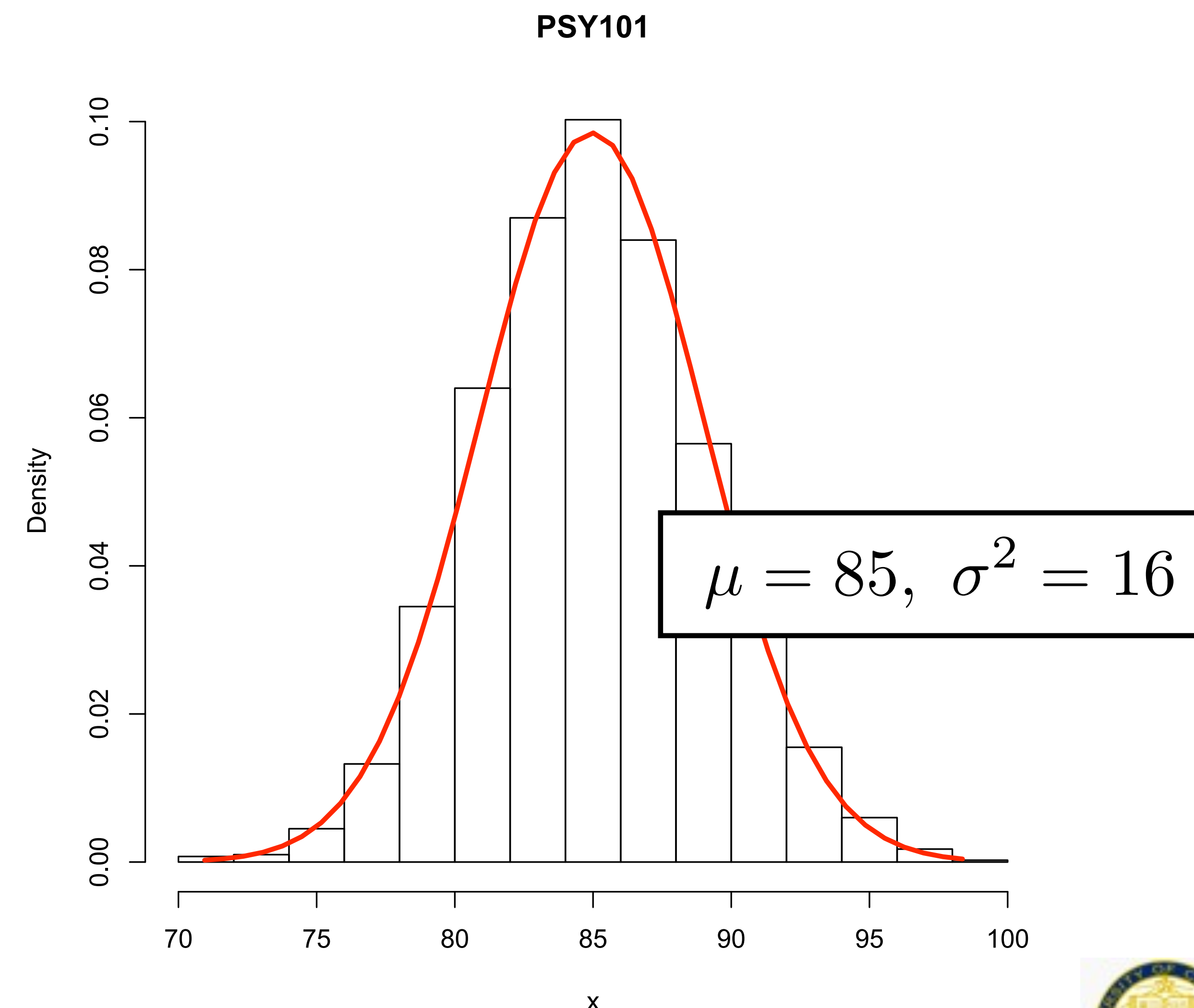
In this case, we have $Z \sim N(0,1)$

Z score

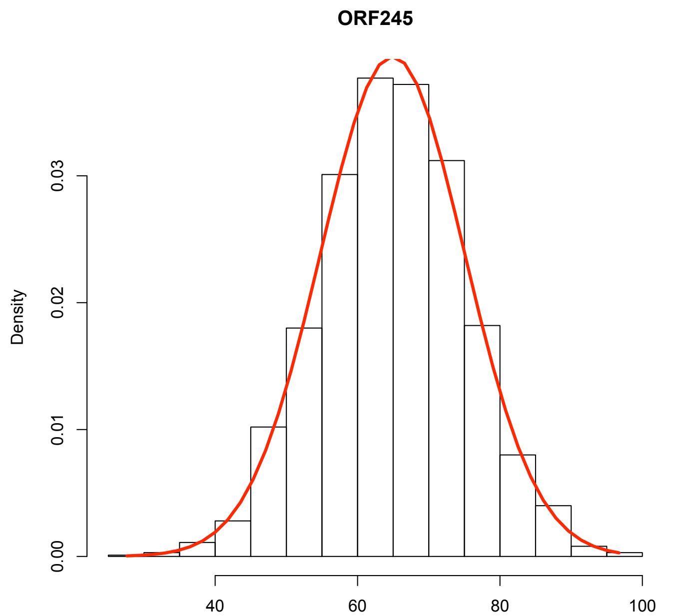
Consider two courses: ORF245 and PSY101

The histogram of historical grades (before curving) in these courses is respectively

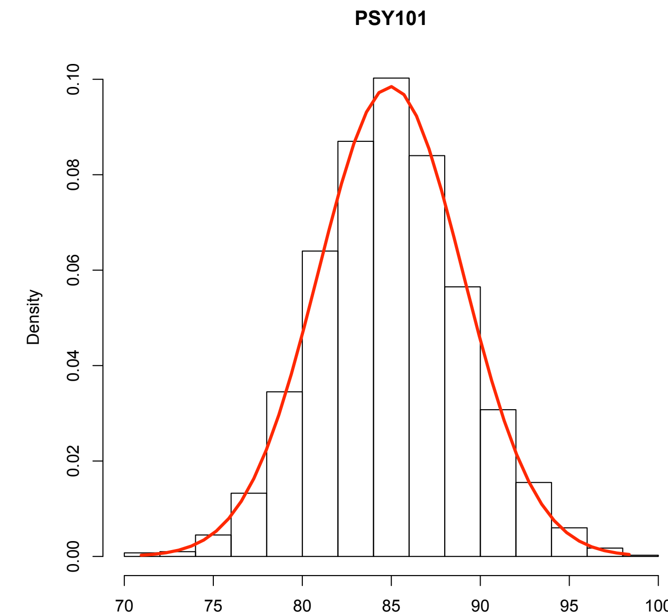




Z score



$$\mu = 65, \ \sigma^2 = 100$$



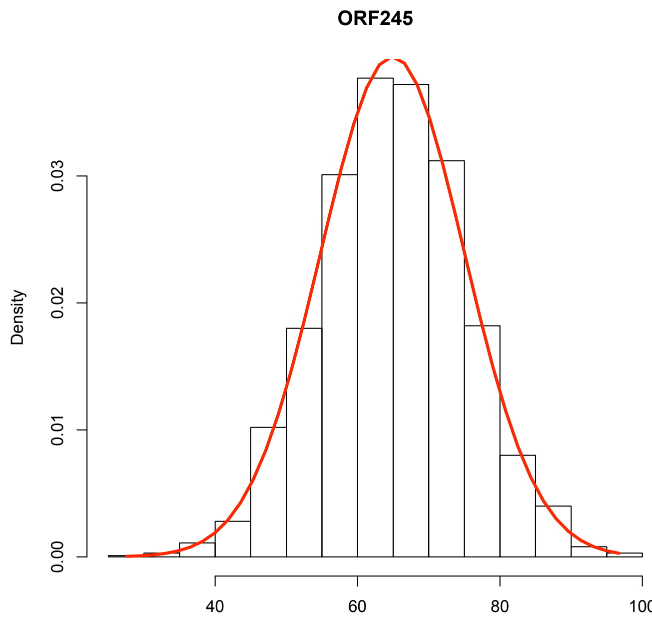
$$\mu = 85, \ \sigma^2 = 16$$

Will, a freshman enrolled in both courses last semester got 75 in MTH183 and 89 in PSY101.

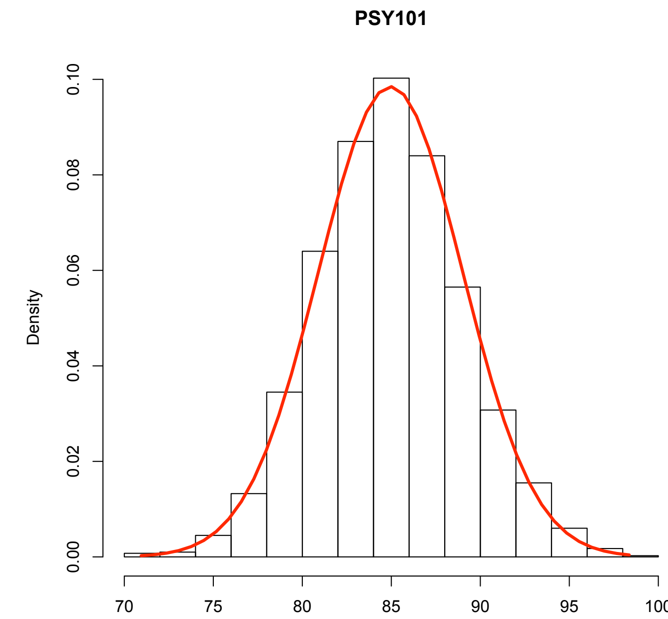
In which course did he do best?

We compute the **Z score** to put both courses on the same scale.

Z score



$$\mu = 65, \ \sigma^2 = 100$$



$$\mu = 85, \ \sigma^2 = 16$$

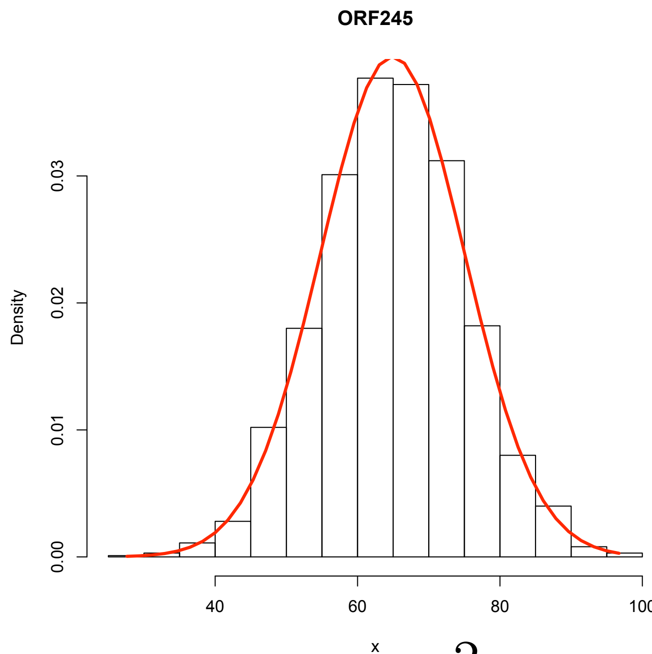
Will, a freshman enrolled in both courses last semester got 75 in MTH183 and 89 in PSY101.

In which course did he do best?

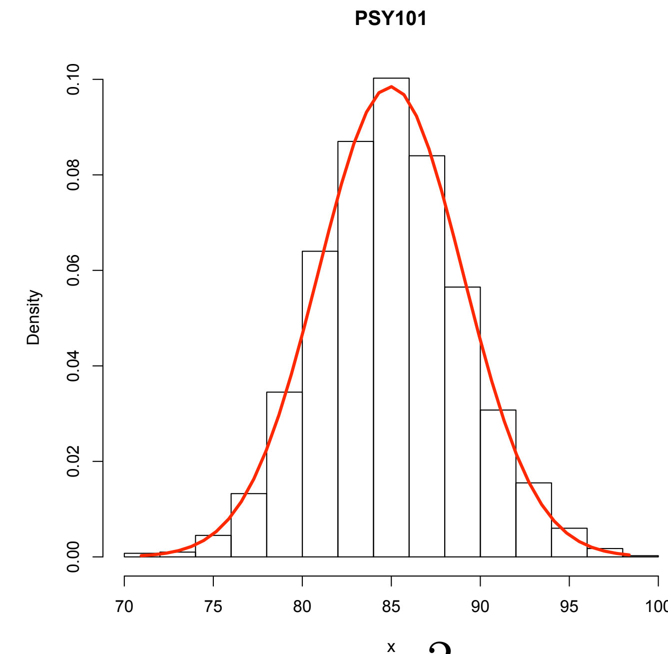
We compute the **Z score** to put both courses on the same scale.

$$Z_{\mathrm{ORF245}} = \frac{75 - 65}{10} = 1$$

Z score



$$\mu = 65, \ \sigma^2 = 100$$



$$\mu = 85, \ \sigma^2 = 16$$

Will, a freshman enrolled in both courses last semester got 75 in MTH183 and 89 in PSY101.
In which course did he do best?
We compute the **Z score** to put both courses on the same scale.

$$Z_{\mathrm{ORF245}} = \frac{75 - 65}{10} = 1 \qquad Z_{\mathrm{PSY101}} = \frac{89 - 85}{4} = 1$$

Z score

In the previous example, Will was better than average in both courses, resulting in positive Z scores in both courses.

If the sign does not matter, we can use the absolute value $|Z|$ to identify unusual observations.

The sales X (in \$) of a hotdog stand on a random day of the summer has distribution $X \sim N(540, 8100)$

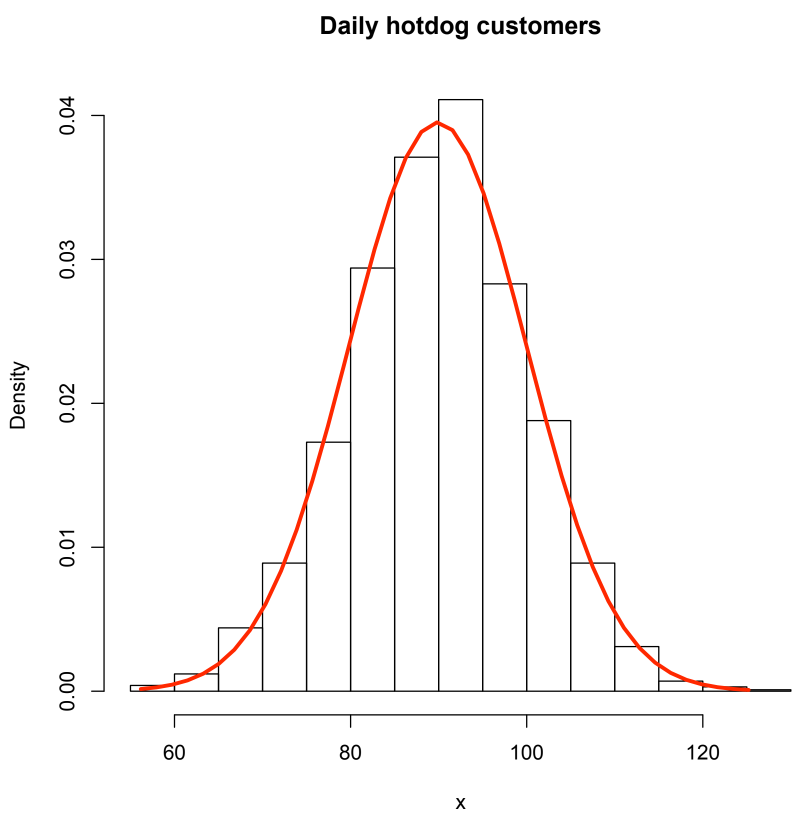
Which amount is the most unusual: 700\$ or 460\$?

$$\left|\frac{700-540}{90}\right| = 1.78 \qquad \left|\frac{460-540}{90}\right| = 0.89$$

Z score

$$\left|\frac{700-540}{90}\right| = 1.78 \quad > \quad \left|\frac{460-540}{90}\right| = 0.89$$

The standard deviation 90 is the same for both scores so we do not need it to compare both numbers.
Assume now that the number of customers Y has (approximate) distribution: $Y \sim N(90, 100)$



Which amount is the most unusual: 700$ or 80 customers

$$\left|\frac{80-90}{10}\right| = 1 < 1.78$$

Standard normal distribution

We can now answer the "relative" question:

which observation is the most unusual?

We want to answer the "absolute" question:

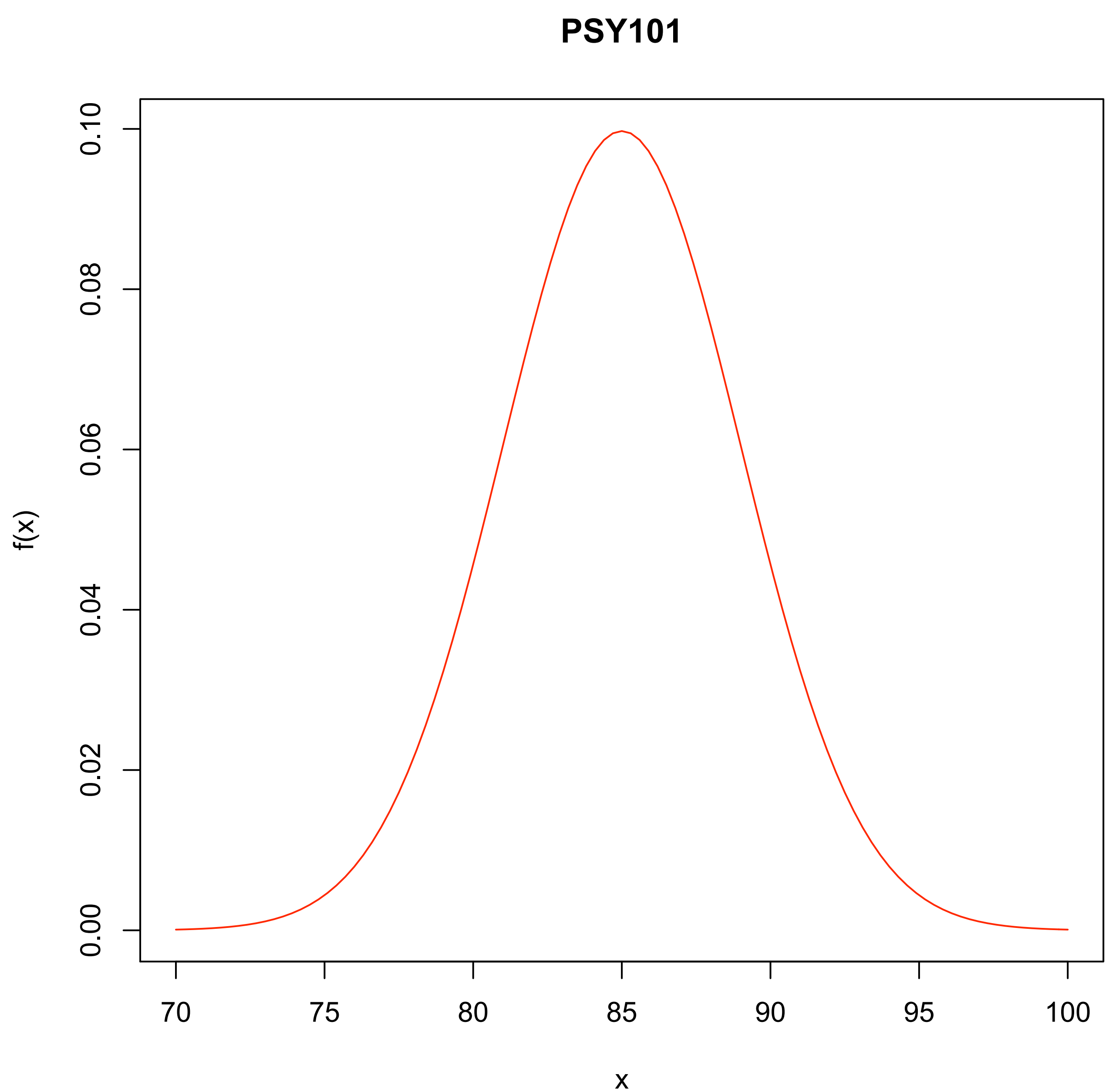
Is 700$ sales unusual?

For that we need to understand better the standard normal distribution and its deviations to its mean ($\mu = 0$)

One way to answer this question is:
If I draw 1000 times a standard normal random variable, how often will it have absolute value greater than 1.78.
In the next slides, we will see that the answer is 7.5% of the time. It is only slightly unusual (slightly unusual: <10%, unusual: <5%, very unusual <1%).

Normal probability table

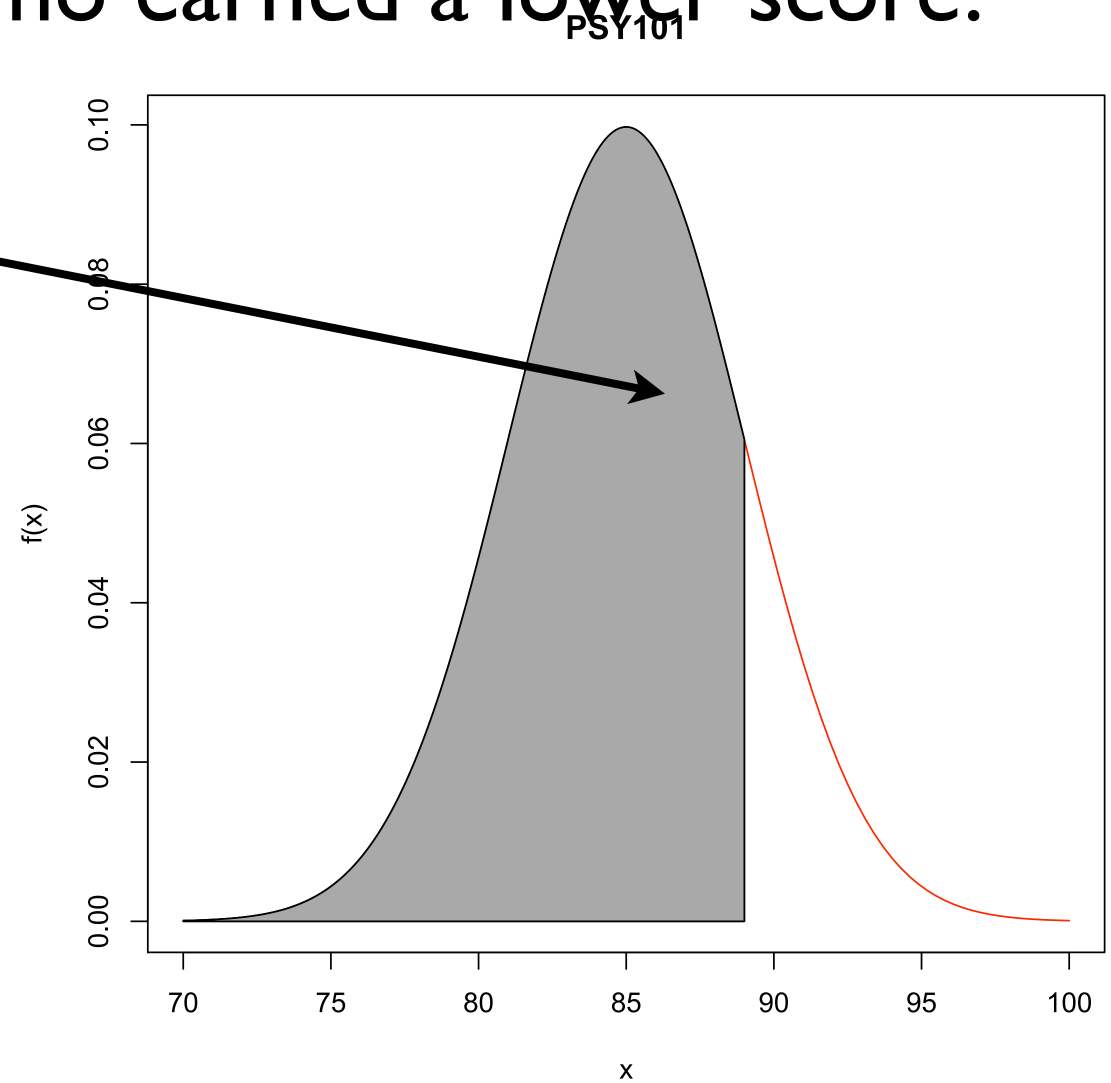
Will scored 85 in PSY101. His percentile is the percentage of students who earned a lower score.



Normal probability table

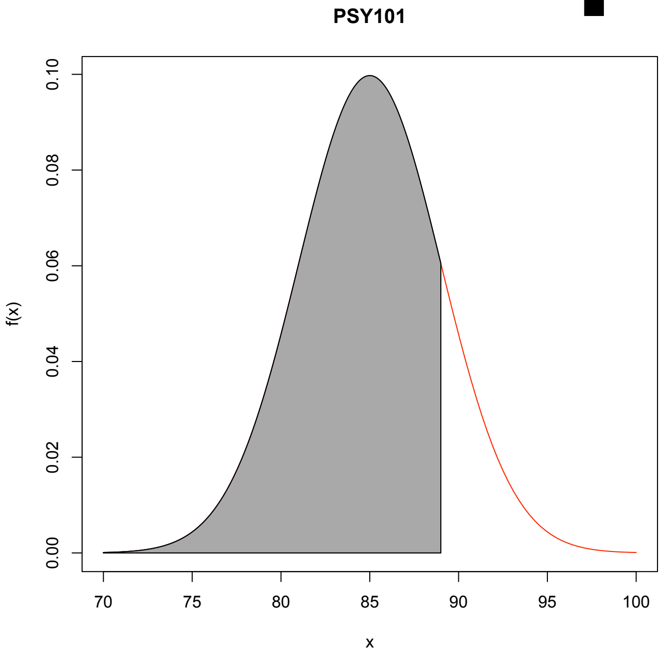
Will scored 85 in PSY101. His percentile is the percentage of students who earned a lower score.

Area under the curve?




```
plot(function(x) dnorm(x, mean=85, sd=4), 70, 100, ylab="f(x)", col=2, main="PSY101")
xx=seq(70, 89, by=0.01; yy=dnorm(xx, mean=85, sd=4)
xx=c(xx, 89, 70); yy=c(yy, 0, 0)
polygon(xx,yy, col='grey60')
```

Normal probability table



Area under the curve? $$\int_{-\infty}^{89} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

Cannot be computed explicitly. We can use either a calculator or a table that gives:

$$\int_{-\infty}^{t} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

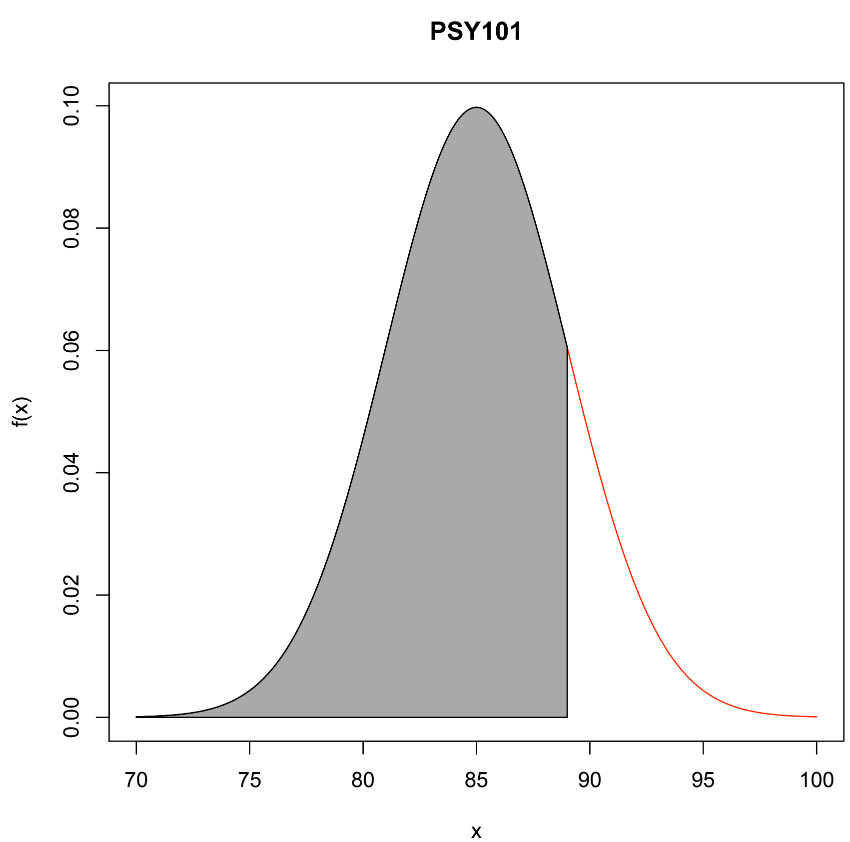
for many values of t (here we are interested in $t = 89$).
Problem: the integral depends on μ and σ^2 an we cannot make one table for each (μ, σ^2) ⇨ too many values!!

Normal probability table

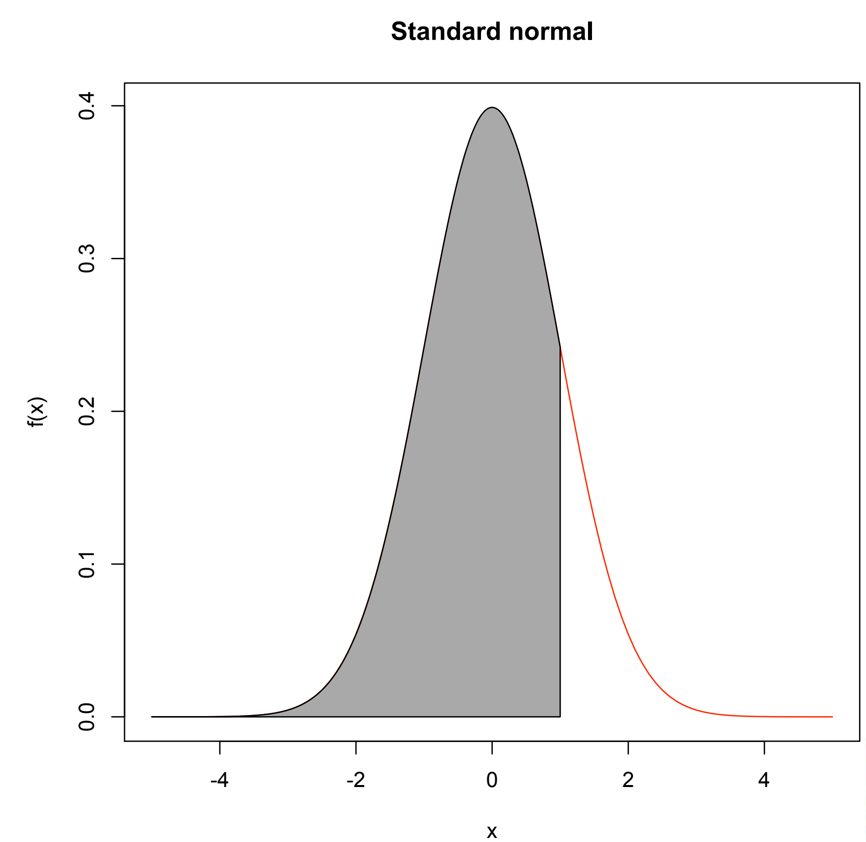
The solution is to standardize:

$$P(X \leq 89) = P\left(\frac{X-85}{4} \leq \frac{89-85}{4}\right) = P(Z \leq 1)$$

where $Z \sim N(0,1)$ is a standard normal random variable



Same shaded area

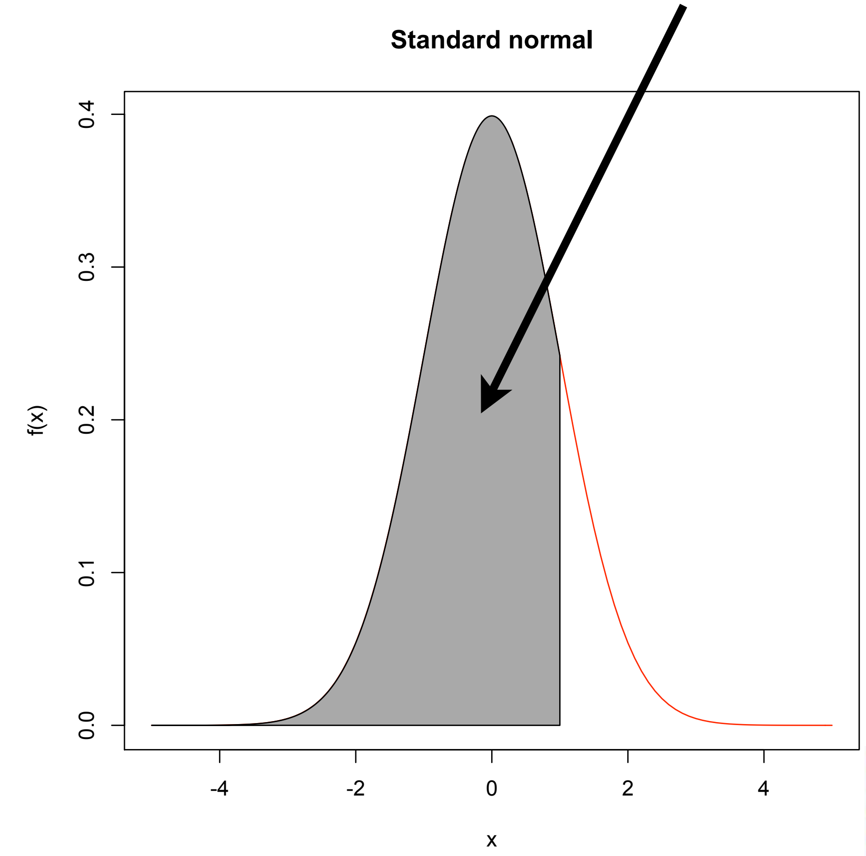


Normal probability table

Therefore, we only need a table for the **standard** normal random variable, that is for the values of the integral

$$\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

Area= $\int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$



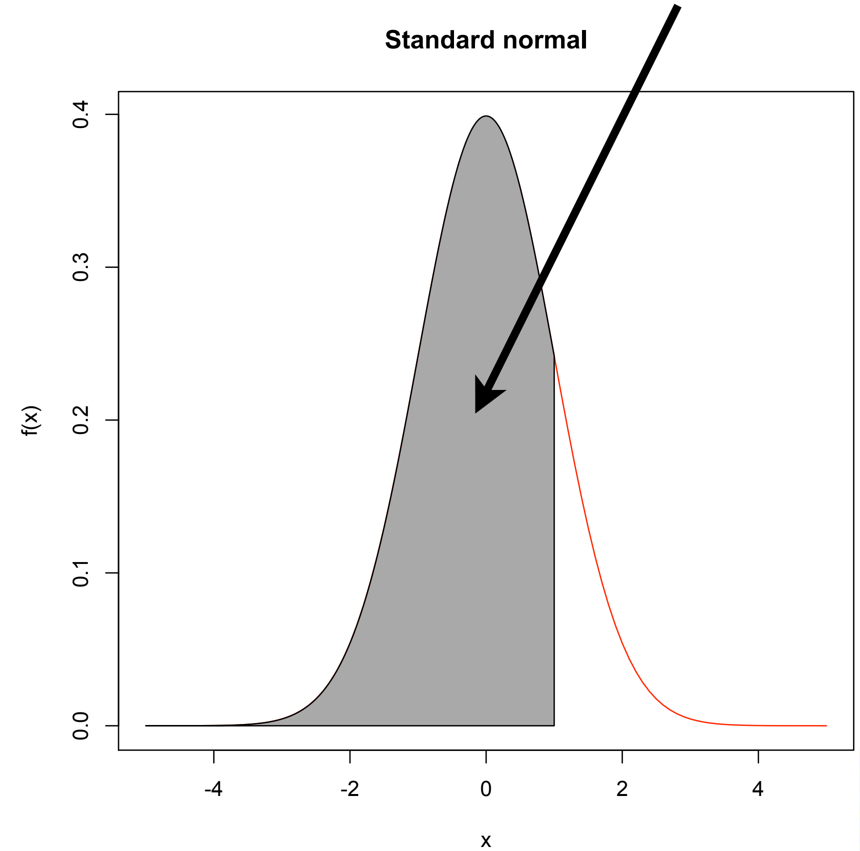
Normal probability table

Therefore, we only need a table for the **standard** normal random variable, that is for the values of the integral

$$\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

Area= $\int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$

These values are given by the **normal probability table** at the end of the book.



Normal probability table

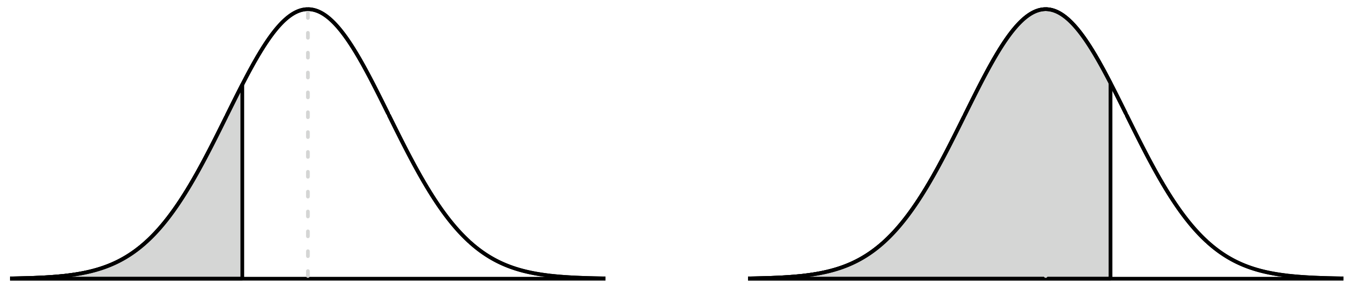


Figure 3.7: The area to the left of Z represents the percentile of the observation.

	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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.3	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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

$P(Z \leq 0.84) =$ $\qquad$ $P(Z \leq 1) =$ $\qquad$ $P(Z \leq 0.43) =$

Normal probability examples

Sarah is a randomly selected student in PSY101. Her score is the random variable $X \sim N(85, 16)$

1. What is the probability that she scores at least 90?
2. The "B-range" is between 80 and 90. What is the probability that earns a B?
3. Eventually, she's at the 85th percentile. What is her score?

Normal probability examples

Sarah is a randomly selected student in PSY101. Her score is the random variable $X \sim N(85, 16)$

1. What is the probability that she scores at least 90?
2. The "B-range" is between 80 and 90. What is the probability that earns a B?
3. Eventually, she's at the 85th percentile. What is her score?

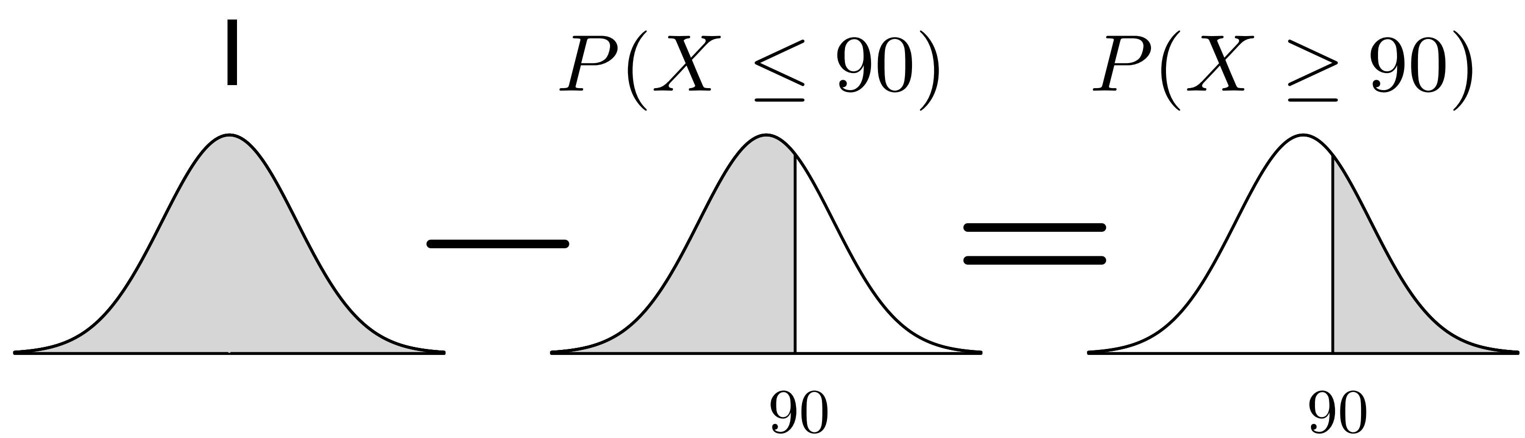
Formally, the questions are:

1. What is $P(X \geq 90)$
2. What is $P(80 \leq X \leq 90)$
3. What is x such that $P(X \leq x) = .85$

Normal probability examples

$$P(X \geq 90) \quad = 1 - P(X \leq 90)$$

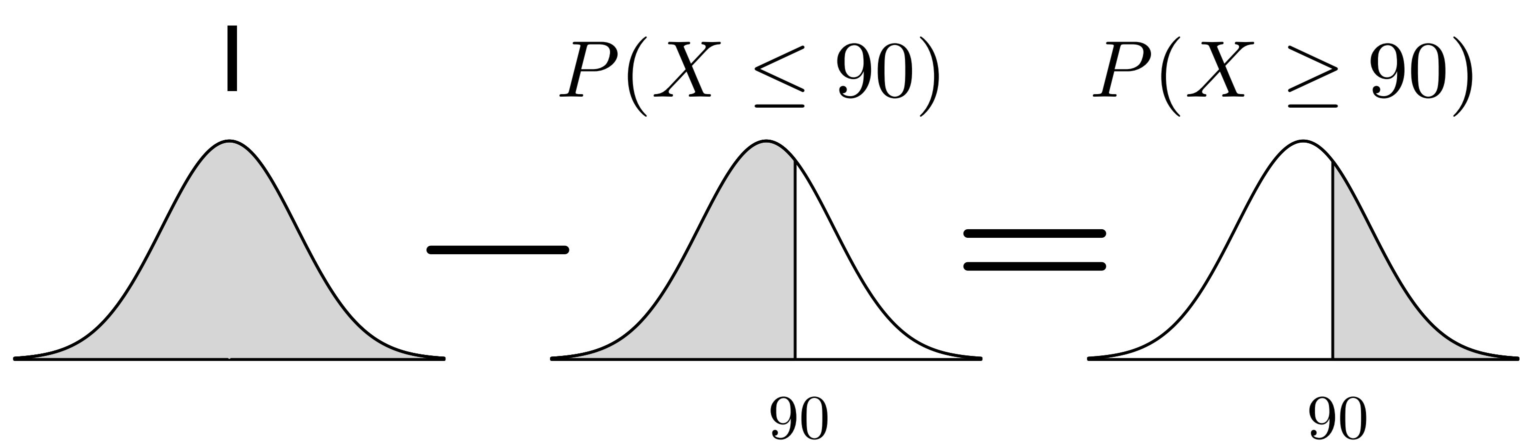
curve of X



Normal probability examples

$$\begin{aligned} P(X \geq 90) &= 1 - P(X \leq 90) \\ &= 1 - P\left(\frac{X - 85}{4} \leq \frac{90 - 85}{4}\right) \\ &= 1 - P(Z \leq 1.25) \end{aligned}$$

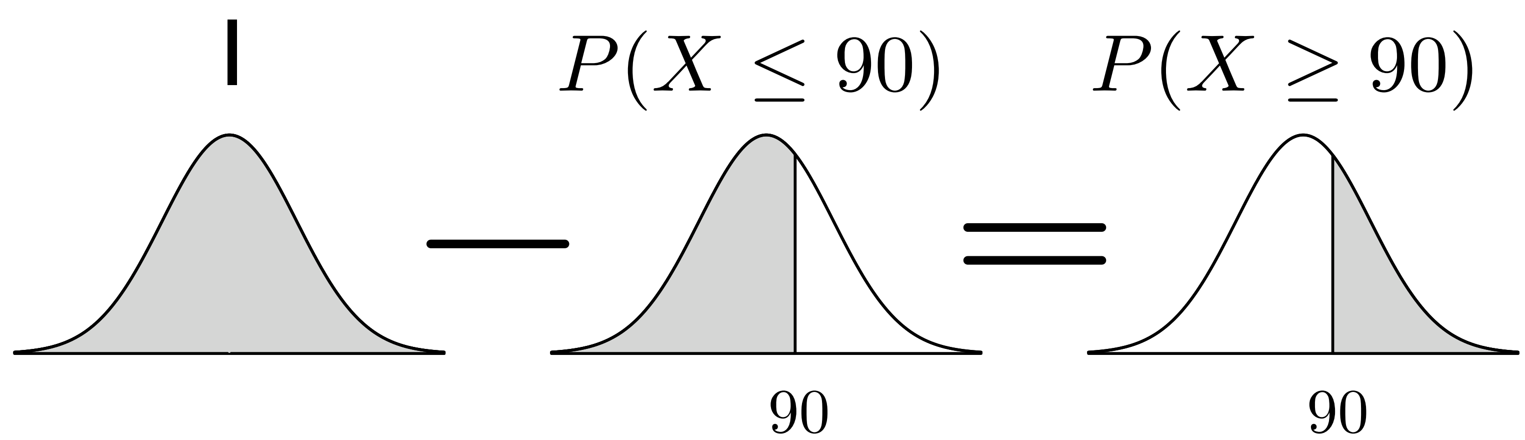
curve of X



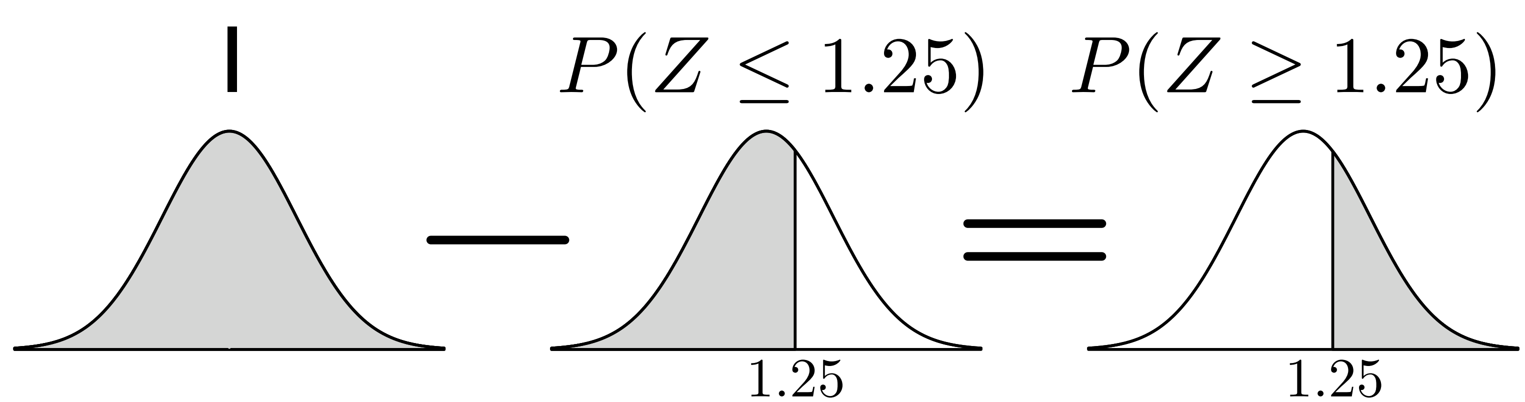
Normal probability examples

$$
\begin{aligned}
P(X \geq 90) &= 1 - P(X \leq 90) \\
&= 1 - P\left(\frac{X - 85}{4} \leq \frac{90 - 85}{4}\right) \\
&= 1 - P(Z \leq 1.25)
\end{aligned}
$$

curve of X

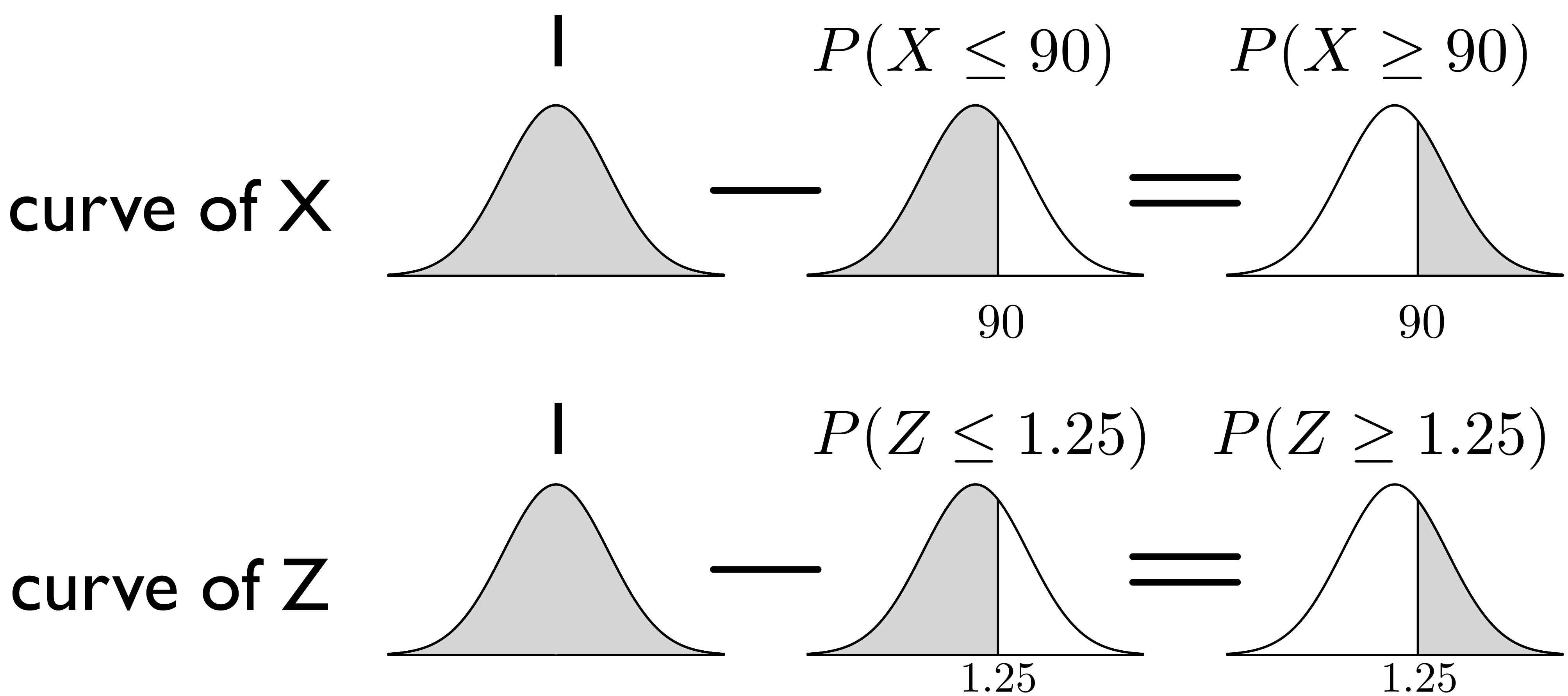


curve of Z



Normal probability examples

$$
\begin{aligned}
P(Z \geq 1.25) = \quad P(X \geq 90) &= 1 - P(X \leq 90) \\
&= 1 - P\left(\frac{X - 85}{4} \leq \frac{90 - 85}{4}\right) \\
&= 1 - P(Z \leq 1.25)
\end{aligned}
$$



Normal probability examples

$$P(Z \leq 1.25) =$$

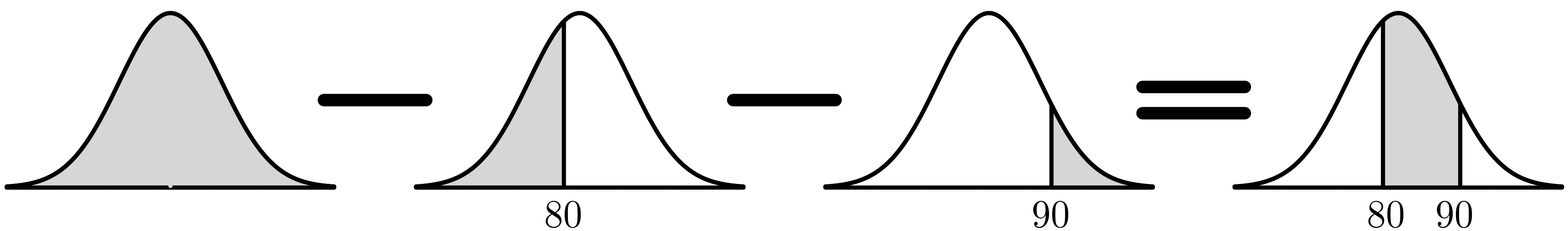
Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	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
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

$$P(X \geq 90) = 0.1056$$

Normal probability examples

2. The "B-range" is between 80 and 90. What is the probability that earns a B?

$$P(80 \leq X \leq 90)$$

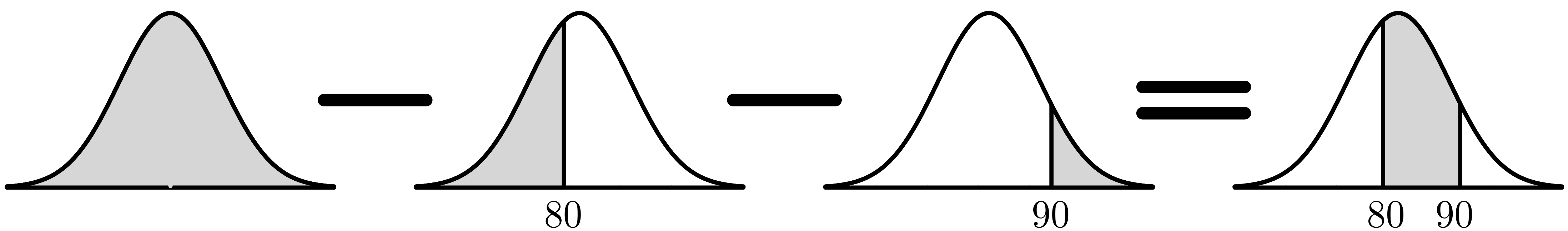


$$P(80 \leq X \leq 90) = 1 - P(X \leq 80) - P(X \geq 90)$$

Normal probability examples

2. The "B-range" is between 80 and 90. What is the probability that earns a B?

$$P(80 \leq X \leq 90)$$



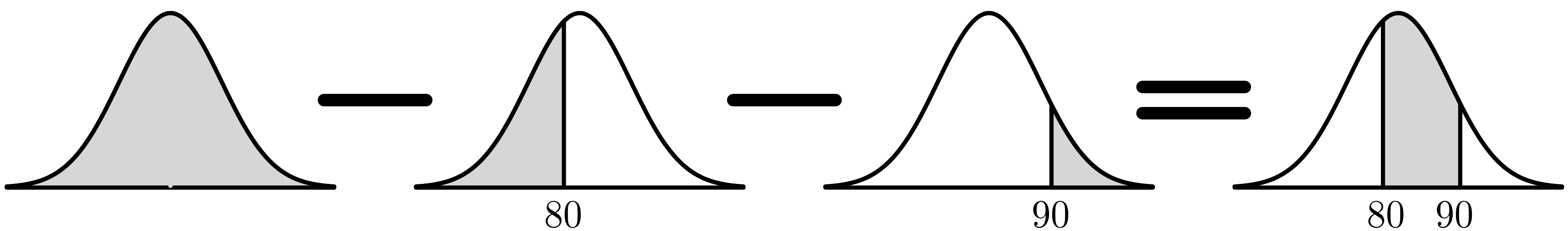
$$P(80 \leq X \leq 90) = 1 - P(X \leq 80) - P(X \geq 90)$$

But we have $P(X \geq 90) = 1 - P(X \leq 90)$

Normal probability examples

2. The "B-range" is between 80 and 90. What is the probability that earns a B?

$$P(80 \le X \le 90)$$



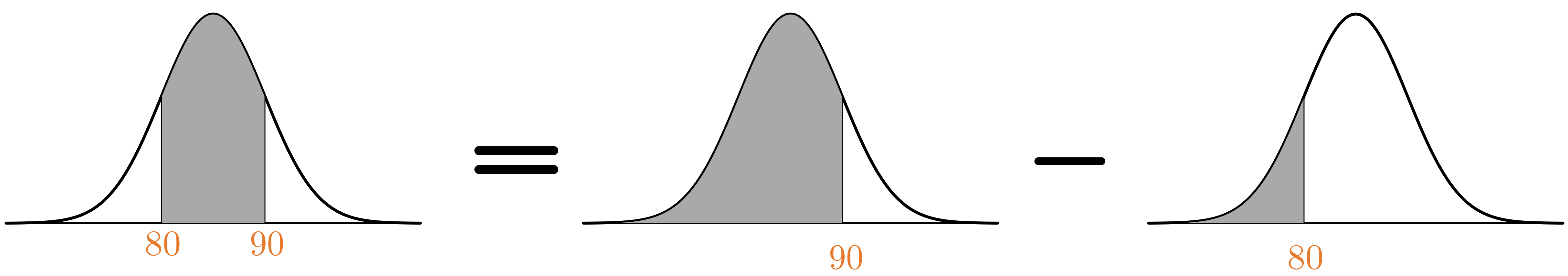
$$P(80 \le X \le 90) = 1 - P(X \le 80) - P(X \ge 90)$$

But we have $P(X \ge 90) = 1 - P(X \le 90)$

$$P(80 \le X \le 90) = P(X \le 90) - P(X \le 80)$$

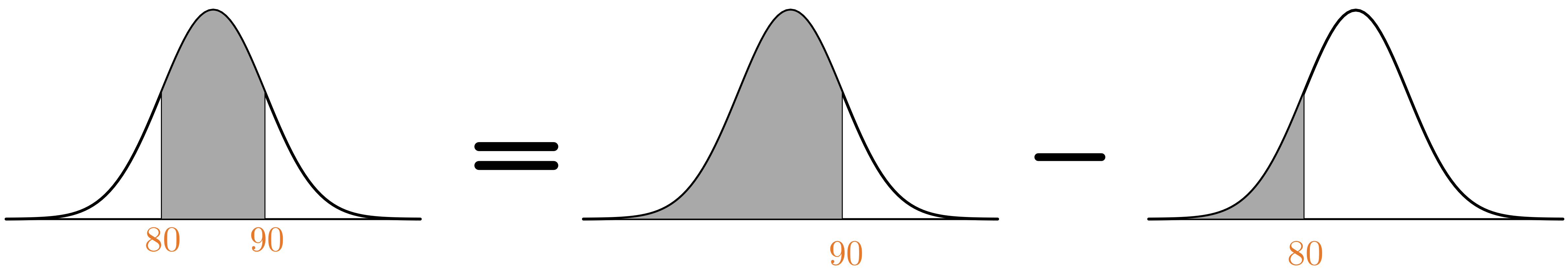
Normal probability examples

$$P(80 \leq X \leq 90) = P(X \leq 90) - P(X \leq 80)$$



Normal probability examples

$$P(80 \leq X \leq 90) = P(X \leq 90) - P(X \leq 80)$$

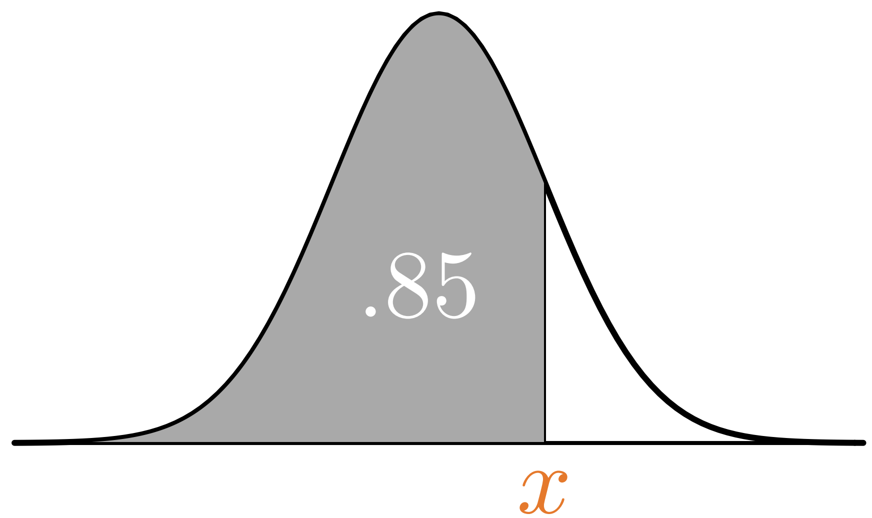


$$\begin{aligned} P(80 \leq X \leq 90) &= P(Z \leq \frac{90-85}{4}) - P(Z \leq \frac{80-85}{4}) \\ &= P(Z \leq 1.25) - P(Z \leq -1.25) \\ &= 0.8944 - 0.1056 \\ &= 0.7888 \end{aligned}$$

Normal probability examples

3. Eventually, she's at the 85th percentile. What is her score?

What is x such that $P(X \leq x) = .85$



$$P(X \leq x) = P(Z \leq \frac{x - 85}{4}) = 0.85$$

Normal probability examples

	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

Normal probability examples

	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

Normal probability examples

	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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

1.04 is such that $P(Z \leq 1.04) \simeq 0.85$ so that

$$\frac{x - 85}{4} = 1.04$$

Normal probability examples

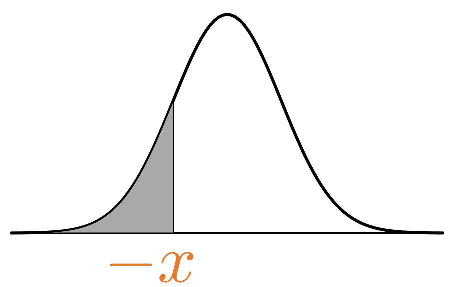
	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

1.04 is such that $P(Z \leq 1.04) \simeq 0.85$ so that

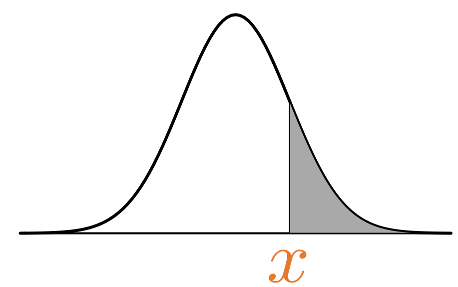
$$\frac{x-85}{4} = 1.04 \quad \longrightarrow \quad x = 89.16$$

A little symmetry

The normal curve is symmetric about is mean. It yields useful shortcuts in the calculation of probabilities and percentiles.

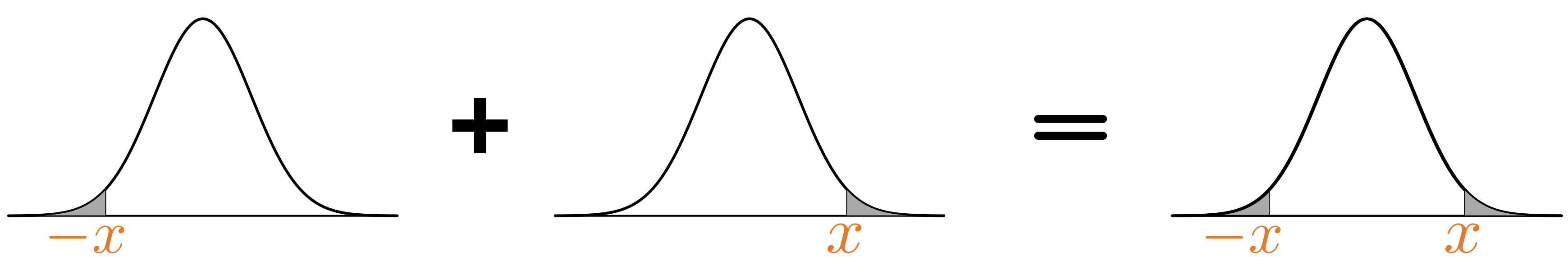


$P(X \leq -x)$



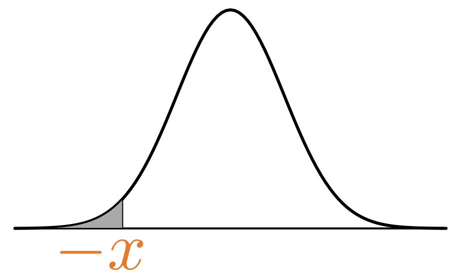
$P(X \geq x)$

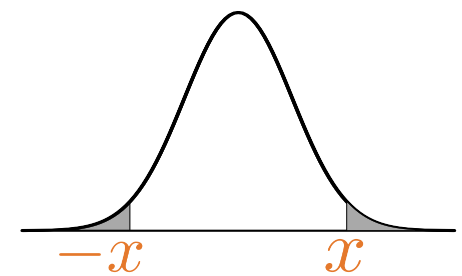
Therefore, if x is the 92nd percentile, then $-x$ is the percentile.



$P(X \leq -x)$ $\quad$ $P(X \geq x)$ $\quad$ $P(|X| \geq x)$

Checking normality

2 x 

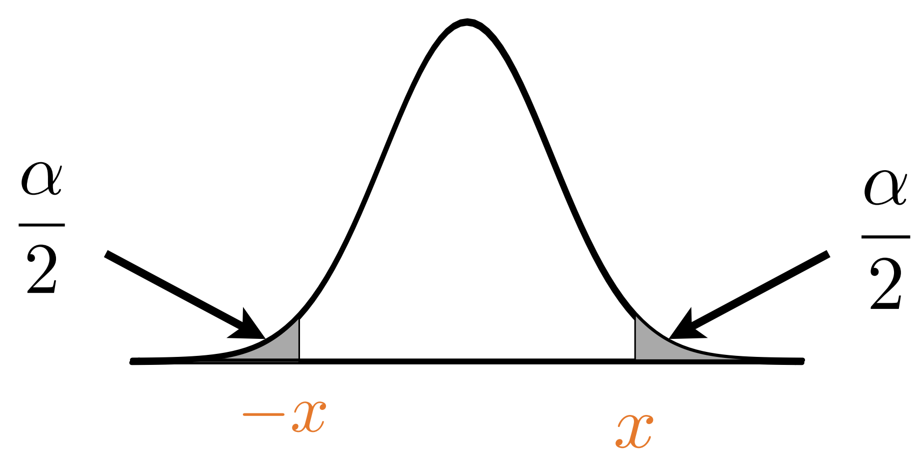
= 

$2P(X \leq -x)$ $\qquad$ $P(|X| \geq x)$

We will often be looking for x such that $P(|X| \geq x) = \alpha$ for some given small α

If we take $-x$ to be the $100\frac{\alpha}{2}$th percentile, then

$$P(|X| \geq x) = 2P(X \leq -x) = 2\left(\frac{\alpha}{2}\right) = \alpha$$



Operations that preserve normality

We have seen already one strong property of normal random variables, that is: if $X \sim N(\mu, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma}$$ is still a normal random variable

More generally, the affine transformation $Y = aX + b$ is a normal random variable with $E(Y) =$ and $var(Y) =$

Operations that preserve normality

We have seen already one strong property of normal random variables, that is: if $X \sim N(\mu, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma}$$ is still a normal random variable

More generally, the affine transformation $Y = aX + b$ is a normal random variable with $E(Y) =$ and $var(Y) =$

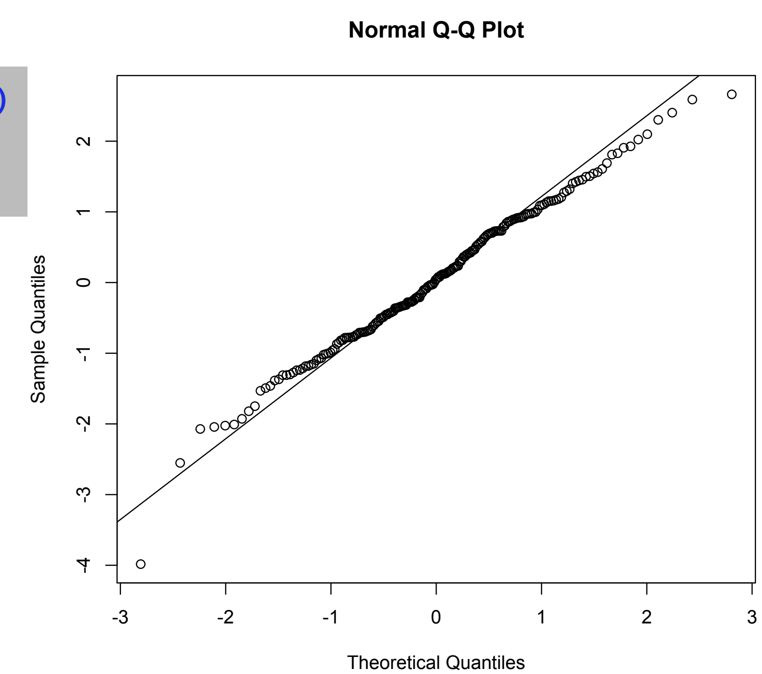
If $X_1, X_2, \ldots, X_n$ are **independent normal** random variables with the same distribution $X_i \sim N(\mu, \sigma^2)$ (i.i.d for independent and identically distributed), then

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$$

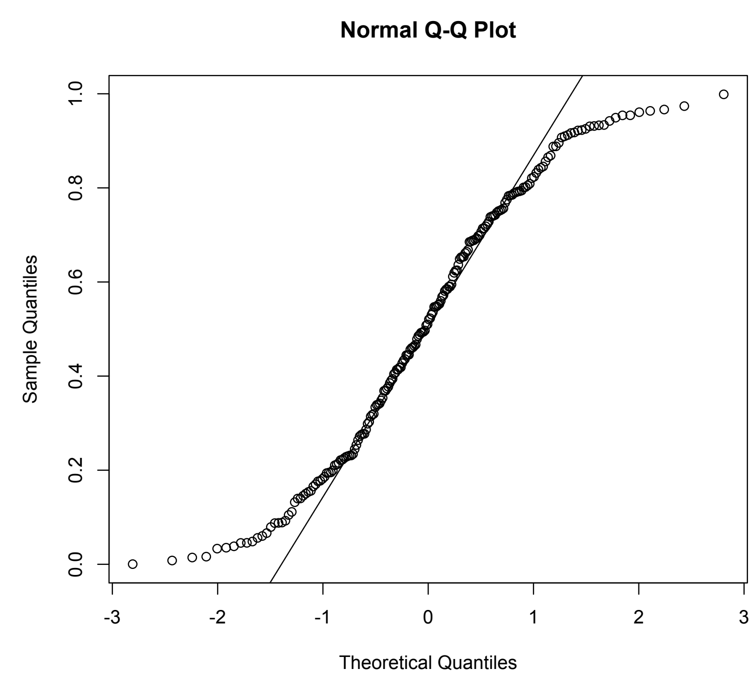
Checking normality

These two properties are valid **only** for normal random variables. It is therefore important to check whether it is plausible that a given set of observation has been drawn from a normal distribution.
A popular and useful tool is normal probability plots (or quantile-quantile plots, or simply Q-Q plots).

```
x=rnorm(200)
qqnorm(x)
qqline(x)
```



normal




```
x=runif(200)
qqnorm(x)
qqline(x)
```

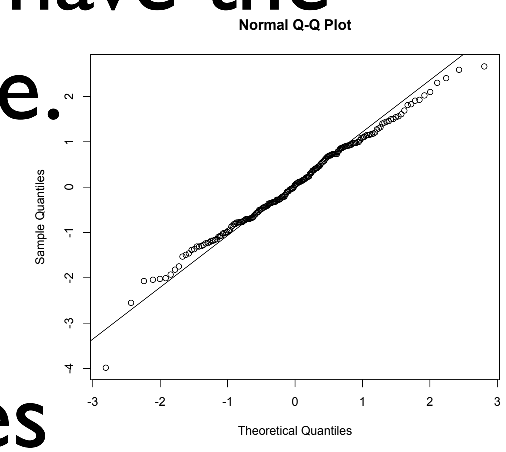
not normal

Normal probability plots

In the axes of the plots, we see quantile. A quantile, is a percentile but not expressed in percentage:

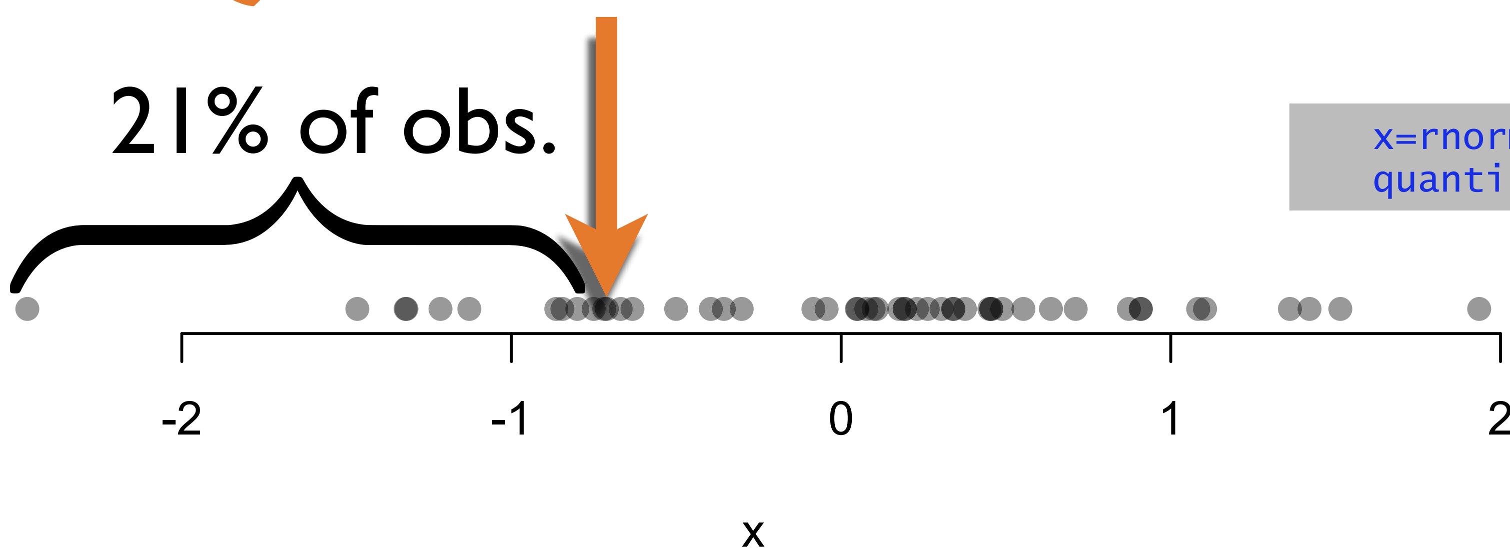
the 82nd percentile is the quantile of order .82

When two distributions are the same, they should have the same quantiles: in the plots, they should be on a line.



The theoretical quantiles can be read (backwards) in a normal probability table. The empirical quantiles are defined just like the median and the quartiles:

Quantile of order .21

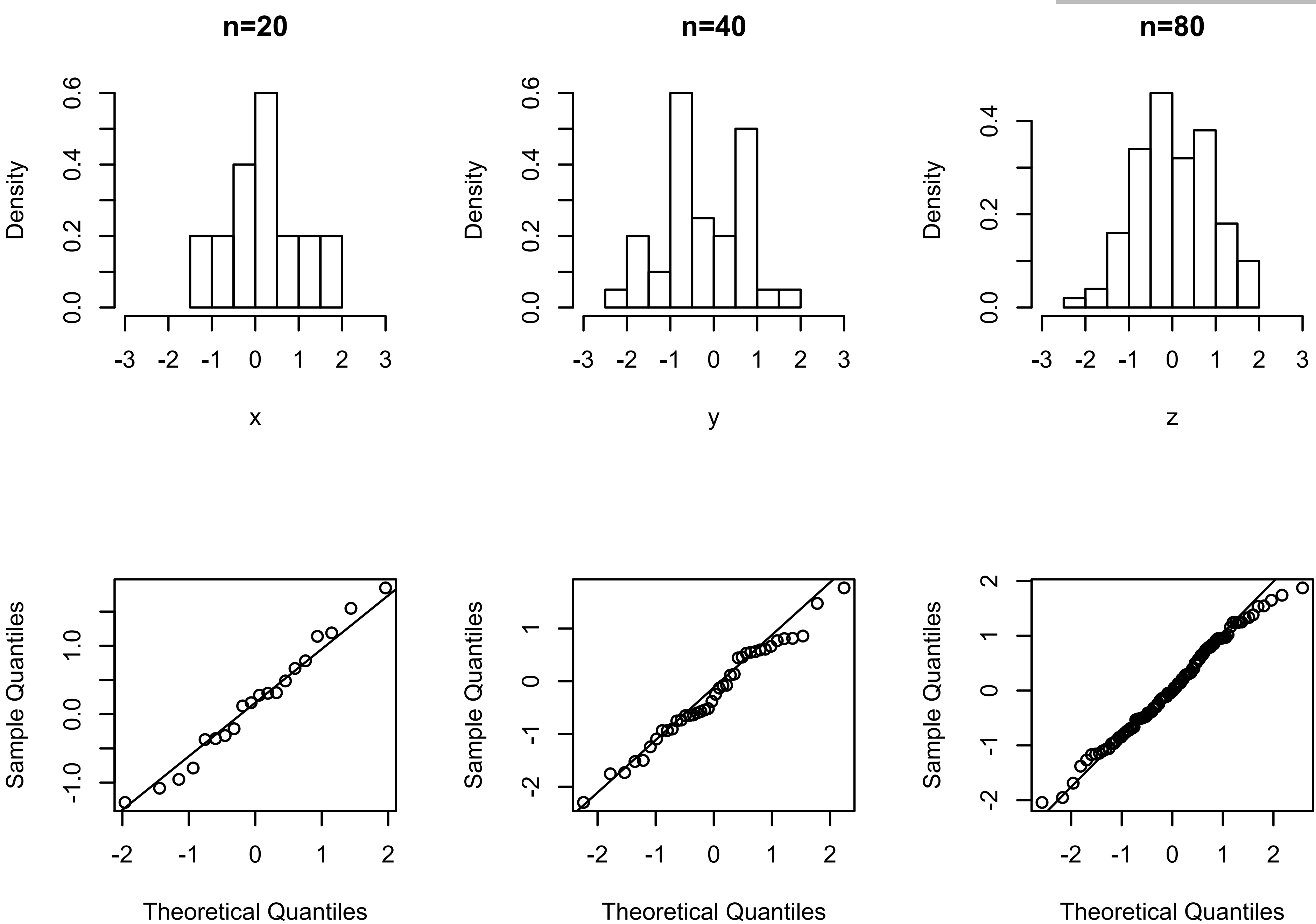



```
x=rnorm(50)
quantile(x, c(0.21, 0.73))
```

QQ-plots and histograms

```
x=rnorm(20)
y=rnorm(40)
z=rnorm(80)
par(mfrow=c(2,3))
hist(x, freq=F, main="n=20", xlim=c(-3,3))
hist(y, freq=F, main="n=40", xlim=c(-3,3))
hist(z, freq=F, main="n=80", xlim=c(-3,3))
qqnorm(x, main=""); qqline(x)
qqnorm(y, main=""); qqline(y)
qqnorm(z, main=""); qqline(z)
```

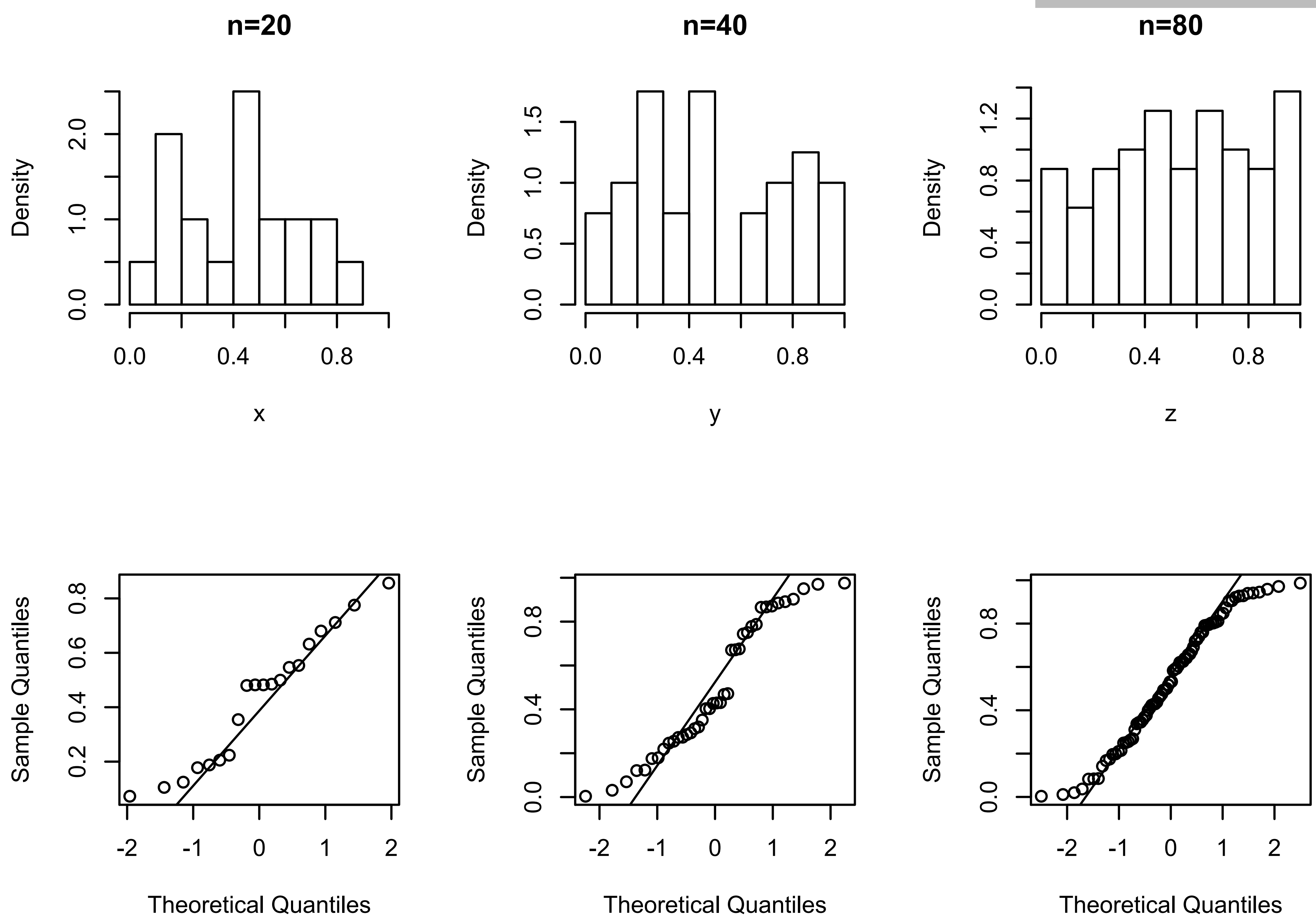
QQ-plots when the observations are simulated from normal distribution



QQ-plots and histograms

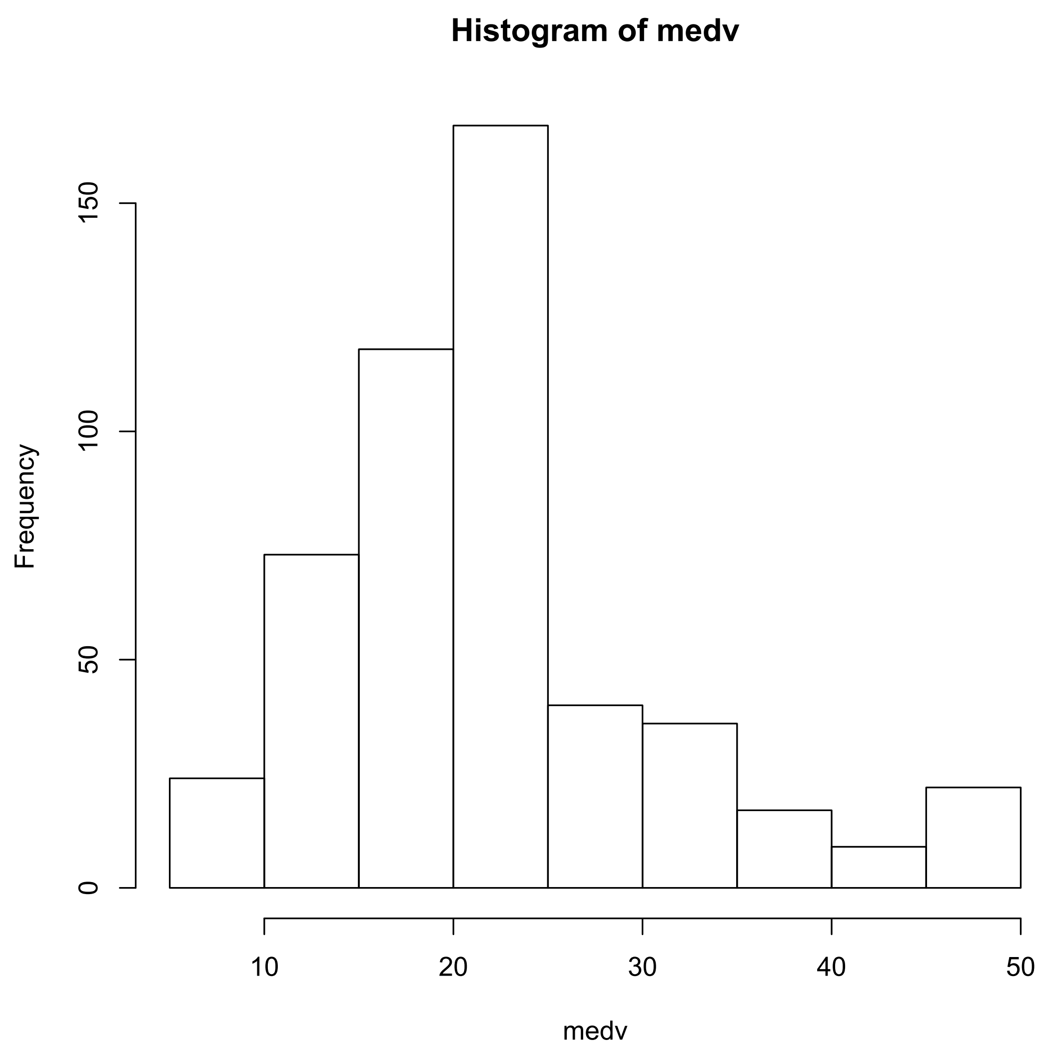
QQ-plots when the observations are not simulated from normal distribution

```
x=runif(20)
y=runif(40)
z=runif(80)
par(mfrow=c(2,3))
hist(x, freq=F, main="n=20", xlim=c(0,1))
hist(y, freq=F, main="n=40", xlim=c(0,1))
hist(z, freq=F, main="n=80", xlim=c(0,1))
qqnorm(x, main=""); qqline(x)
qqnorm(y, main=""); qqline(y)
qqnorm(z, main=""); qqline(z)
```

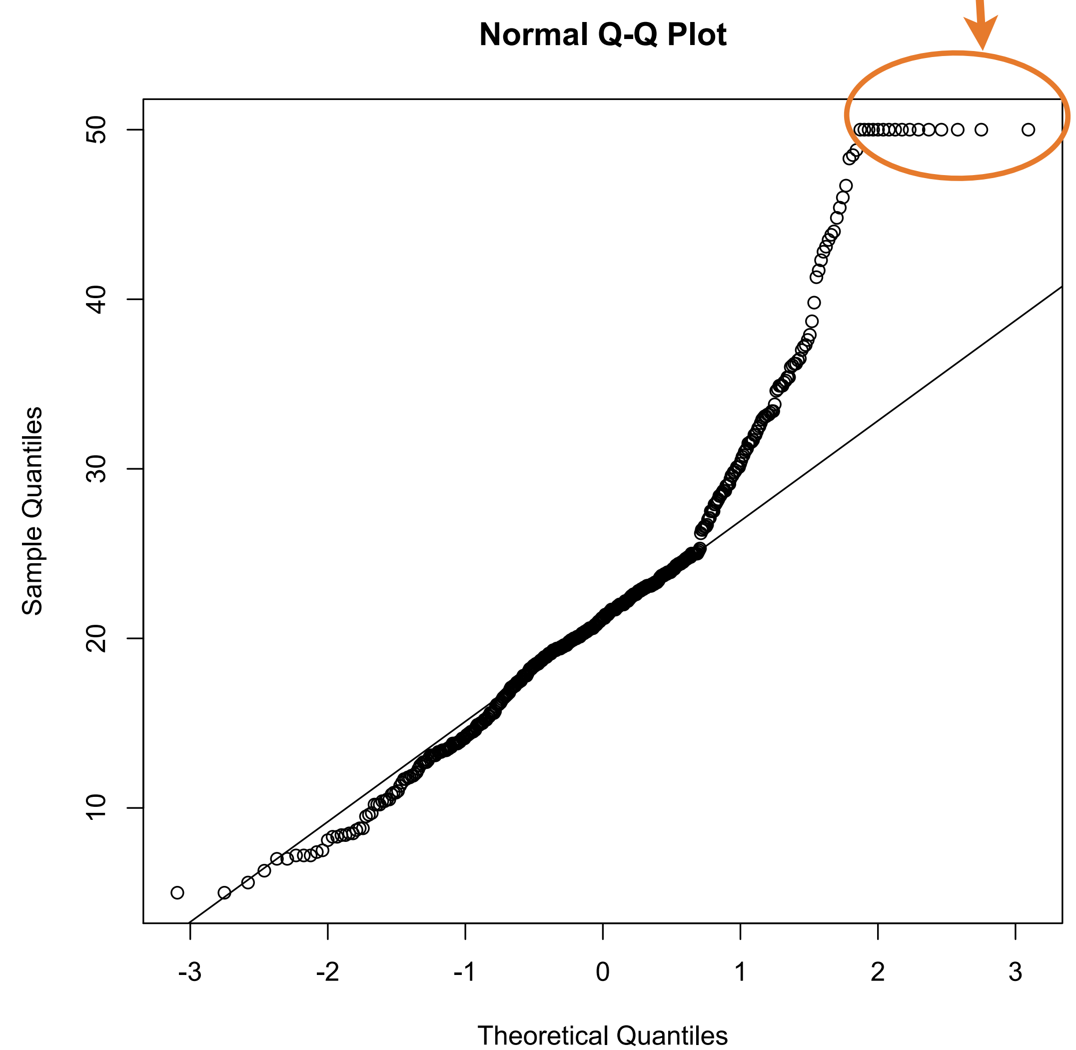


Real data: medv

```
library(mlbench)
data(BostonHousing2)
attach(BostonHousing2)
hist(medv)
qqnorm(medv); qqline(medv)
```



truncated observations

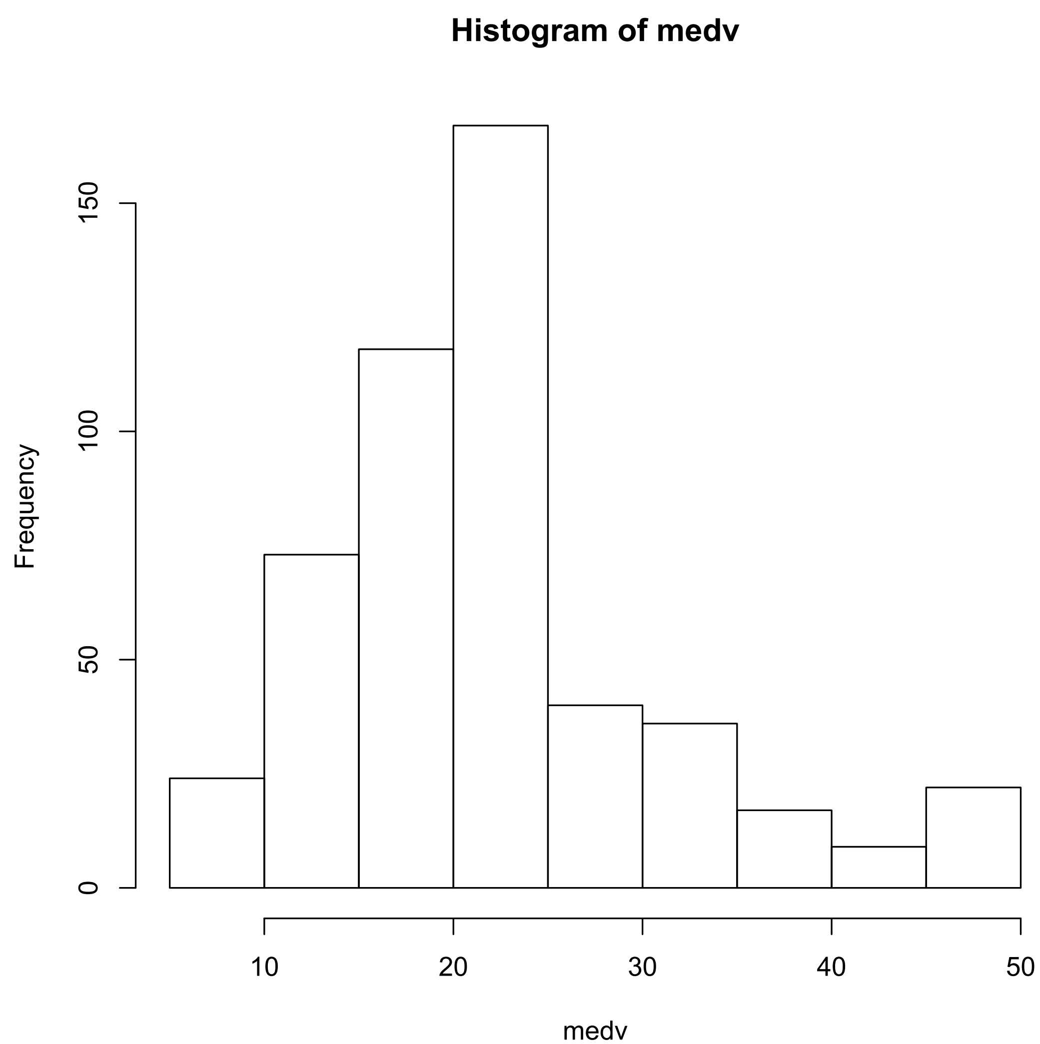


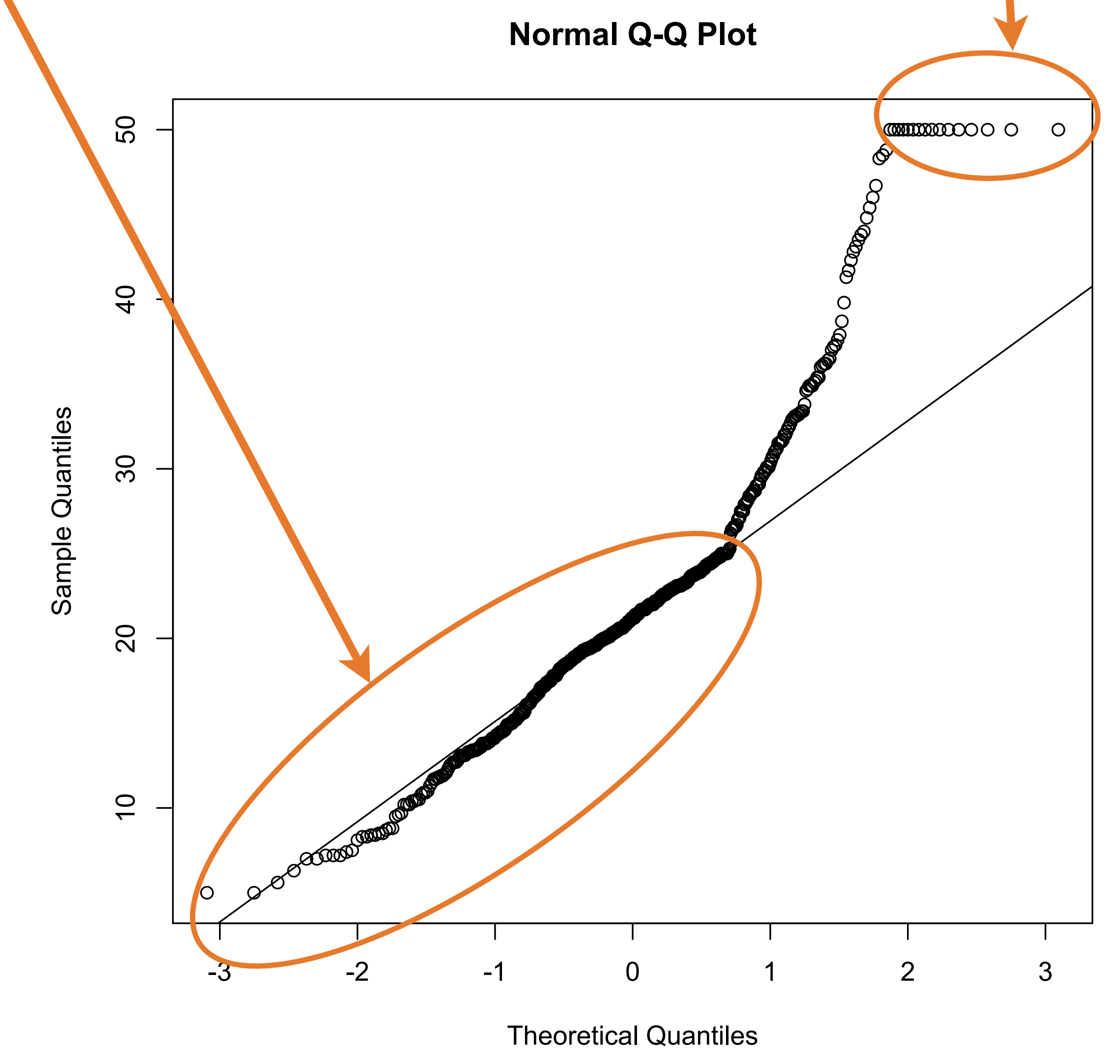
Real data: medv

```
library(mlbench)
data(BostonHousing2)
attach(BostonHousing2)
hist(medv)
qqnorm(medv); qqline(medv)
```

truncated observations

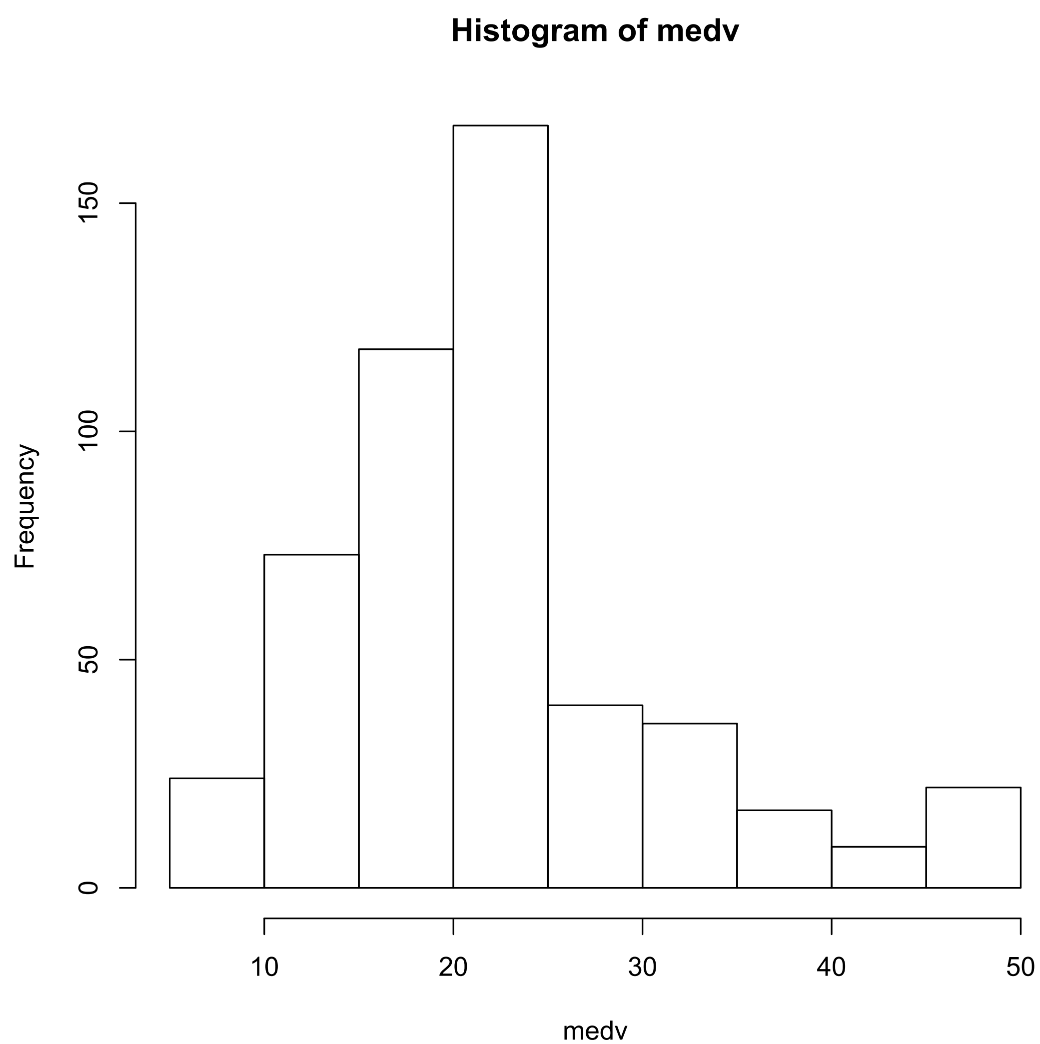
normal regime



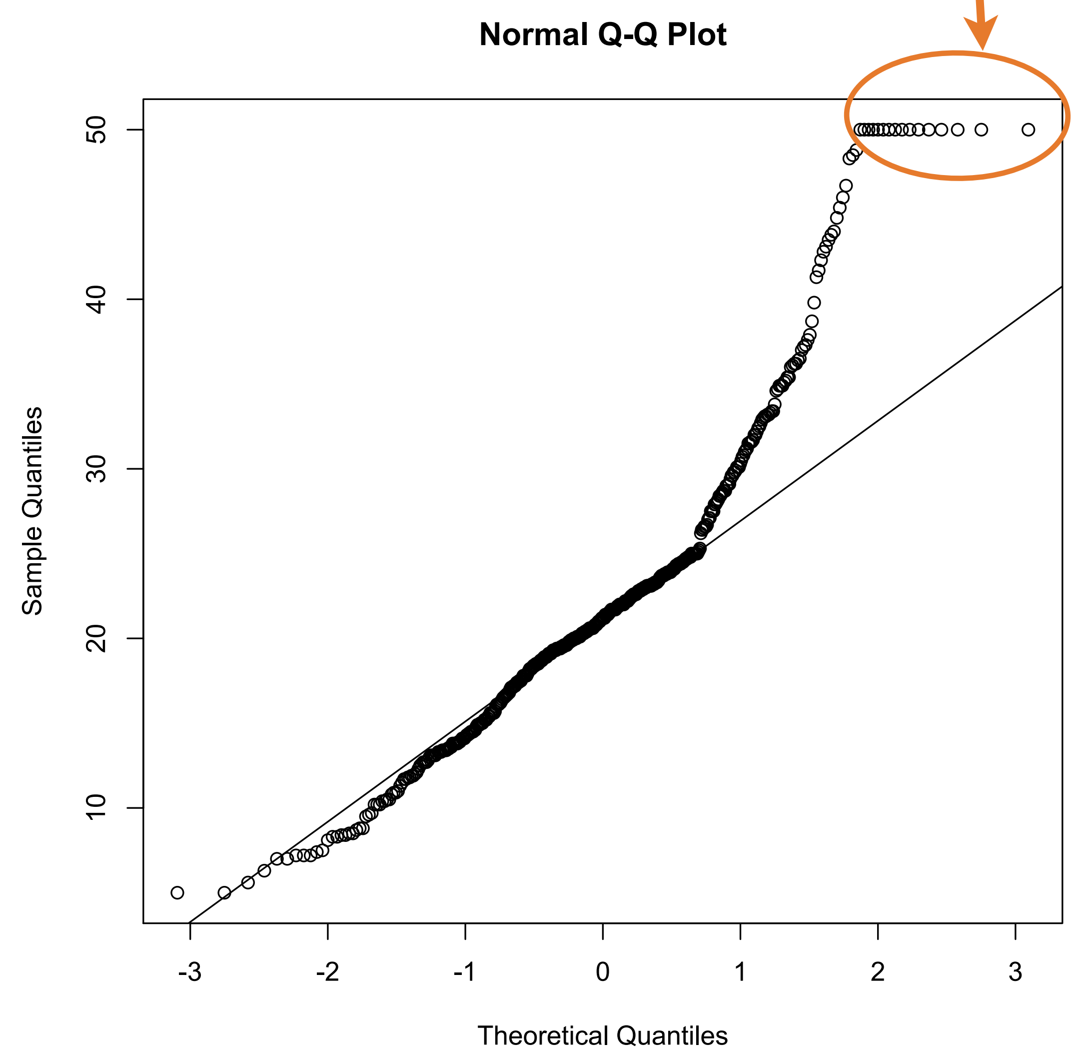


Real data: medv

```
library(mlbench)
data(BostonHousing2)
attach(BostonHousing2)
hist(medv)
qqnorm(medv); qqline(medv)
```



truncated observations

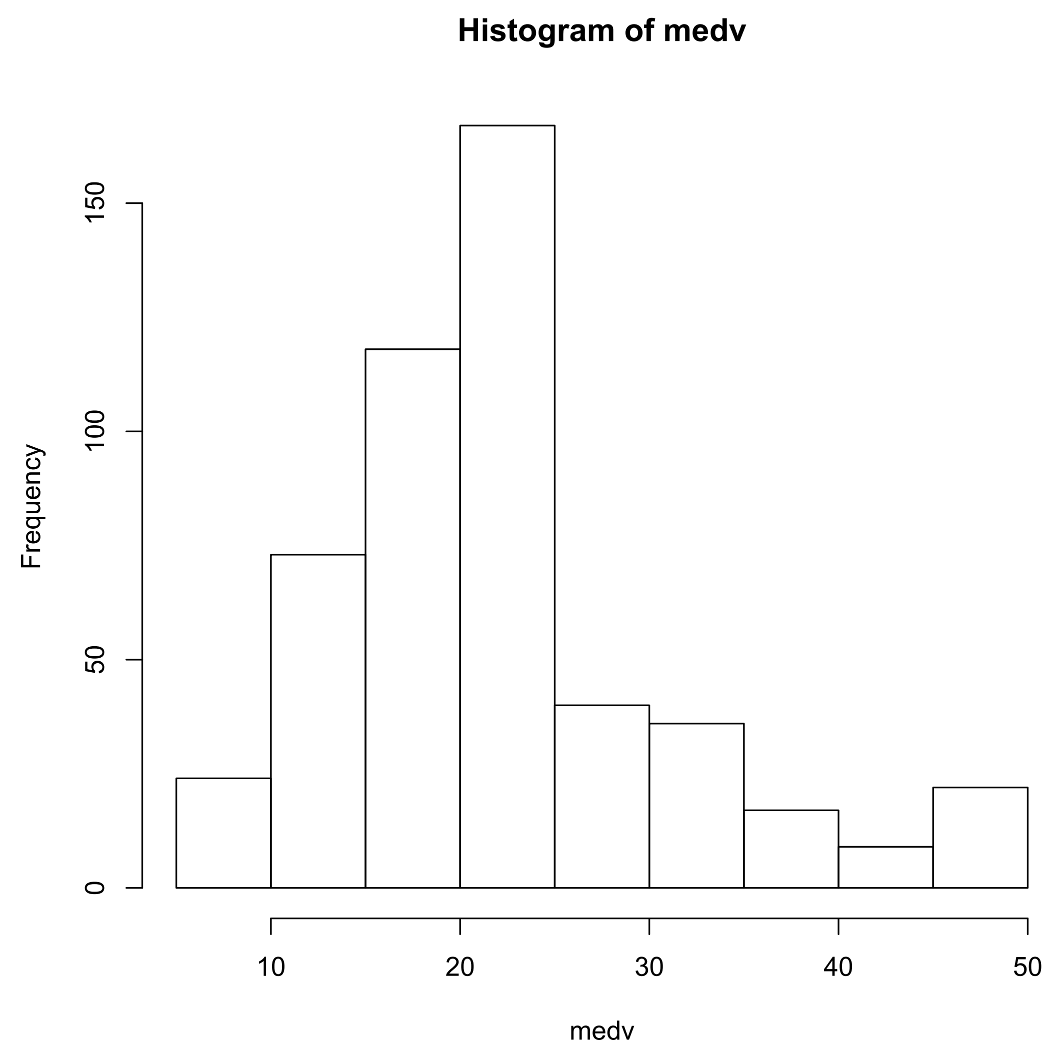


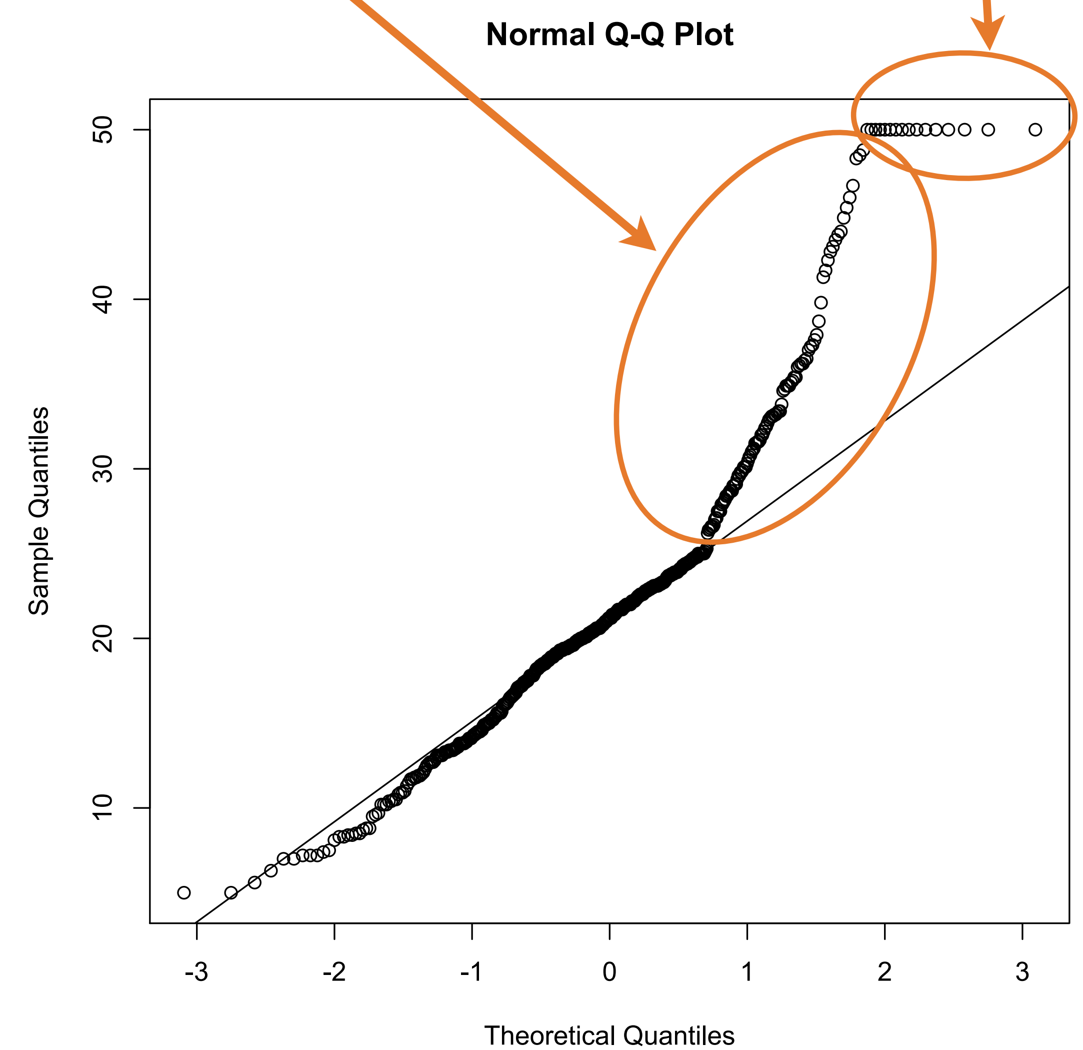
Real data: medv

```
library(mlbench)
data(BostonHousing2)
attach(BostonHousing2)
hist(medv)
qqnorm(medv); qqline(medv)
```

tails lighter than normal

truncated observations

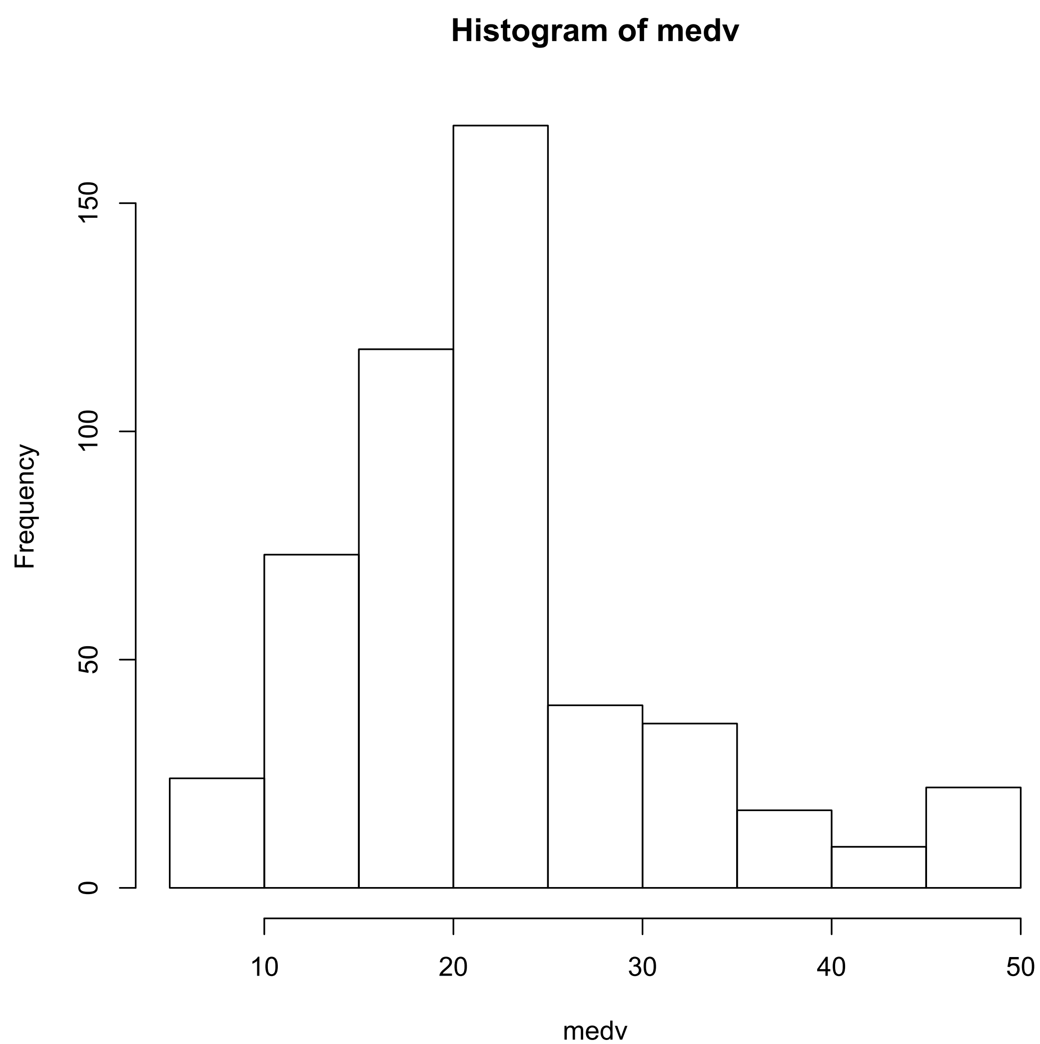


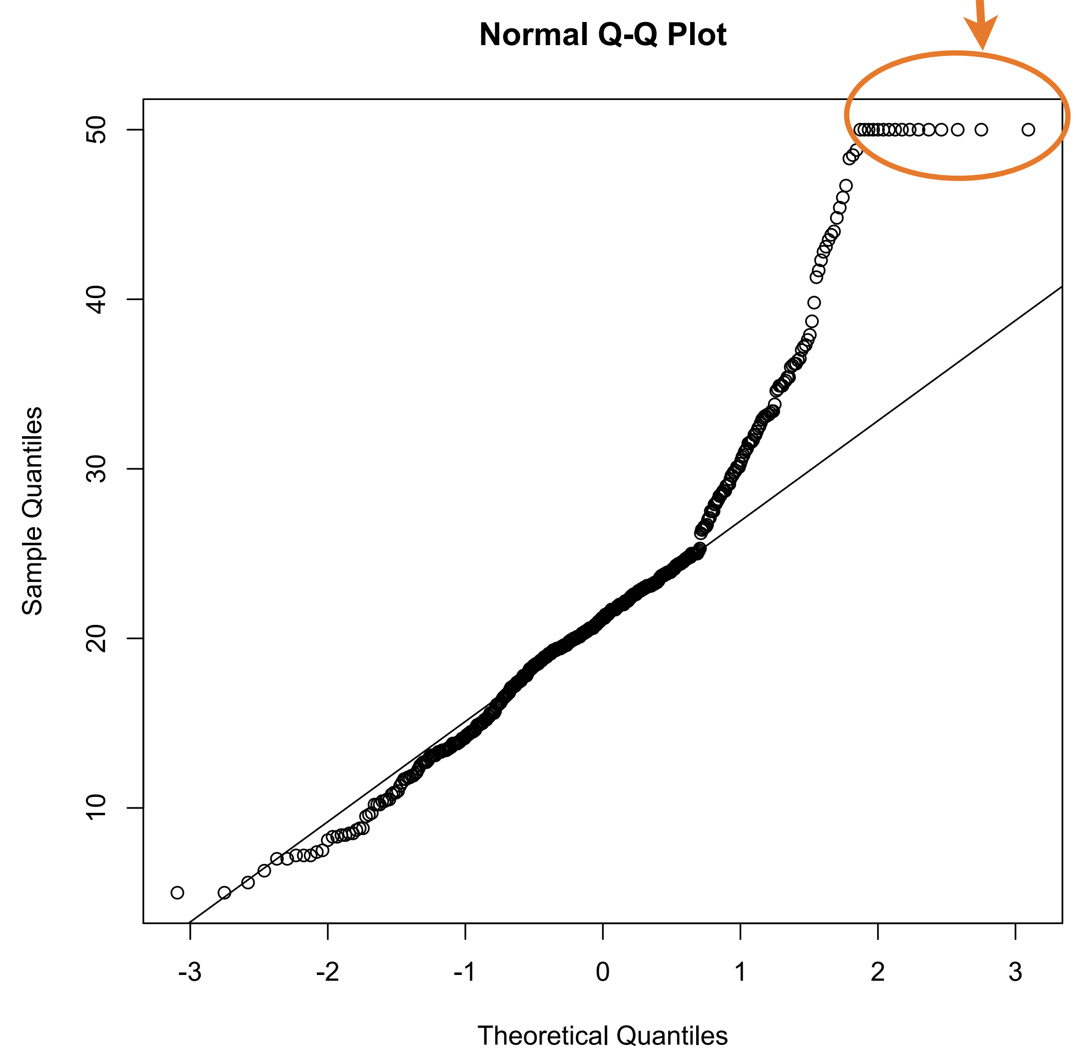


Real data: medv

```
library(mlbench)
data(BostonHousing2)
attach(BostonHousing2)
hist(medv)
qqnorm(medv); qqline(medv)
```

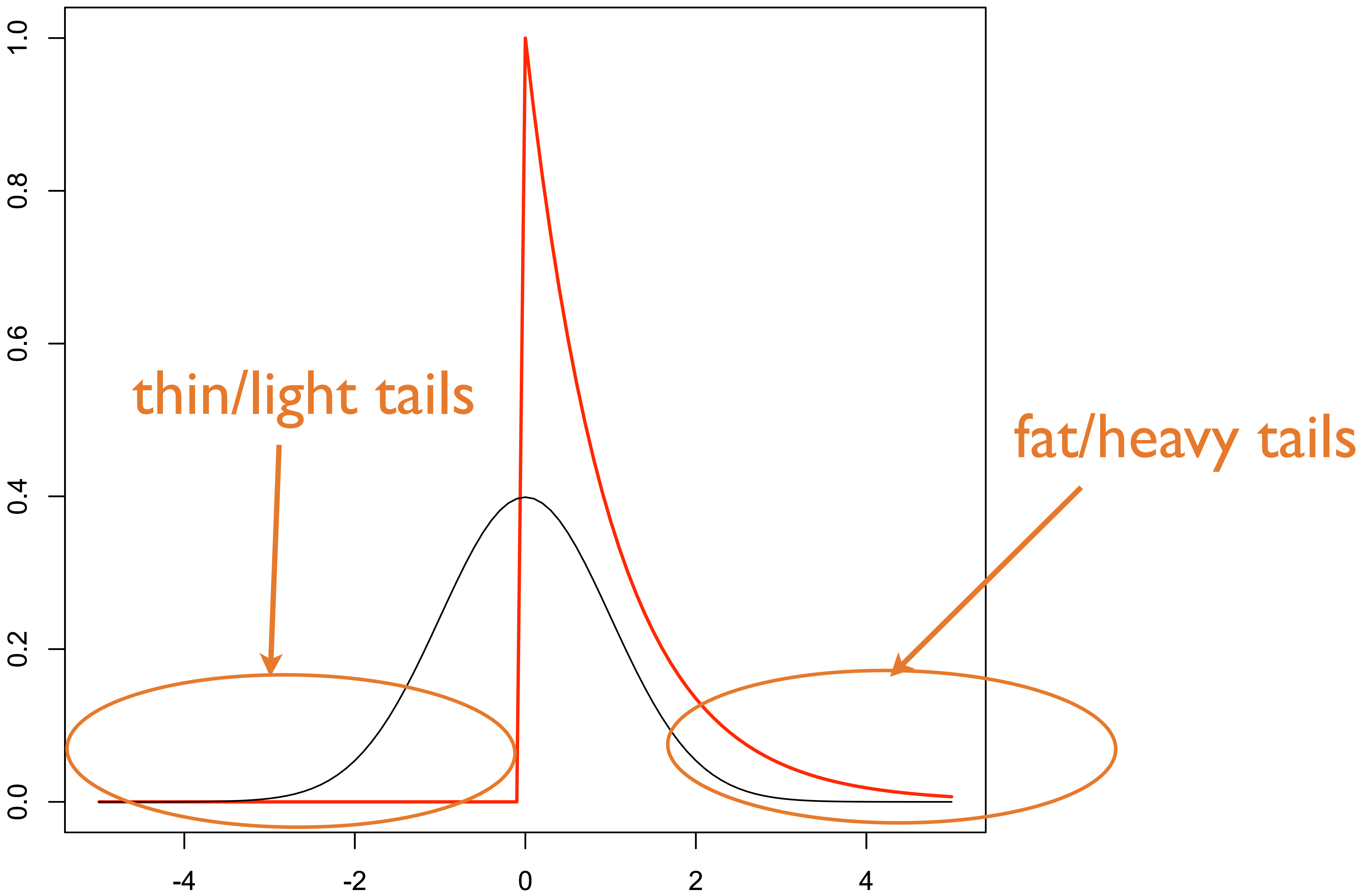
truncated observations





Light or heavy tails?

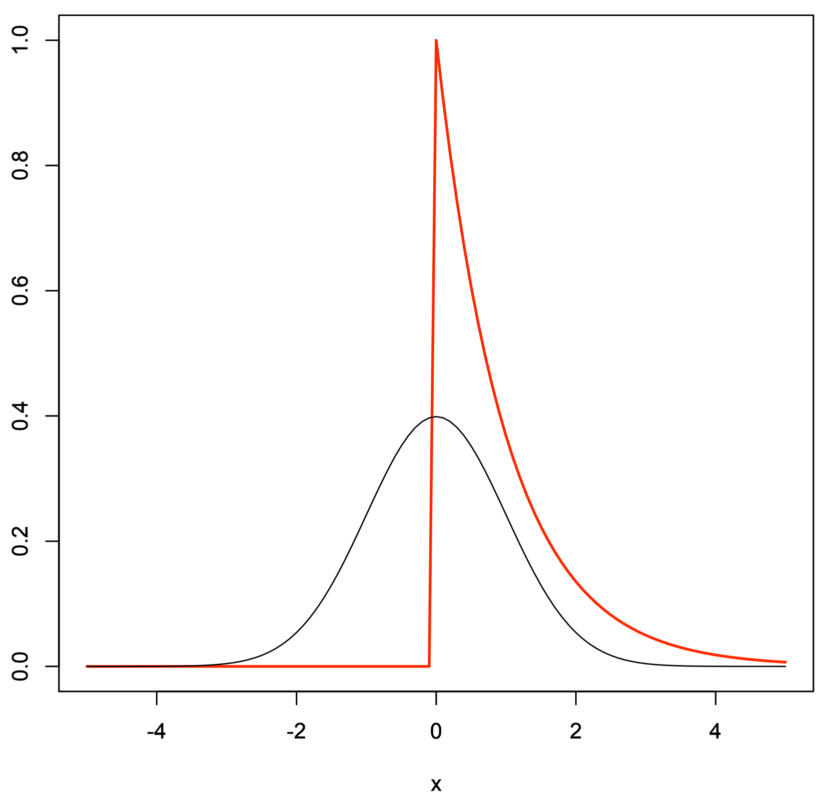
The thickness/weight of the tails can be read on a QQ-plot

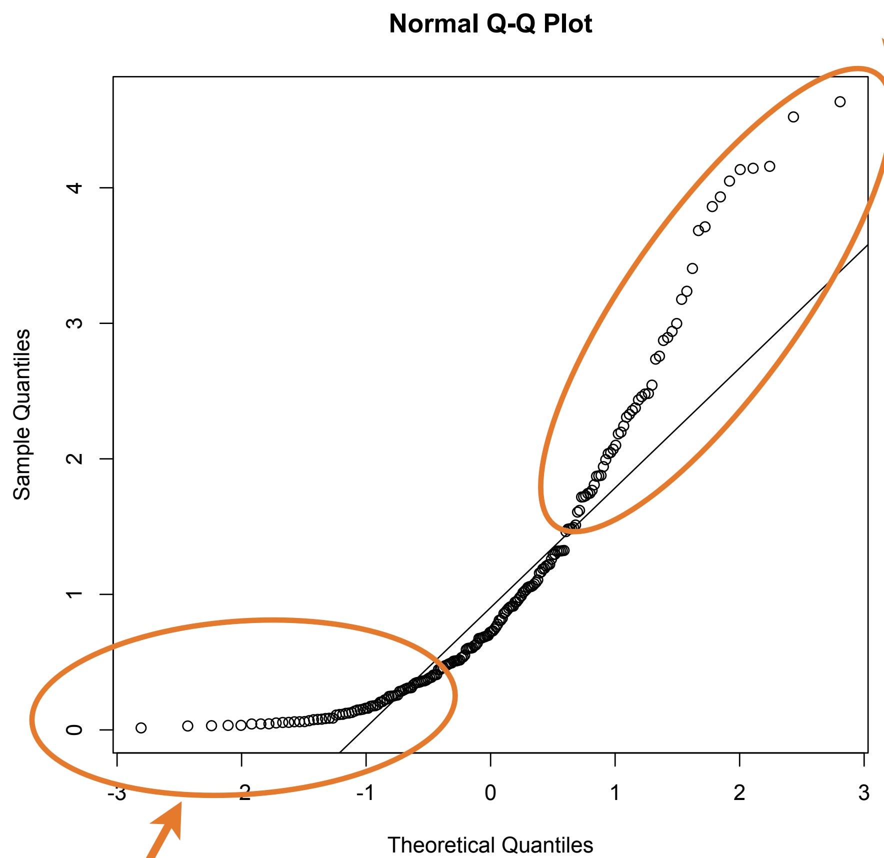


Light or heavy tails?

Light/thin or heavy/fat tails is **always** meant with respect to normal tails.

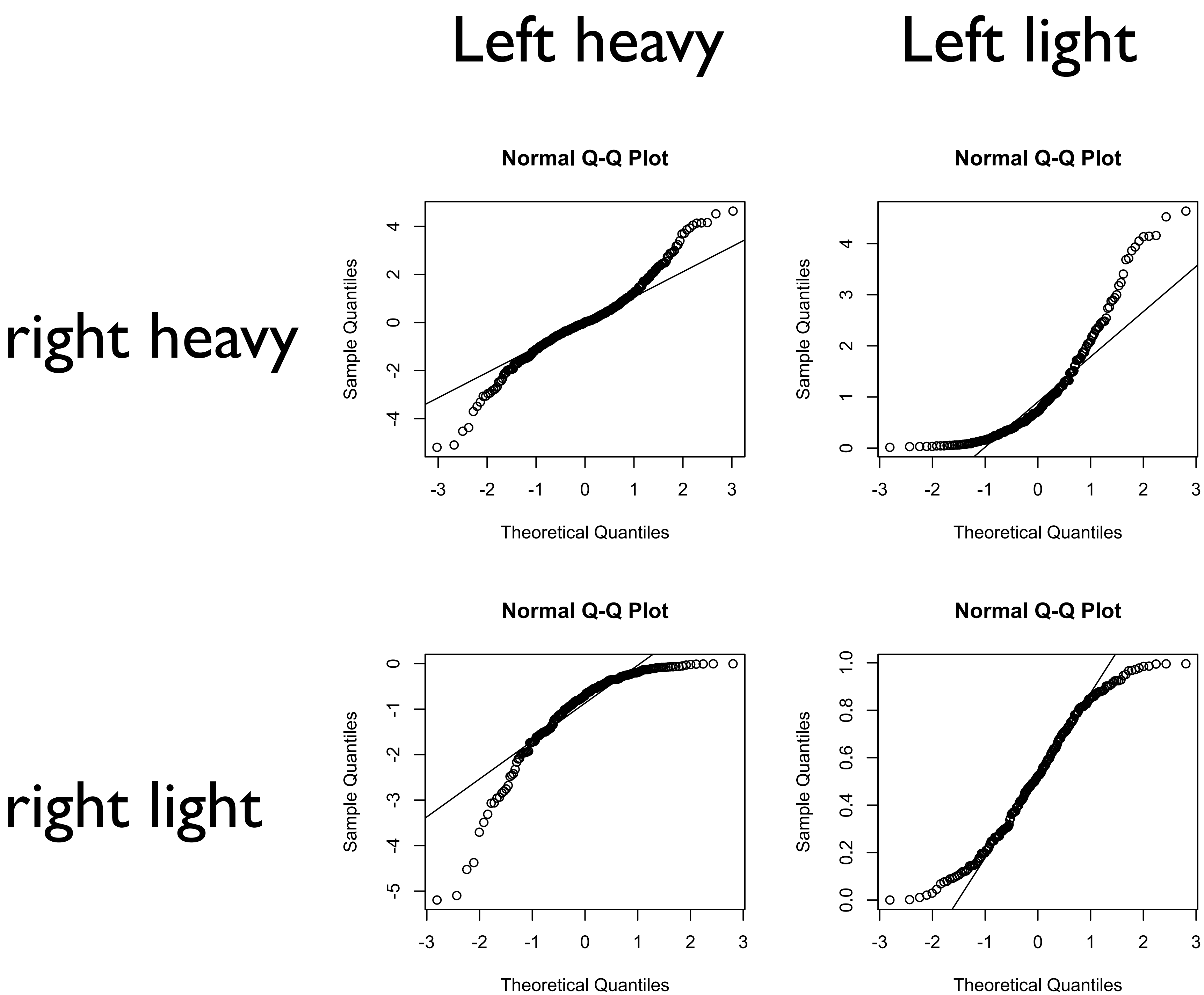
heavy right tail





light left tail

Light or heavy tails?



Other commonly used distributions

For the rest of the chapter, we are going to review commonly used distributions that are not the normal distribution. As usual, we split them into discrete and continuous ones.

Discrete:

- Bernoulli
- Binomial
- Geometric (in book, sec. 3.3.2)
- Poisson (in lab)

Continuous:

- Uniform
- Exponential (in lab)
- Chi-square

Chi-square distribution

The Chi-square distribution is obtained by a transformation of normal random variables that are not affine.

Let $X_1, X_2, \ldots, X_k$ be k i.i.d standard normal random variables then the random variable

$$Z = X_1^2 + X_2^2 + \cdots + X_k^2$$

Has Chi-square distribution with k degrees of freedom and we write

$$X \sim \chi_k^2$$

Chi-square distribution

The probability density function for the chi-square distribution is complicated and not very useful.

The important points about $X \sim \chi_k^2$ are:

$$X \geq 0$$

$$E(X) = k$$

$$var(X) = 2k$$

and the table for the quantiles of a chi-square distribution.