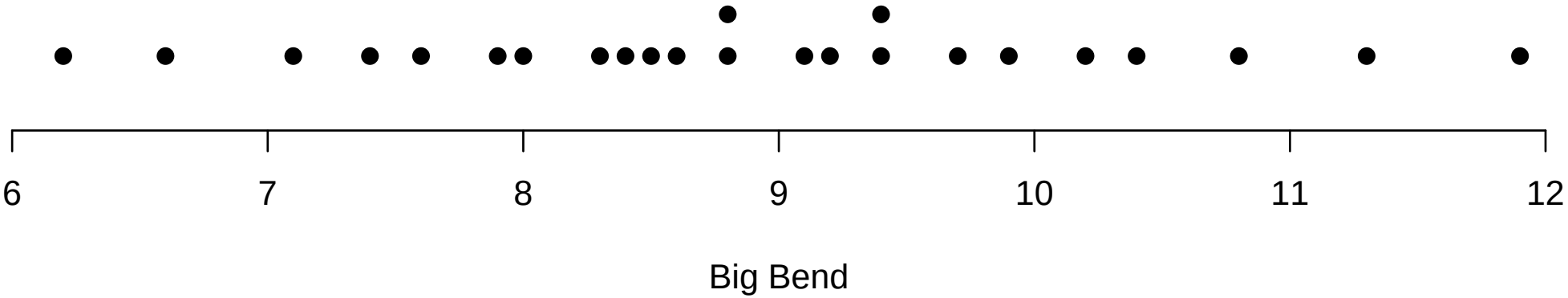
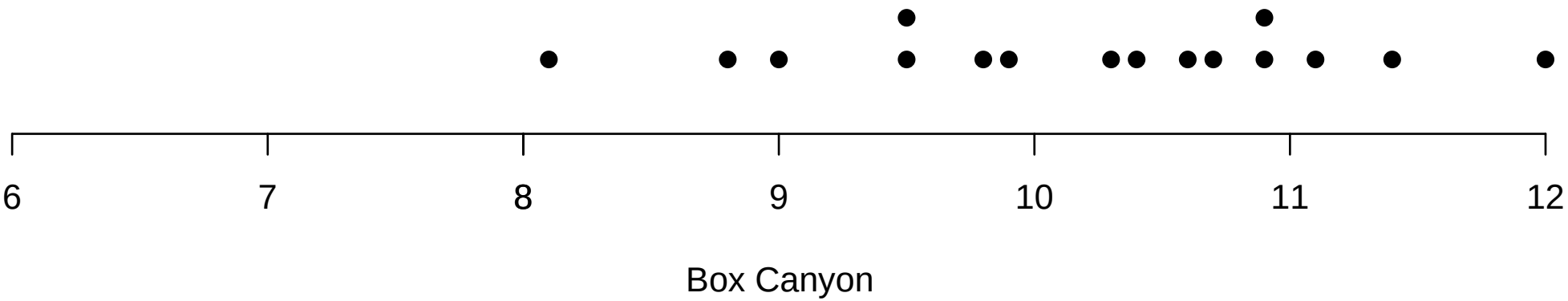
Outline

Big bend lizards tail length





Big bend lizards tail length

We want to know μ, the mean tail length in the entire Big Bend population of adult males of that lizard species.

```
> bigbend
 [1]  8.8  9.7 10.8  7.1  6.6  9.9 10.2  8.6 10.4 11.9  7.6  8.0  8.5
[16]  7.4  8.3  9.1  9.2  7.9  8.4 11.3  6.2  8.8
> mean(bigbend)
[1] 8.895833
> sd(bigbend)
[1] 1.429953
> length(bigbend)
[1] 24
```

$\bar{y} = 8.896$ cm is our best estimate for μ.

How good is this estimate? How far is μ from 8.896 cm?

Standard error of the mean

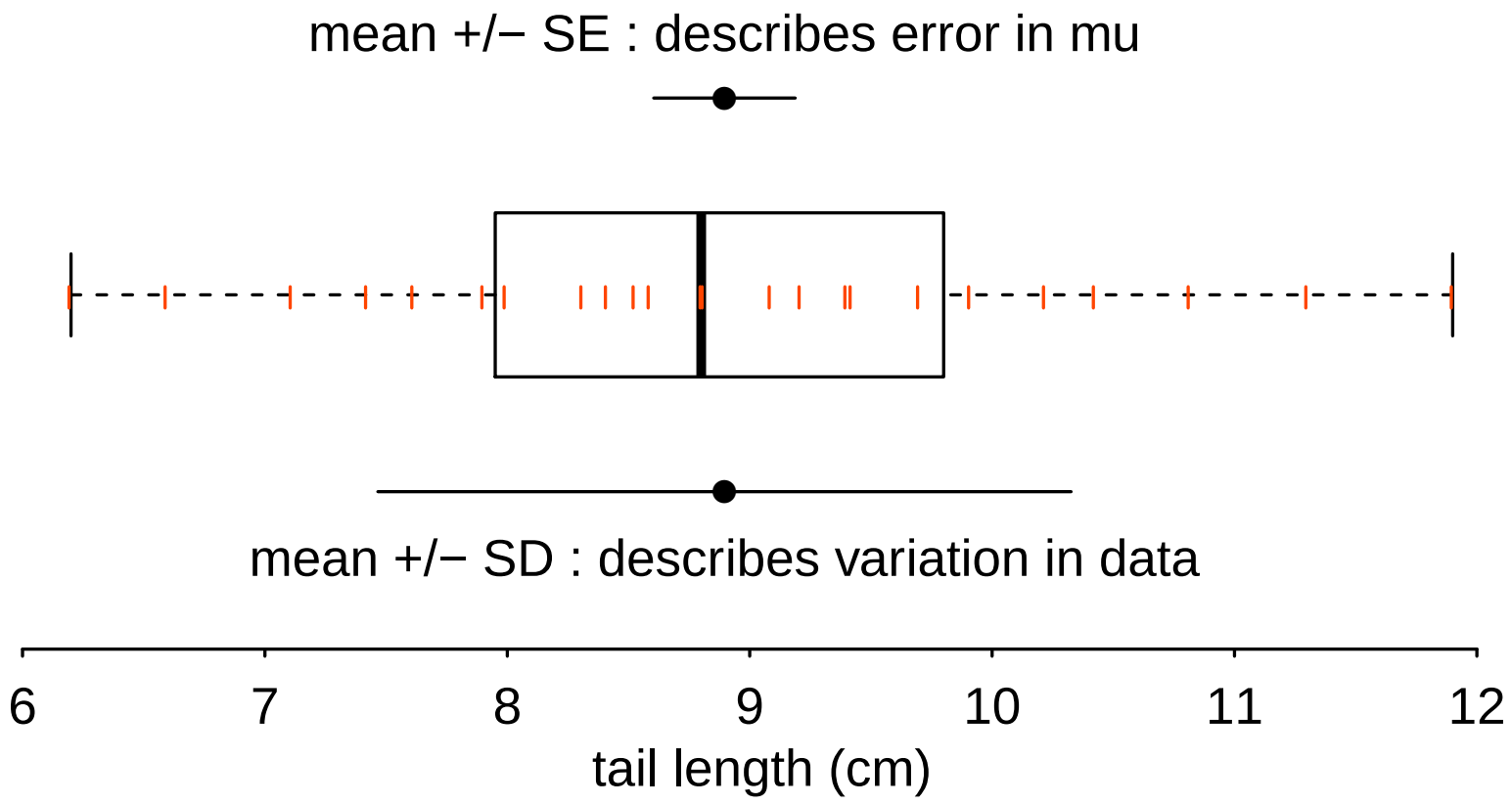
We know the standard deviation of $\bar{Y}$ is $\sigma/\sqrt{n}$. But we don't know σ. Hopefully, the standard deviation of the sample, s, is close to σ.

$\text{SE}_{\bar{y}} = \dfrac{s}{\sqrt{n}}$ is the standard error of the mean.

$\text{SE}_{\bar{y}}$ is an estimate of the standard deviation of $\bar{Y}$, from expt to expt: gives us an idea of how far $\bar{y}$ is from μ typically (in a typical experiment).

Here: $s = 1.43$ and $n = 24$, so $\text{SE}_{\bar{y}} = 1.43/\sqrt{24} = 0.292$

SD of the data and SE of the mean



What happens to *s* (SD) when the sample size increases?

What happens to $SE_{\bar{y}}$ when the sample size increases?

The t-distribution

If $Y_1, \ldots, Y_n$ have a normal distribution, $\bar{Y}$ has one too, and

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)\,.$$

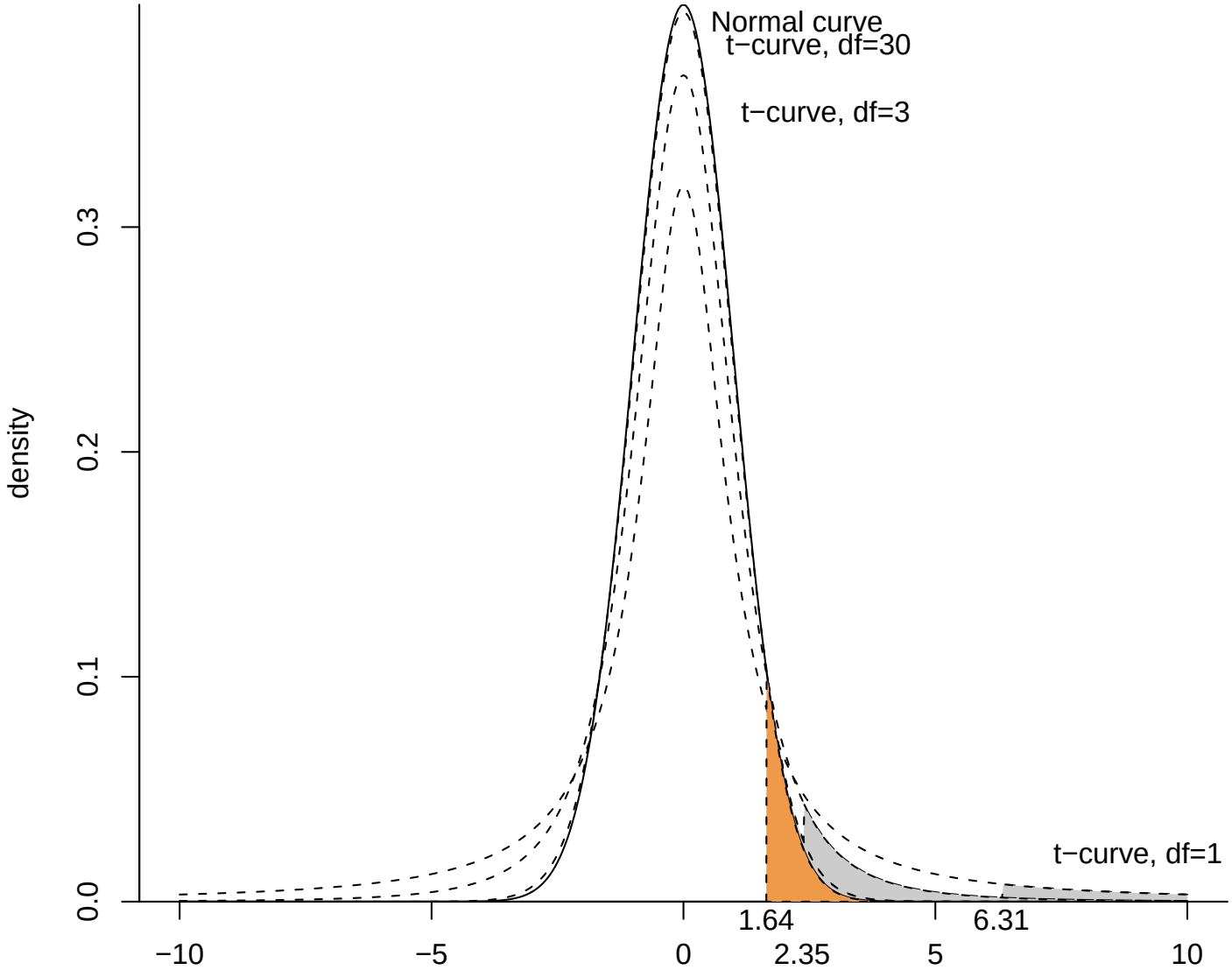
Application: $\bar{Y}$ and μ are no more than 2 $\sigma/\sqrt{n}$ apart in 95% experiments.

When we replace $\sigma/\sqrt{n}$ by SE$= s/\sqrt{n}$,

$$\frac{\bar{Y} - \mu}{\text{SE}} = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim \text{ t-distribution, } n - 1 \text{ degrees of freedom.}$$

Application: $\bar{Y}$ and μ are no more than ?how many? SE apart in 95% experiments?

The t-distribution



Mechanics of a confidence interval

1. **Choose a confidence level.** Typically, 95%. Polls use 90% or 95%.
2. **Find the value** t such that $\mathbb{P}\{-t \leq T \leq t\} =$ confidence level. It also means

$$\mathbb{P}\{T \geq t\} = (1 - \text{confidence level})/2$$

   use Table C, with degree of freedom df$= n - 1$.
3. **Construct the interval:** $\bar{y} \pm t\text{SE}_{\bar{y}}$ i.e.

$$(\bar{y} - t\text{SE}_{\bar{y}}, \bar{y} + t\text{SE}_{\bar{y}})$$

4. **Conclude:**

We are 95% confident that the mean tail length of all adult male lizards from this Big Bend population is between 8.29 cm and 9.50 cm.

Confidence interval: Big Bend lizards' tail length

1. **Confidence level.** We will do both 90% and 95%.
2. **Find the value** t: such that $\mathbb{P}\{T \geq t\} = .05$ for level 90% and .025 for level 95%.
   Degree of freedom: df=$24 - 1 = 23$.
   t-Table gives: $t = 1.71$ for 90% confidence and $t = 2.07$ for 95% confidence. With R:

```
> qt(.950, df=23)      > qt(.975, df=23)
[1] 1.713872           [1] 2.068658
```

3. **Interval:** We had $\bar{y} = 8.896$, $s = 1.430$ and $\text{SE}_{\bar{y}} = 1.43/\sqrt{24} = 0.292$.
   Radius of interval (bull's eye): $t * \text{SE}_{\bar{y}} = 0.500$ (90% confidence) and 0.604 (95% confidence).
   The interval is 8.896 ± 0.500 or 8.896 ± 0.604, i.e.

$$(8.396, 9.396) \quad \text{for 90\% confidence}$$
$$(8.292, 9.500) \quad \text{for 95\% confidence}$$

4. **Conclude.**

Degree of freedom: $n - 1$

Recall that

$$s^2 = \frac{1}{n-1}\left((y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\right)$$

Remember example $n = 3$ and $y_1 - \bar{y} = 3$, $y_2 - \bar{y} = 5$. Then no choice: $y_3 - \bar{y}$ had to be -8.

The last deviation is completely specified by the first $n - 1$.
The variance is completely specified by $n - 1$ deviations, or $n - 1$ pieces of information.

df = # pieces of information needed for computing s^2.

Imagine a sample with a single observation.

R: t.test() for confidence interval from raw data

```
> bigbend
 [1]  8.8  9.7 10.8  7.1  6.6  9.9 10.2  8.6 10.4 11.9  7.6  8.0  8.5
[16]  7.4  8.3  9.1  9.2  7.9  8.4 11.3  6.2  8.8

> t.test(bigbend, conf.level = .90)

        One Sample t-test
data:  bigbend
t = 30.4769, df = 23, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 8.395575 9.396092
sample estimates:
mean of x
 8.895833
```

```
> t.test(bigbend)
...
95 percent confidence interval:
 8.292017 9.499649
```

```
> t.test(boxcanyon)
...
95 percent confidence interval:
  9.631525 10.730975
```

!Warning! do not use 1-sample CIs for comparing 2 samples.

True or False?

95% CI for the mean tail length in Big Bend: 8.29-9.50 cm.

With the same data, a 99% confidence interval would be larger.

In a second sample of same size (24 lizards), there is a 95% chance that the new sample mean will be in (8.29, 9.50).

The probability is 95% that the sample mean is in (8.29, 9.50).

The probability is 95% that the population mean is in (8.29, 9.50).

The confidence is 95% that the population mean is in (8.29, 9.50).

In the population, 95% of all adult male lizards are in (8.29, 9.50).

In the sample, 95% of all adult male lizards are in (8.29, 9.50).

Planning a study: how big should *n* be?

When planning a study, it is always a question we ask.

- How many people am I going to interview?
- How many blood samples to I need?
- How many plants to I need to grow?

Trade-off between accuracy and cost.
We want just the right number *n* to reach the conclusion.

We need to set a goal.

- Polls: “margin of error” at least as small as 1%.
- Lizards: pilot study, expt to be repeated in 10 different populations. We might want interval length ≤ 0.5 cm for each.
- Or: require SE no greater than a given size: SE ≤ 0.25 cm.

Planning a study: how big should n be?

Solving this problem requires a guess for the population SD.
It usually involves preliminary data.

Lizard tail length: guess is that SD $= s = 1.43$ mg.
Aim: SE ≤ 0.25 mg.
Then we solve SE $=$ SD $/\sqrt{n}$

$$n = \left(\frac{\text{guessed SD}}{\text{desired SE}} \right)^2$$

$n = (1.43/0.25)^2 = 32.72$ (no unit). We would sample 33 lizards for the next experiment / location.

Conditions for validity

1. Most importantly: the sampling process needs to be like random sampling. **Independence** of observations, sampled from the **target** population. At the end, we should draw conclusions about the adequate population.

   If the sampling process is biaised, the confidence interval will greatly overstate the confidence we should have.

   Example: milk yield quality (e.g. Somatic Cell Count). If sampling biased toward large farms, confidence interval likely to be unreliable.

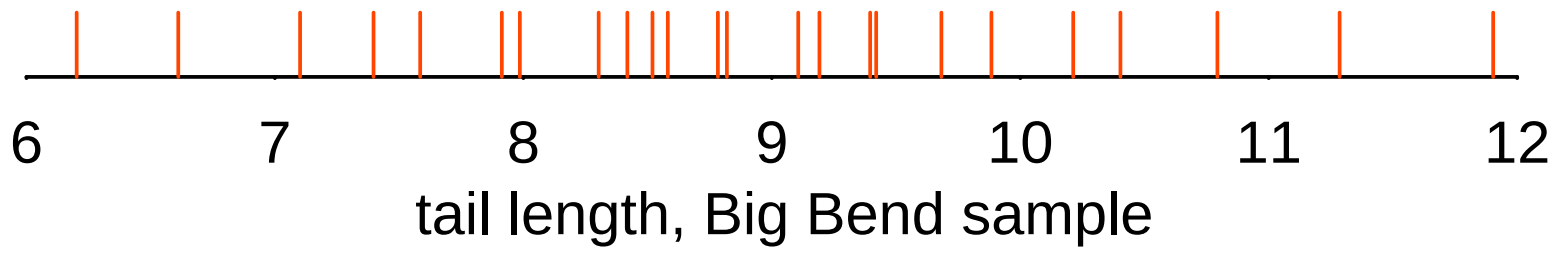
2. The observations $Y_1, \ldots, Y_n$ should be from a normal distribution if n is small, so that $\bar{Y}$ is approximately normal.

   How can we tell?

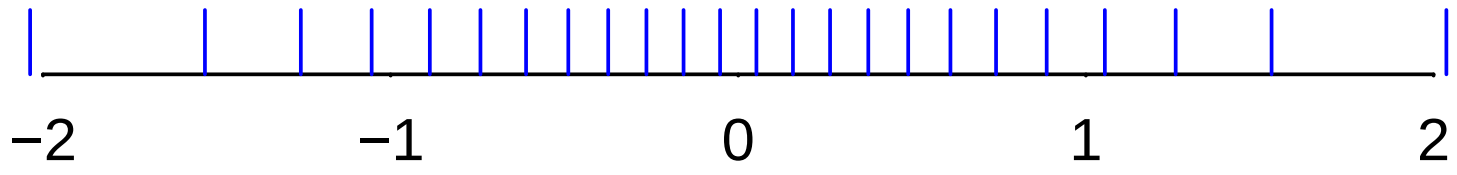
Detecting non-normality - Normal quantile plot

(section 13.1 in W&S)

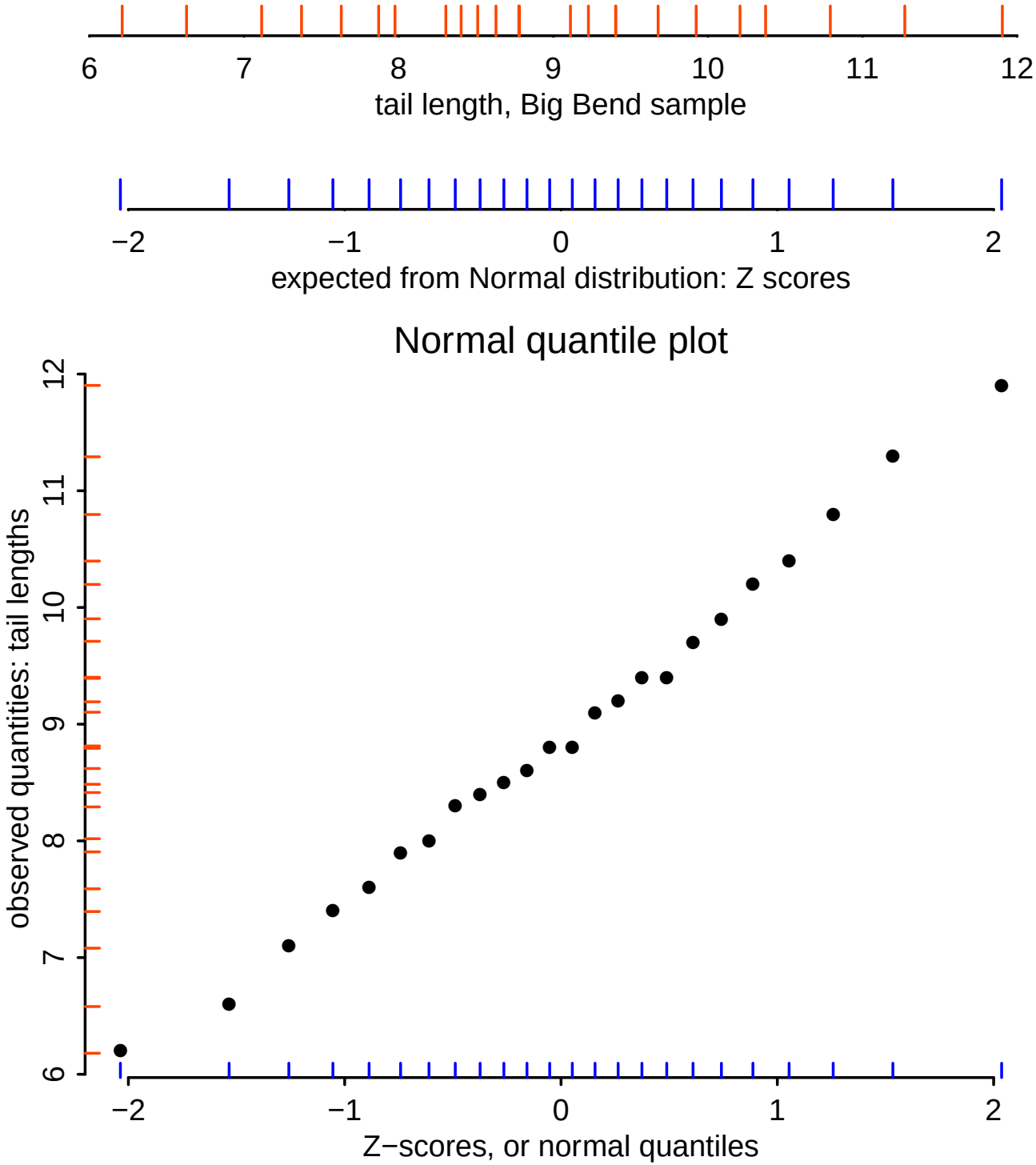
Compare spacing among observations with that expected from normal distribution:



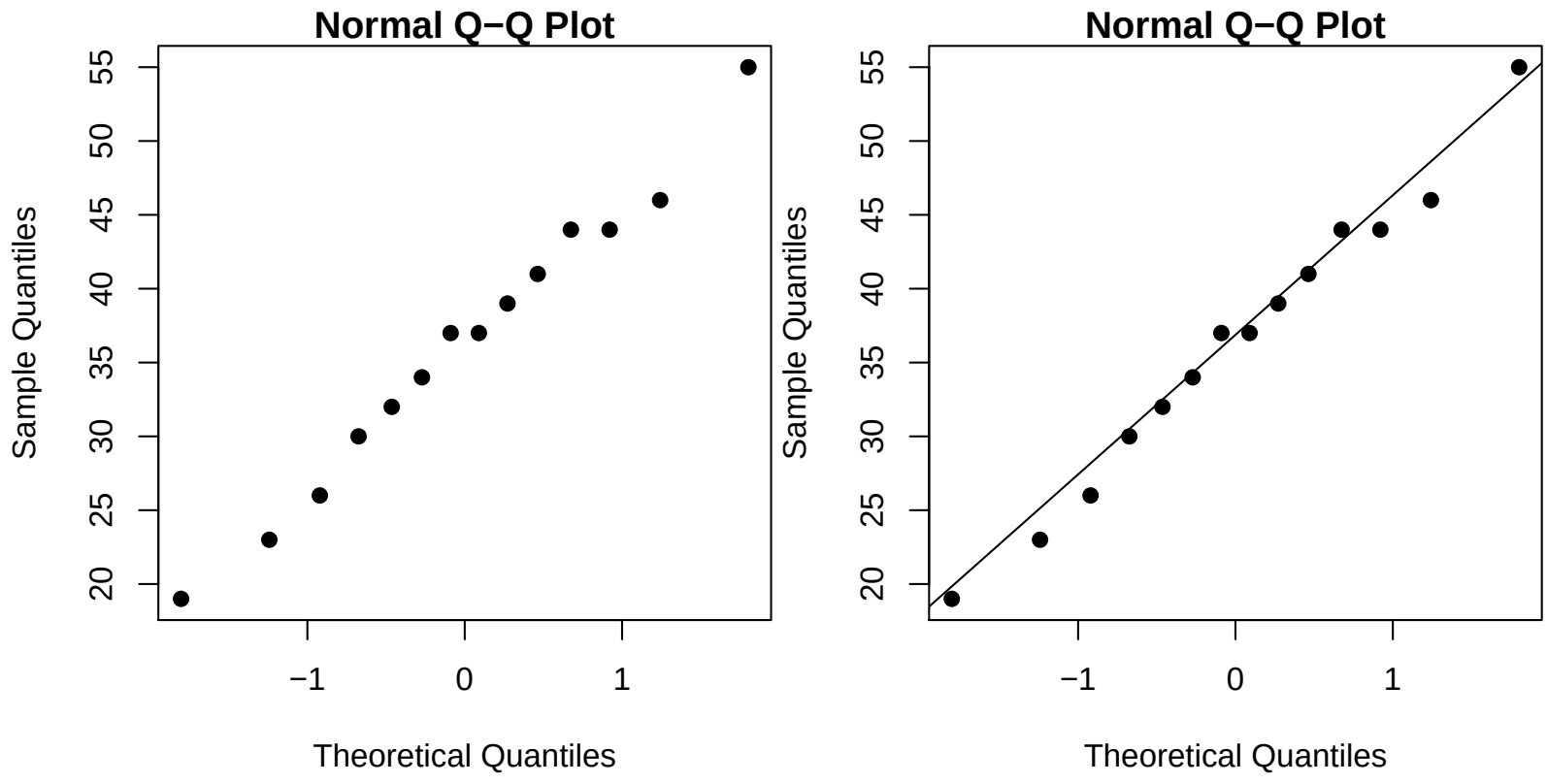
tail length, Big Bend sample



expected from Normal distribution: Z scores


6
7
8
9
10
11
12
tail length, Big Bend sample
−2
−1
0
1
2
expected from Normal distribution: Z scores
Normal quantile plot
observed quantities: tail lengths
6
7
8
9
10
11
12
−2
−1
0
1
2
Z−scores, or normal quantiles

Detecting non-normality - Normal quantile plot



If the points are close to a line, then we can say the data are normally distributed.

It is easier to tell with a normal quantile plot than with a histogram, even for small samples.

Detecting non-normality: qqnorm() - R demo

```
layout(matrix(1:2,1,2))
skewed.data = rgamma(10, shape=0.5)
hist(skewed.data, col="wheat")
qqnorm(skewed.data)

# non-linearity more pronounced with larger samples:
skewed.data = rgamma(30, shape=0.5)
hist(skewed.data, breaks=10, col="wheat")
qqnorm(skewed.data)

norm.dat = rnorm(10, mean=20)  # fake data
hist(norm.dat,col="wheat"); qqnorm(norm.dat)
# repeat to get a sense of how 'linear' the plot
# typically is for truly normal data

# less variation along the line with larger data sets:
norm.dat = rnorm(30, mean=20)  # fake data
hist(norm.dat,col="wheat"); qqnorm(norm.dat)
```

Outline

Paired vs. Independent samples

Treatments: A and B.

Paired samples: each observation on trt A is naturally paired with an observation on trt B. Related or same experimental units are used for both treatments.

Independent samples: no direct relationship between an observation on trt A and an observation on trt B.

Choice of paired versus independent sample is an important design issue. Data analysis follows the design.

Examples of two-sample comparisons

Compare tail length in 2 distant populations of the same lizard species

Compare taste of cheese from cows on two different diets (organic in the open vs. non-organic, hay/pellets)

Compare cholesterol level of patients before and after a drug treatment

Baby weight at birth among smoking/non-smoking women

When, why should samples be paired?

Cholesterol example:

1. Cholesterol level of 10 patients before and after a drug treatment.
2. Cholesterol level of 10 patients before treatment and of another 10 patients after treatment.

Baby weight example: pairing women according to certain traits. Effective only if it controls variability.

Paired sample studies usually preferred, because of increased precision (i.e. reduced variability) in estimating treatment differences.

If 3 or more treatments, blocking replaces pairing.

Paired samples - Blood pressure example

Question of interest: is there any evidence that a particular drug has an effect on blood pressure?

Experiment: on 15 middle-aged male hypertension patients. For each patient, blood pressure is measured at time of enrollment and again after 6 months of the drug treatment.

Blood pressure (mm Hg)

Subject	Before (Y_1)	After (Y_2)	Difference ($D = Y_1 - Y_2$)
1	90	88	2
2	100	92	8
3	92	82	10
4	96	90	6
5	96	78	18
6	96	86	10
7	92	88	4
8	98	72	26
9	102	84	18
10	94	102	-8
11	94	94	0
12	102	70	32
13	94	94	0
14	88	92	-4
15	104	94	10

$\mu_1 =$ population mean blood pressure before the drug trt
$\mu_2 =$ population mean blood pressure after the drug trt
$\mu_D = \mu_1 - \mu_2 =$ population mean of the difference

Confidence interval from paired samples

A $(1-\alpha)$ CI for the difference of means $\mu_1 - \mu_2 = \mu_D$ is

$$\bar{d} - t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}} \leq \mu_D \leq \bar{d} + t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}}$$

Assumptions:

- random sample of subjects (independence),
- D values have a normal distribution, or large sample size.

Check normal quantile plot of D.

No normality assumption about Y_1, or about Y_2.

Y_1 and Y_2 are not independent due to pairing: that's okay.

Confidence interval for the difference

From the D values of the $n = 15$ subjects:

$$\bar{d} = 8.80 \text{ mm Hg}, s_d = 10.98$$

t multiplier for 95% confidence: use t-distribution with df$= 15 - 1 = 14$: $t_{.025,14} = 2.145$.

Standard error of the mean: SE$= s_d/\sqrt{15} = 2.835$ mmHg

interval:

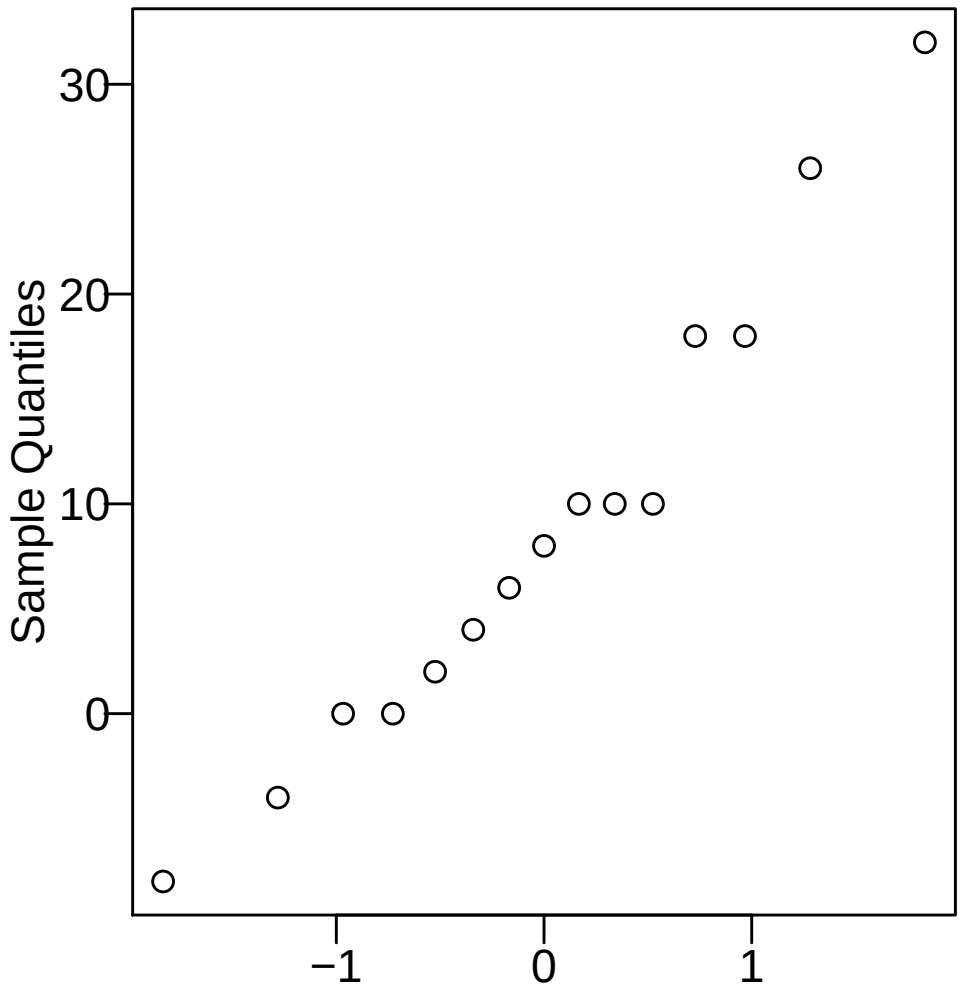
$$\leq \mu_D \leq$$

We are 95% confident that the population mean decrease in blood pressure after 6 months of treatment lies between 2.72 and 14.88 mm Hg (or 8.80 ± 6.08).

Checking the normality assumption

```
bpbefore =c(90,100,92,96,96,96,92,98,102,94,94,102,94,88,104)
bpafter = c(88, 92,82,90,78,86,88,72,84,102,94, 70,94,92, 94)
bpdiff = bpbefore - bpafter
qqnorm(bpdiff, main="Normal quantile plot for D values")
```

Normal quantile plot for D values



The t-test for paired samples

$D = Y_1 - Y_2$ is the blood pressure difference.

Paired samples

Testing $\mu_1 = \mu_2$ or $\mu_1 \neq \mu_2$ is equivalent to testing

$$H_0 : \mu_D = 0 \quad \text{vs} \quad H_A : \mu_D \neq 0.$$

A one-sample t-test can be used on the differences:

$$T = \frac{\bar{D} - 0}{S_D/\sqrt{n}}$$

If H_0 is true, $T \sim$ t-distribution on df$= n - 1 =$ # pairs -1.

Same assumptions as CI:

random sample of n subjects,

normal distribution for D, or large sample size.

The t-test on blood pressure

1. H_0: $\mu_D = 0$ and 2-side test H_A: $\mu_D \neq 0$.
2. If H_0 is true, $T = \dfrac{\bar{D} - 0}{S_D/\sqrt{n}}$ has a t-distribution on df =
3. We observed $\bar{d} = 8.80$ mm Hg, $s_d = 10.98$ and SE$= 10.98/\sqrt{15} = 2.835$.
   The observed t-value is

   $$t = \qquad\qquad = 3.10$$

   As extreme: $T = -3.10$ or 3.10, more extreme: $T > 3.10$ or $T < -3.10$. The p-value is $2\mathbb{P}\{T_{14\text{ df}} \geq 3.10\}$, which is between 0.002 and 0.01 from Table C.
4. There is strong evidence against H_0: the drug is deemed beneficial.

Statistical significance vs. biological importance

Warning: This t-test of H_0: $\mu_D = 0$ tells us only about statistical significance.

The confidence interval tells us also about biological importance: average improvement between 2.72mmHg and 14.88.

If a difference of 5 mm Hg is needed for biological significance, we could test $H_0 : \mu_D = 5$ vs. $H_A : \mu_D \neq 5$. For this, use $T = \frac{\bar{D}-5}{S_D/\sqrt{n}}$.

Direct relationship between CI and t-test

0 mmHg outside the 95% CI $\leftrightarrow$ "$\mu_D = 0$" is rejected at $\alpha = 0.05$.
5 mmHg outside the 95% CI $\leftrightarrow$ "$\mu_D = 5$" is rejected at $\alpha = 0.05$.

R commands: t.test()

```
> # first enter the data
> bpbefore =c(90,100,92,96,96,96,...,102,94,88,104)
> bpafter = c(88, 92,82,90,78,86,..., 70,94,92, 94)
>
> # Now do the paired t-test and 95% CI
> t.test( bpbefore - bpafter )

        One Sample t-test

data:  bpbefore - bpafter
t = 3.1054, df = 14, p-value = 0.00775
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  2.722083 14.877917
sample estimates:
mean of x
      8.8
```

R commands: t.test()

Or use t.test with both sets of original values (before & after), and the option `paired=TRUE`:

```
> t.test(bpbefore, bpafter, paired=TRUE)

        Paired t-test

data:  bpbefore and bpafter
t = 3.1054, df = 14, p-value = 0.00775
alternative hypothesis: true difference in means is not
95 percent confidence interval:                  equal to 0
  2.722083 14.877917
sample estimates:
mean of the differences
                    8.8
```

What if the normality assumption is not met?

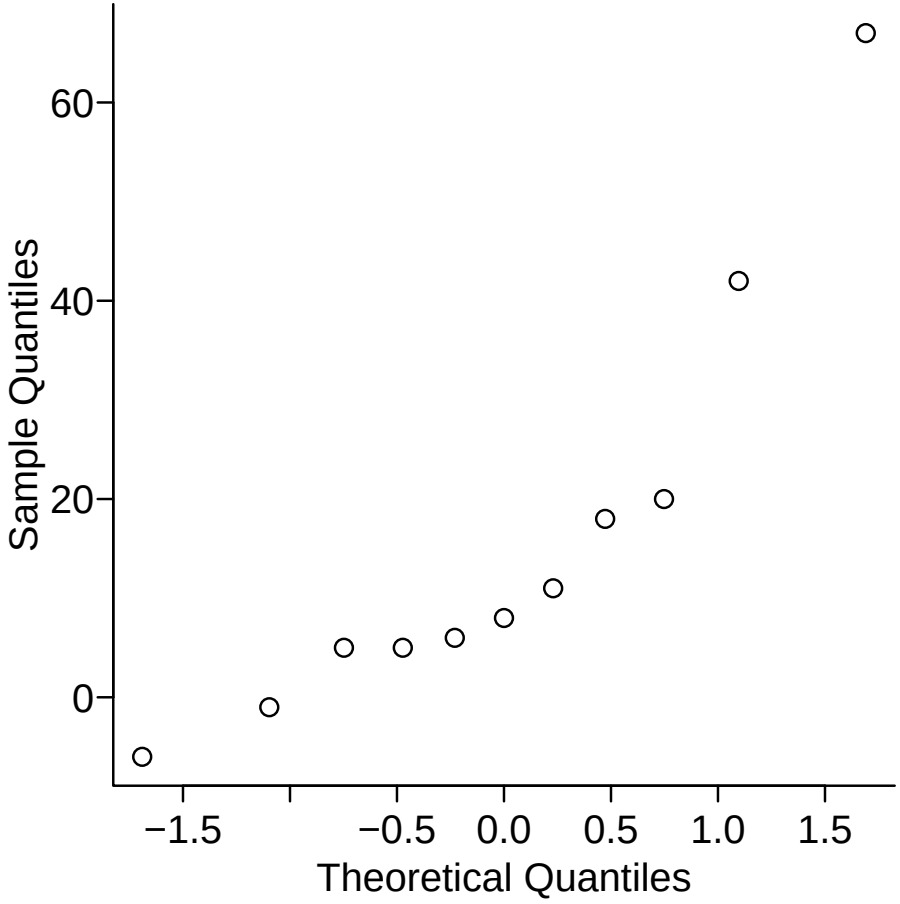
Skin graft: skin from cadavers can provide temporary skin grafts for severely burned patients. The longer the graft survives before its inevitable rejection, the more the patient benefits. Investigate the usefulness of matching graft to patient w.r.t. HL-A antigen system. Each received 2 grafts: one with close HL-A compatibility, the other with poor compatibility. Graft survival times in days:

Patient	1	2	3	4	5	6	7	8	9	10	11
close: Y_1	37	19	≥ 57	93	16	23	20	63	29	≥ 60	18
poor: Y_2	29	13	15	26	11	18	26	43	18	42	19
d: $Y_1 - Y_2$	8	6	42+	67	5	5	-6	20	11	18+	-1

2 incomplete observations: patient 3 died before one graft was rejected, unspecified reason for patien 10.

Problem: incomplete data & non-normality

Normal quantile plot for
incomplete d values:



signs of d well determined:

+ + + + + + − + + + −

9 of 11 patients had longer time with graft of close HL-A compatibility. Is this difference (9 vs. 2) significant?

The sign test

Idea: simply look at the signs of differences, i.e. at which treatment worked best (regardless of how much better) for each patient.

H_0: median of D is 0, i.e.
$\mathbb{P}\{\text{close HL-A comp. is better}\} =$
$\mathbb{P}\{\text{poor HL-A comp. is better}\} = 0.5.$

Here a one-sided alternative is appropriate (we already know about HL-A compatibility):

H_A: median of D is positive, i.e.
$\mathbb{P}\{D > 0\} = \mathbb{P}\{\text{close HL-A comp. is better}\} > 0.50.$

The sign test: binomial test on # of > 0 differences

Test statistic: $Y_+ =$ # of $+$ signs. 9 here, out of 11.

If H_0 is true, for each subject there is a 50% chance to observe $+$. So

$$Y_+ \sim \mathcal{B}(n, 0.5) \qquad \text{if } H_0 \text{ is true}$$

p-value: More extreme than $Y_+ = 9$ (or as extreme as) means $Y_- = 9, 10, 11$ (one-sided test)

$$\text{p-value} = \mathbb{P}\{Y_+ \geq 9\} = 0.0327$$

```
sum(dbinom(9:11, size=11, prob=.5))
```

Conclusion: Reject H_0. There is moderate evidence that the skin grafts tend to last longer when the HL-A compatibility is close than when it is poor.

The sign test: What if there are ties?

Tie: $y_1 = y_2$ for some subject. Then $d = 0$ for this subject. No sign!!

Exclude all zeros, and decrease the sample size accordingly.

Example: If the differences were

Patient	1	2	3	4	5	6	7	8	9	10	11
close: Y_1	37	13	≥ 57	93	16	23	20	63	29	≥ 60	18
poor: Y_2	29	13	15	26	11	18	26	43	18	42	19
d: $Y_1 - Y_2$	8	0	42+	67	5	5	-6	20	11	18+	-1
sign	$+$	0	$+$	$+$	$+$	$+$	$-$	$+$	$+$	$+$	$-$

$\mathbf{Y}_- = \mathbf{2}$ as before, but $\mathbf{Y}_+ = \mathbf{8}$ and $\mathbf{n} = 2 + 8 = 10$, not 11.

p-value $= \mathbb{P}\{Y_+ \geq 8 | n = 10\} = 0.0547$.

```
sum(dbinom(8:10, size=10, prob=.5))
```

R: sign(), table() and binom.test()

```
> y.close = c(37,19,57,93,16,23,20,63,29,60,18)
> y.poor  = c(29,13,15,26,11,18,26,43,18,42,19)
> ydiff = y.close - y.poor

> sign(ydiff)
 [1]  1  1  1  1  1  1 -1  1  1  1 -1

> table(sign(ydiff))

-1  1
 2  9

> binom.test(9, 9+2, alternative="greater")
        Exact binomial test

data:  9 and 9 + 2
number of successes = 9, number of trials = 11, p-value = 0.03271
alternative hypothesis: true prob. of success is greater than 0.5
95 percent confidence interval:
 0.5299132 1.0000000
sample estimates:
probability of success
             0.8181818
```

Limitations and Wilcoxon signed-rank test

The **sign test** has

- low power: less powerful than the t-test
- but does not assume anything other than independence (random sample). No distribution assumption.

Alternative: the **Wilcoxon signed-rank** test. Uses the signs and part of the magnitude information.

- More powerful than the sign test, but assumes a symmetric distribution.
- Less powerful than the t-test, but does not assume the normal distribution

```
> wilcox.test(y.close, y.poor, paired=T, alternative="greater")
        Wilcoxon signed rank test with continuity correction

data:  y.close and y.poor
V = 60.5, p-value = 0.008131
alternative hypothesis: true location shift is greater than 0
```