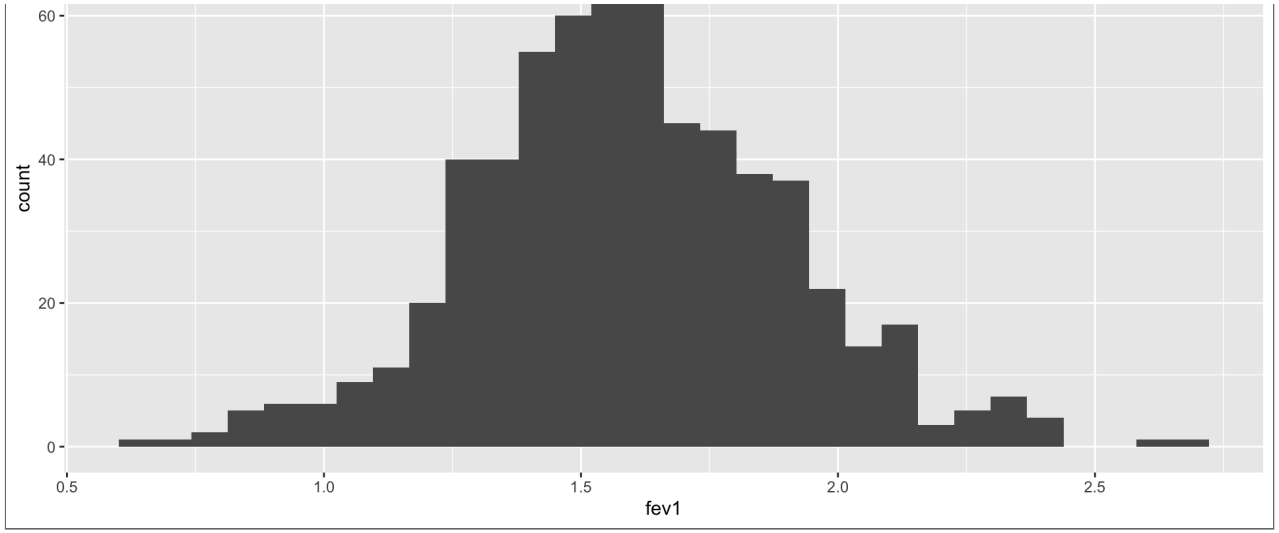
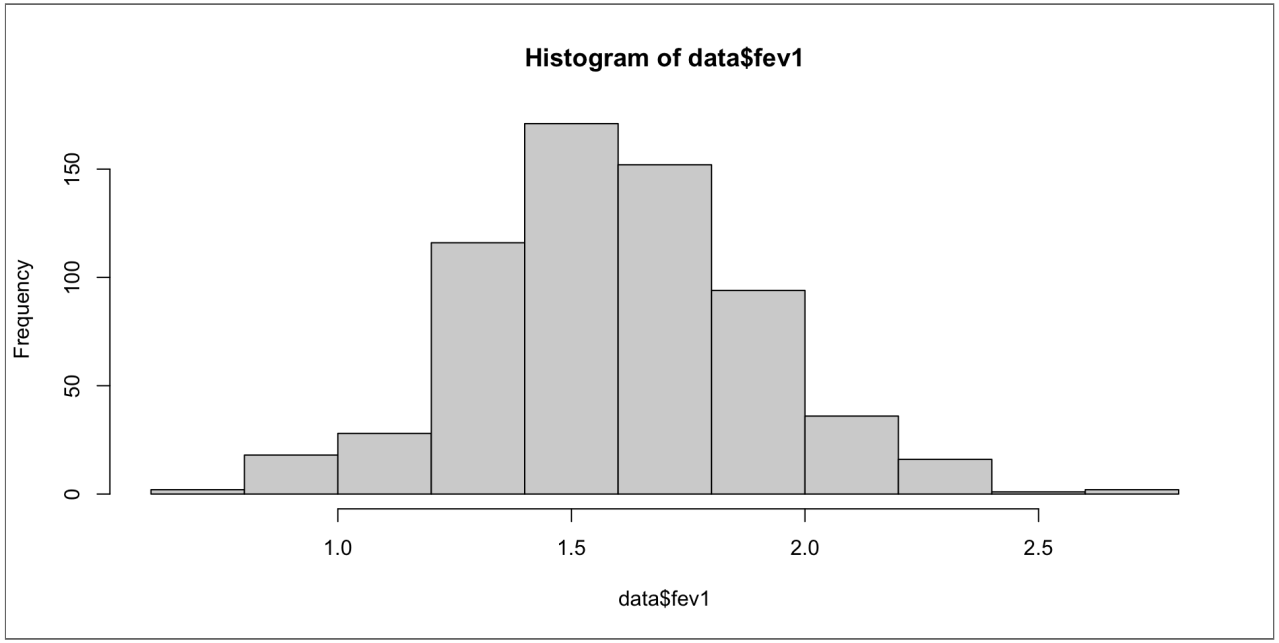
Y


60
40
20
0
count
0.5
1.0
1.5
2.0
2.5
fev1

```r
hist(data$fev1)
```

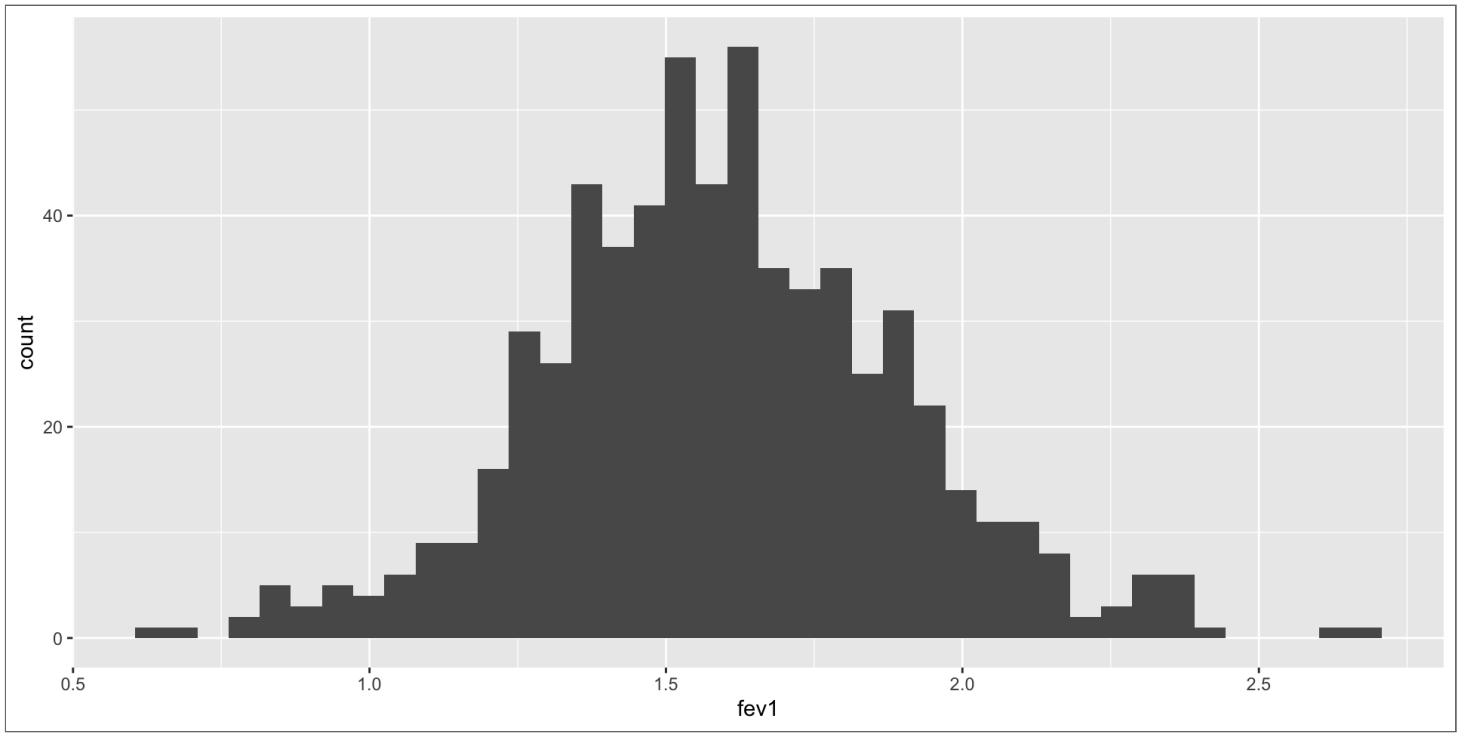


The histogram

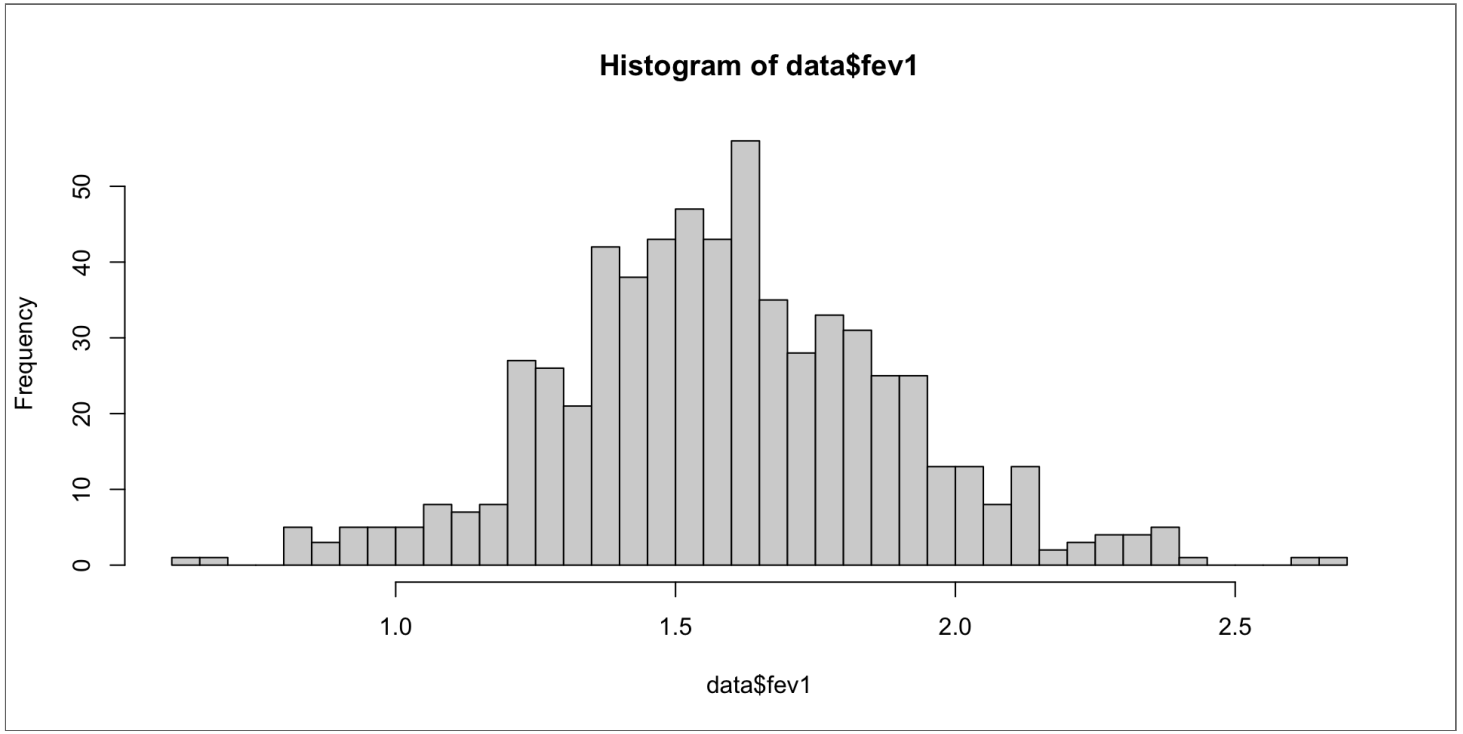
Setting the number of bins

with ggplot2 | **in base R**

```
1 ggplot(data, aes(x=fev1)) +
2   geom_histogram(bins=40)
```



```r
hist(data$fev1, breaks=40)
```

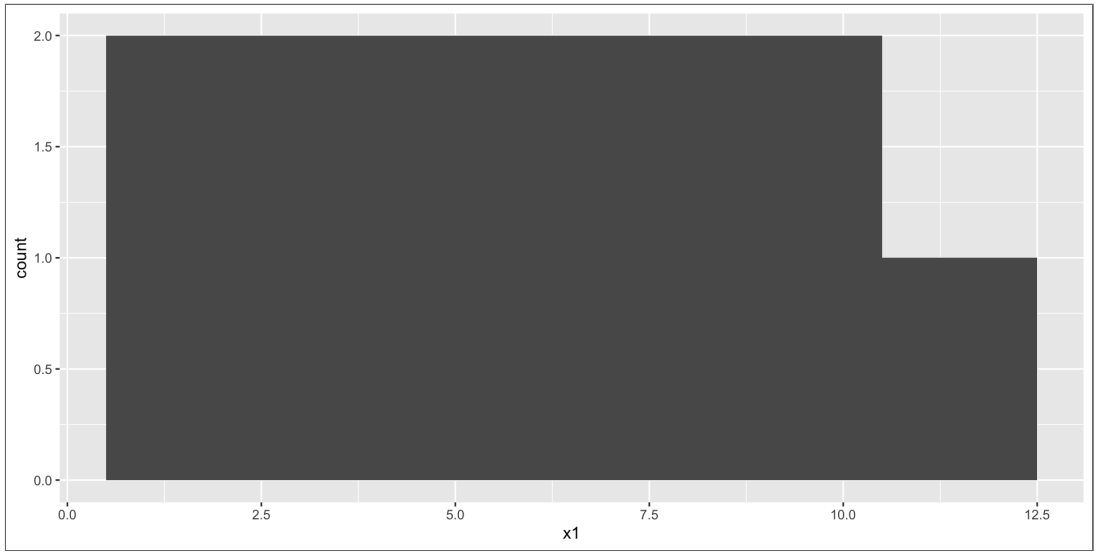


The histogram

The histogram is sensitive to the choice of bins

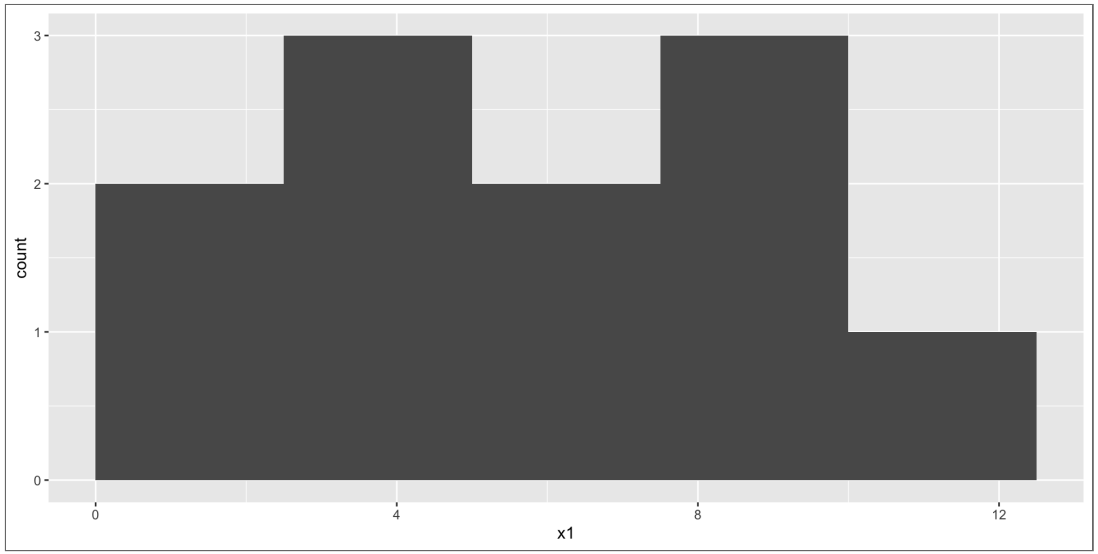
binwidth=2, start at 0.5 | **binwidth=2.5, start at 0**

```
1  toy_dat←data.frame(x1=1:11)
2  breaks←seq(from=0.5,to=12.5,by=2)
3  ggplot(toy_dat, aes(x=x1)) +
4    geom_histogram(breaks=breaks)
```



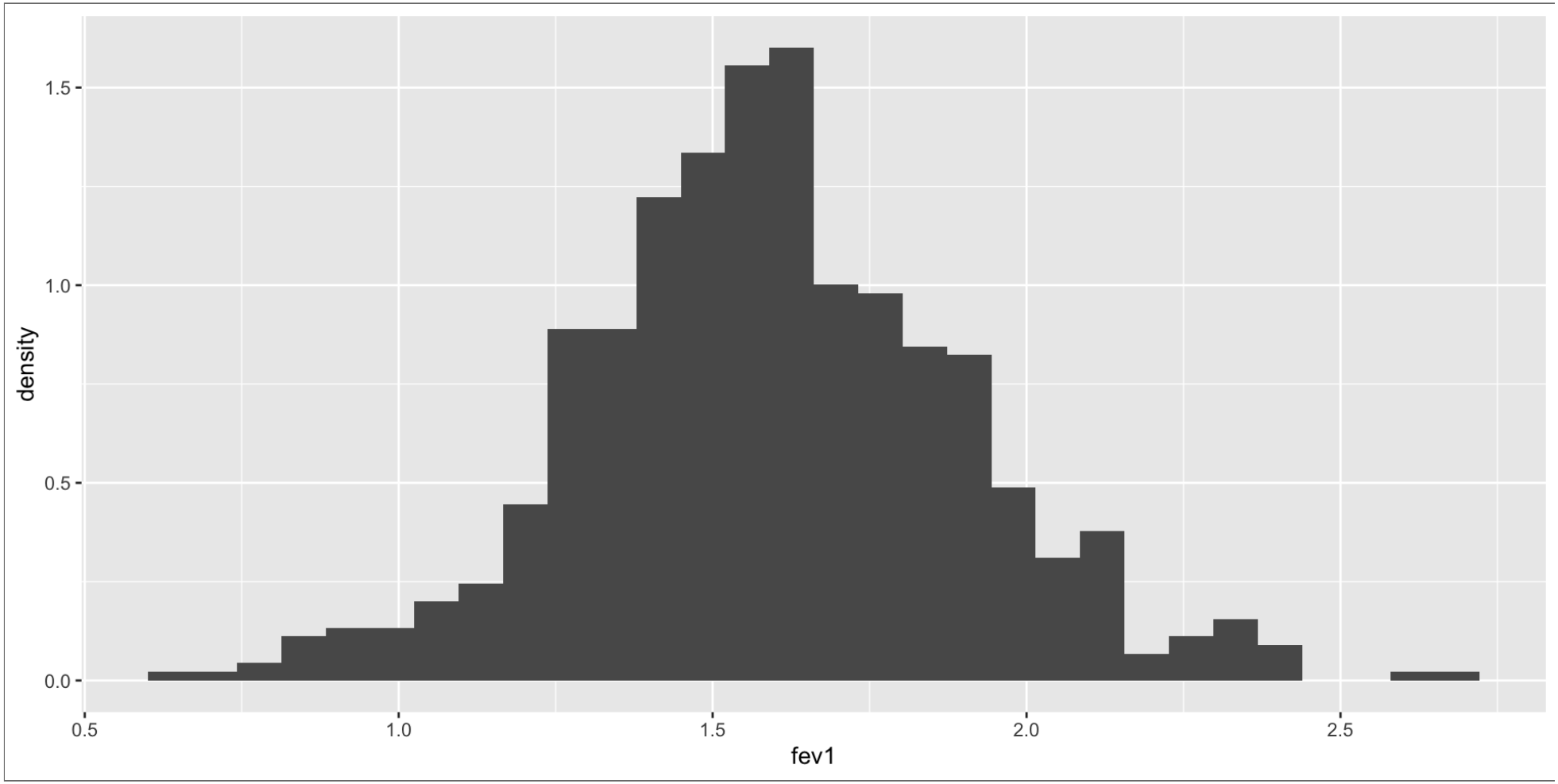

```
1  toy_dat←data.frame(x1=1:11)
2  breaks←seq(from=0,to=12.5,by=2.5)
```

```
3 ggplot(toy_dat, aes(x=x1)) +
4   geom_histogram(breaks=breaks)
```



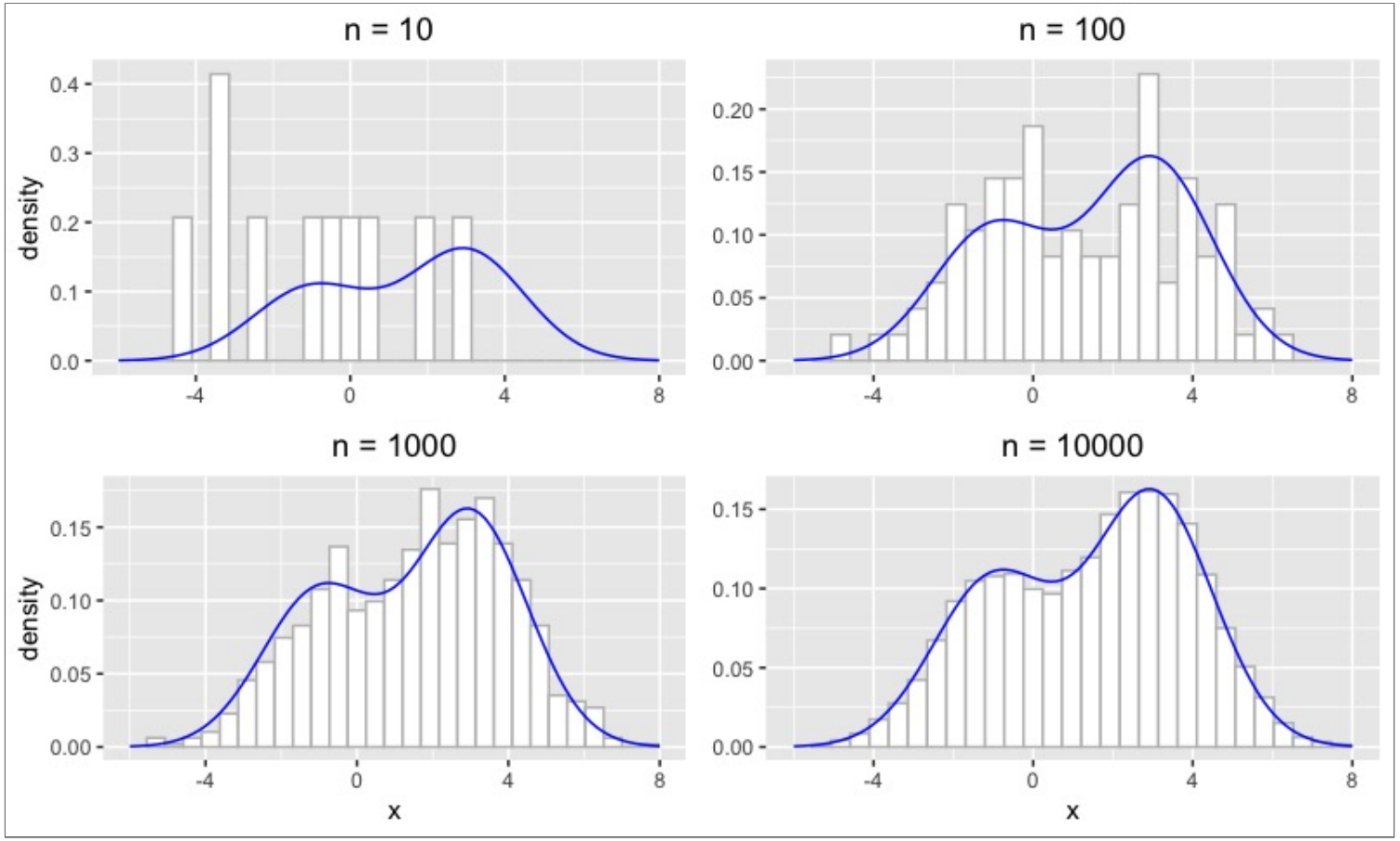
The histogram as a density estimator

```
ggplot(data, aes(x=fev1)) +
  geom_histogram(aes(y=after_stat(density)))
```


1.5
1.0
0.5
0.0
density
0.5
1.0
1.5
2.0
2.5
fev1

The histogram as a density estimator

The histogram approaches the true distribution (probability density function) as the sample size increases.



Descriptive statistics

A **statistic** = a quantity calculated from the data
A **descriptive statistic** = a statistic summarizing features of the sample, typically features of the distribution of a variable
Descriptive statistics = the process of describing the sample using such summary statistics
IMPORTANT: describing the **sample** … as opposed to the **population**

Descriptive statistics

An easy way to obtain some important descriptive statistics on your variables.

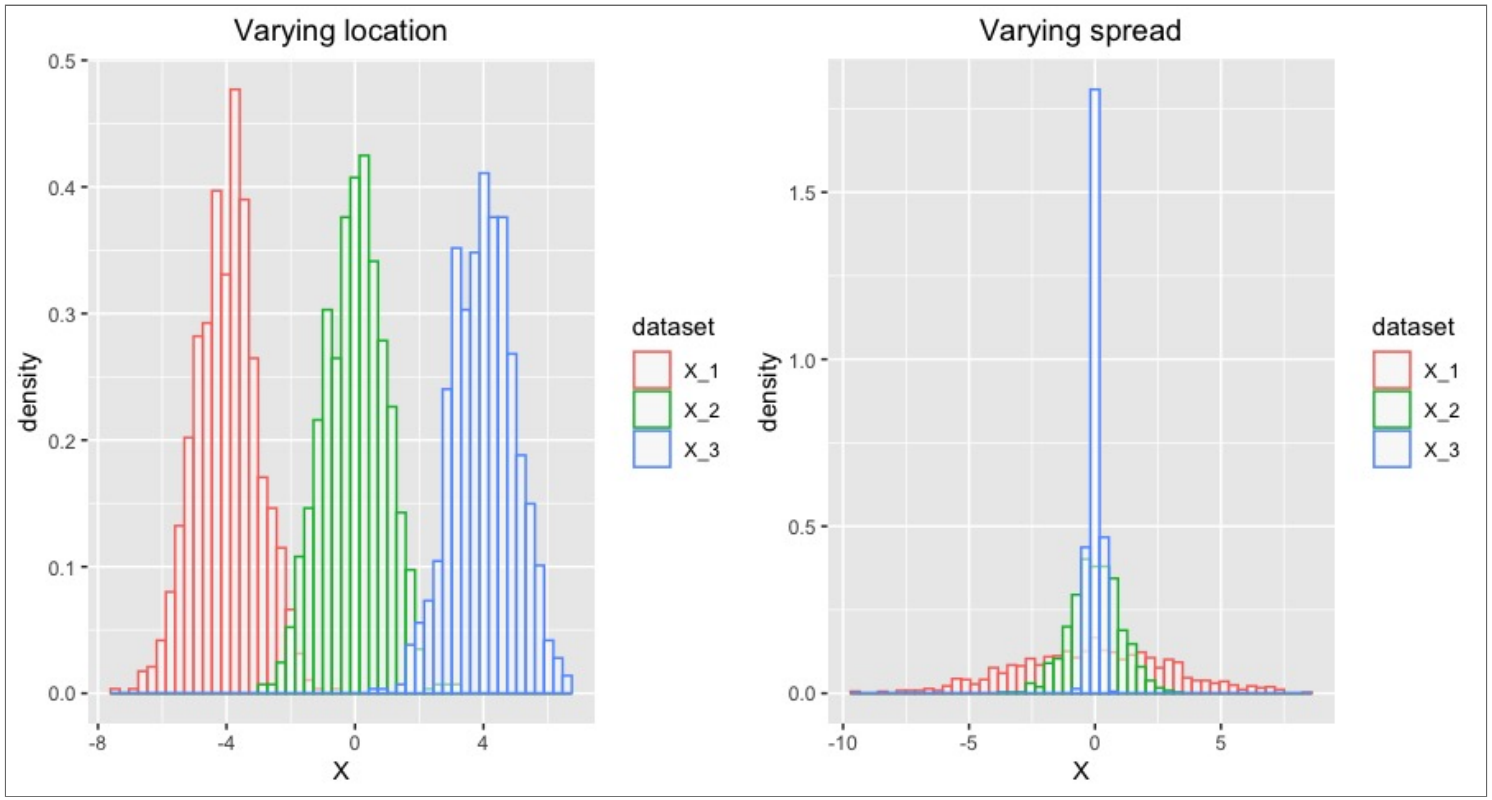
```
1 summary(data)
```

```
      id              fev1            age              height      sex
Min.   :  1.0   Min.   :0.640   Min.   : 7.116   Min.   :105.6   f:335
1st Qu.:160.8   1st Qu.:1.397   1st Qu.: 8.493   1st Qu.:119.9   m:301
Median :319.5   Median :1.580   Median : 8.909   Median :124.0
Mean   :319.8   Mean   :1.595   Mean   : 8.984   Mean   :124.1
3rd Qu.:479.2   3rd Qu.:1.790   3rd Qu.: 9.627   3rd Qu.:128.0
Max.   :638.0   Max.   :2.690   Max.   :10.440   Max.   :149.0
respsymptoms          asthma_hist
no :491      current asthma : 59
yes:145      never          :450
             previous asthma:127
```

Note: There are no missing values in this data set. These would appear in the lowest row as NA: number of missings.

Location and spread

Location: Measures where on the x-axis an average value would lie (central tendency).
Spread: Measures how widely values vary.



Sample mean

To compute the sample mean simply

- Sum up all values
- Divide by n

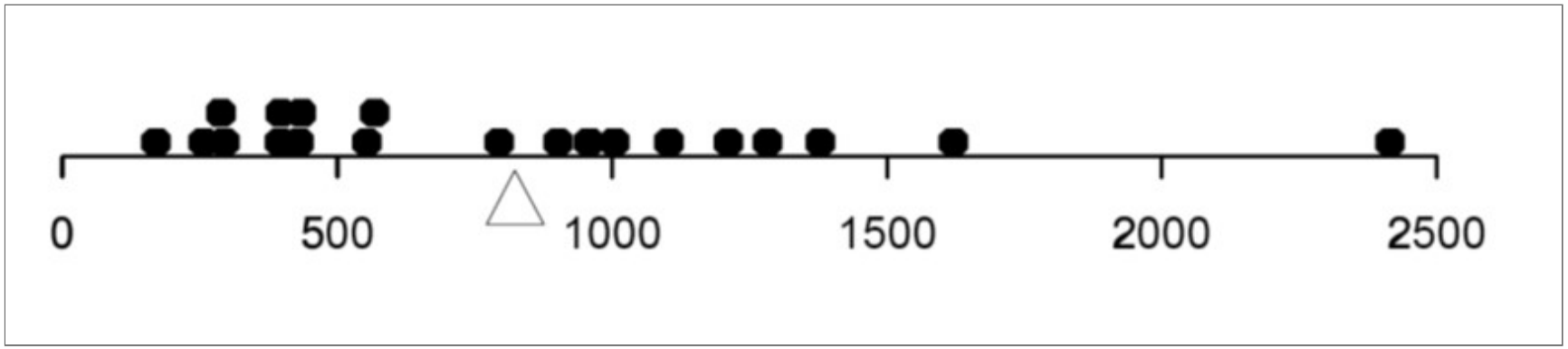
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \ldots + x_n}{n}$$

Whats the mean of these values? 6, 8, 11, 3, 5, 6

Sample mean

The sample mean “balances out” the data:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$



The median

The median separates the lower half from the higher half of the data sample

To calculate, first order the values: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

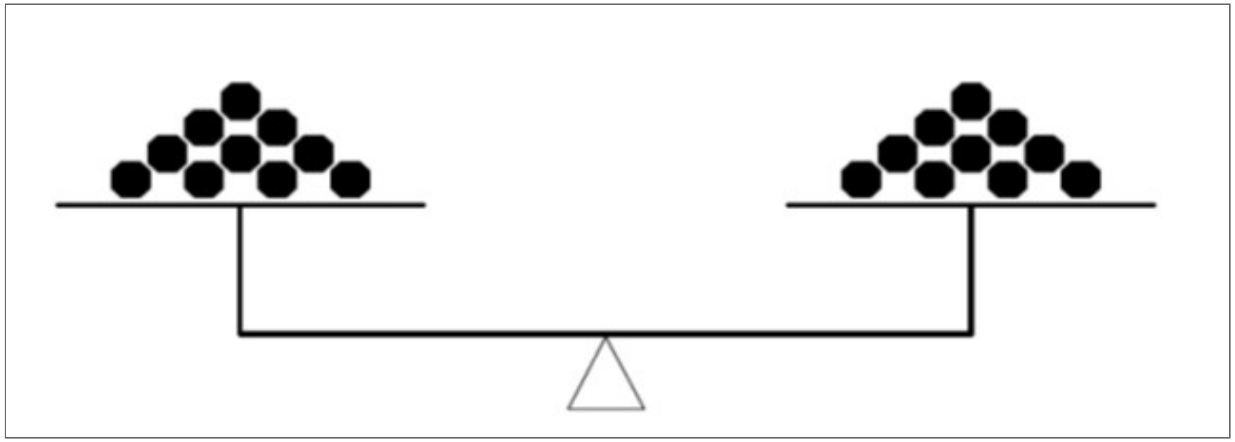
The median is given by:

$$x_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd,} \\ \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right) & \text{if } n \text{ is even.} \end{cases}$$

What's the median of these values? 6, 8, 11, 3, 5, 6

Der Median

The median “balances out” the data on a beam balance.



The median is a *robust* measure of central tendency because it is not affected by outliers.

Sensitivity to outliers

Here some toy data:

```
toy_dat←tibble(x1=1:11,
               x2=c(1:10, 100))
toy_dat
```

```
# A tibble: 11 × 2
      x1    x2
   <int> <dbl>
 1     1     1
 2     2     2
 3     3     3
 4     4     4
 5     5     5
 6     6     6
 7     7     7
 8     8     8
 9     9     9
10    10    10
11    11   100
```

The mean is strongly affected by the outlier in x_2.

The median remains the same

```r
summary(toy_dat)
```

```
       x1              x2          
 Min.   : 1.0   Min.   :  1.00  
 1st Qu.: 3.5   1st Qu.:  3.50  
 Median : 6.0   Median :  6.00  
 Mean   : 6.0   Mean   : 14.09  
 3rd Qu.: 8.5   3rd Qu.:  8.50  
 Max.   :11.0   Max.   :100.00  
```

Skewed and symmetric distributions

The median will differ from the mean in the case of skewed distribution

The empirical CDF

What proportion of values are less or equal to a given value t?
The empirical cumulative distribution function (ECDF):

$$F_n(t) = \frac{\#\{x_i \leq t\}}{n}$$

```
1 ggplot(data, aes(fev1)) +  stat_ecdf()
```

The ECDF and quantiles

We can use ECDF to obtain quantiles

Quartile

Quartiles divide the values into 4 (almost) equally sized groups

- $x_{0.25}$: lower or first quartile,
- $x_{0.5}$: median or second quartile,
- $x_{0.75}$: upper or third quartile

```
quantile(data$fev1, c(0.25,0.5,0.75))
```

```
    25%    50%    75%
1.3975 1.5800 1.7900
```

Note: There is some ambiguity involved, see different methods for calculating quantiles **here**. Use the option “type” to select a method.

Measuring dispersion

Simple idea: Let's take the average of the deviations from the mean. But these sum to zero.

So let's take the absolute deviations, resulting in the mean absolute deviation (MAD):

$$\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{x}\right|$$

… or the squared deviations, resulting in the mean sum of squares:

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

Sample variance

The latter is more commonly used and is related to the concept of variance in probability theory.

Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

alternative notation: s^2, s_x^2, Var, Var_x, VAR

Sample standard deviation

To get back to the original scale of the variable x we need to take the square root.

Standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

Alternative notation: s_x, SD, SD_x

Why divide by $n-1$?

The sample mean is closer to the data points than the true mean would be.

The mean sum of squares underestimates the true variance. Dividing by $n-1$ (*degrees of freedom*) corrects for the bias.

Degrees of freedom (df)

- Relative to the true mean μ all n deviations $(x_i - \mu)$ are free to vary during sampling ($\mathrm{df} = n$).
- Relative to the sample mean $\overline{x}$ only $n - 1$ deviations $(x_i - \overline{x})$ are free to vary ($\mathrm{df} = n - 1$).

  We “use up”” one df by calculating $\overline{x}$. More generally, one df is lost for every fitted parameter used to calculate the mean (relevant in regression modelling).

Other measures of dispersion

Range: difference between highest (maximum) and lowest value (minimum)

$\text{Range} = x_{(n)} - x_{(1)}$

Interquartile range (IQR): Difference between upper and lower quartile (includes 50% of the observations).

$\text{IQR} = x_{0.75} - x_{0.25}$

The box plot

The box plot

```
1 ggplot(data, aes(y=fev1, color=sex)) +
2   geom_boxplot() +
3   scale_x_discrete()
```

All in one publication type table

Using the package `gtsummary`

```
library(gtsummary)
tabl_summary ← data ▷
  select(where(is.double)) ▷
  tbl_summary(
    type = all_continuous() ~ "continuous2",
    statistic = all_continuous() ~ c("{mean} ({sd})",
                                     "{median} ({p25}, {p75})",
                                     "{min}, {max}")
  )
```

Note that IQR and range are mostly reported as intervals rather than the length of these intervals.

```
tabl_summary
```

Characteristic	N = 636
fev1	
Mean (SD)	1.59 (0.30)
Median (IQR)	1.58 (1.40, 1.79)
Range	0.64, 2.69

Characteristic	N = 636
age	
Mean (SD)	8.98 (0.72)
Median (IQR)	8.91 (8.49, 9.63)
Range	7.12, 10.44
height	
Mean (SD)	124 (6)
Median (IQR)	124 (120, 128)
Range	106, 149

Standardizing variables

If the values $x_1, x_2, \ldots, x_n$ have mean $\overline{x}$ and standard deviation s_x
then the standardized values z_i

$$z_i = \frac{x_i - \overline{x}}{s_x}$$

have sample mean $\overline{z} = 0$ and standard deviation $s_z = 1$.

Normal distribution

We often use the normal distribution to model the distribution of continuous variables.

Normal distribution

If X follows a normal distribution distribution with mean μ and variance σ^2 its probability density function (pdf) $f(x)$ is given by

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

The standardized variable $Z = \frac{X-\mu}{\sigma}$ then follows a **standard normal distribution** with mean 0 and variance 1.

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

Normal distribution

To calculate the probability that $X < a$ we need the area under the standard normal density below the value $\frac{a-\mu}{\sigma}$.

Example

Body height of men

Example

Probability that a randomly selected man is taller than 180cm:

$$\Pr(X > 180) = 1 - \Pr(X \leq 180) = 1 - \Pr\left(Z \leq \frac{180 - 171.5}{6.5}\right)$$

We can use the cumulative distribution function `pnorm` in R

```r
# Using z transformation and standard normal
z←(180-171.5)/(6.5)
1-pnorm(z)
```

```
[1] 0.09548885
```

```r
# or
pnorm(z, lower.tail = FALSE)
```

```
[1] 0.09548885
```

```r
# or directly using the distribution of X
1-pnorm(180, mean=171.5, sd=6.5)
```

```
[1] 0.09548885
```

Remember 1.96

Some commonly used quantiles of the standard normal distribution (you will shortly see why).

```
1 qnorm(c(0.95, 0.975, 0.995))
```

```
[1] 1.644854 1.959964 2.575829
```

Inference about the mean

Population and samples

Inferential statistics

Inferential statistics = Inferring properties of the population based on the sample

Parameteric statistics: We make specific (parametric) assumptions about the true distribution of a variable. Example:

- Assumptions: FEV_1 is normally distributed within children of the same sex (i) and age (j) with means $\mu_{i,j}$ and variances $\sigma_{i,j}$.
- Inference: We want to *estimate* the means $\mu_{i,j}$ and *test* whether these differ between sexes.

Non-parametric statistics: We make no parametric assumptions about the true distribution. Example:

- Using the histogram to *estimate* probability density function

What’s the problem?

We never observe the full population, only 1 sample. Any inference about the population is subject to error.
Sampling variation: Chance of variation between samples.
Sampling distribution: The distribution of a statistic across different samples.

Accuracy and precision

Properties of good estimators:

- **Precision**: Low variance
- **Accuracy**: Low bias

Source: https://doi.org/10.3390/app11052191

Simulating the sampling distribution

Simulation steps:

1. Specify a distribution for X
2. Draw a sample of size n: $x_1, \ldots, x_n$
3. Calculate the mean: $\overline{x}$
4. Repeat steps 2 and 3 many times
5. Plot a histogram of the resulting means

Underlying distribution bimodal

Histograms of means for increasing sample sizes

Underlying distribution skewed

Underlying distribution uniform

Observations

Regardless of the underlying distribution, the sampling distribution becomes more and more …

- narrow
- unimodel
- symmetrical
- focused around the true mean

Note: The sample mean is unbiased estimator. Its sampling distribution is always centered around the mean regardless of sample size. The last point is a visual impression due to the narrowing of the curve.

The sample mean is unbiased

Denote the population mean as μ. Then

$$
\begin{aligned}
E(\overline{x}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}\cdot n \cdot \mu = \mu
\end{aligned}
$$

Standard error of the mean

The standard deviation of the sampling distribution is called the standard error.
Assuming that the observations x_i are independent, the **standard error of the mean** is

$$SE = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the underlying distribution.
As σ is usually unknown, we estimate it using the sample standard deviation s

$$\hat{SE} = \frac{s}{\sqrt{n}}$$

Central limit theorem (CLT)

As sample size increases, the sampling distribution of the mean approaches a normal distribution.
Specifically, the distribution of

$$z = \frac{\overline{x} - \mu}{SE}$$

approaches the *standard normal distribution*. Here μ is the mean of the underlying distribution, i.e. the true mean.

Central limit theorem (CLT)

Histograms of standardized means z for increasing sample sizes

FEV_1 by age and sex

```r
data$age_cat←cut(data$age,seq(6,12,2), right=FALSE)
tbl_fev1 ← data  %>%
  group_by(sex, age_cat) %>%
  summarise(n=n(), mean=mean(fev1), sd=sd(fev1))
tbl_fev1
```

```
# A tibble: 6 × 5
# Groups:   sex [2]
  sex   age_cat     n  mean    sd
  <fct> <fct>   <int> <dbl> <dbl>
1 f     [6,8)      29  1.30 0.232
2 f     [8,10)    275  1.53 0.267
3 f     [10,12)    31  1.83 0.300
4 m     [6,8)      24  1.35 0.257
5 m     [8,10)    254  1.67 0.296
6 m     [10,12)    23  1.85 0.256
```

FEV_1 by age and sex

```
ggplot(data, aes(x=age_cat, y=fev1, fill=sex)) +
  geom_boxplot() +
  xlab("Age group") +
  ylab("FEV1") +
  facet_wrap(~sex)
```

95%-Confidence intervals

Assuming we have a large sample, we can use the CLT to construct an approximate 95%-confidence interval (95%-CI) for the mean.

$[\overline{x} - 1.96 \cdot \hat{SE},\ \overline{x} + 1.96 \cdot \hat{SE}]$

Defining feature: A 95%-CI is defined such that, under repeated sampling, 95% of the intervals are expected to contain the true mean.

95%-Confidence intervals

```r
tbl_fev1$SE <-tbl_fev1$sd/sqrt(tbl_fev1$n)
tbl_fev1$CI_norm_lb<-tbl_fev1$mean - qnorm(0.975)*tbl_fev1$SE
tbl_fev1$CI_norm_ub<-tbl_fev1$mean + qnorm(0.975)*tbl_fev1$SE
print(tbl_fev1, digits=3)
```

```
# A tibble: 6 × 8
# Groups:   sex [2]
  sex   age_cat     n  mean    sd     SE CI_norm_lb CI_norm_ub
  <fct> <fct>   <int> <dbl> <dbl>  <dbl>      <dbl>      <dbl>
1 f     [6,8)      29  1.30 0.232 0.0431       1.21       1.38
2 f     [8,10)    275  1.53 0.267 0.0161       1.50       1.56
3 f     [10,12)    31  1.83 0.300 0.0538       1.73       1.94
4 m     [6,8)      24  1.35 0.257 0.0525       1.25       1.45
5 m     [8,10)    254  1.67 0.296 0.0186       1.63       1.71
6 m     [10,12)    23  1.85 0.256 0.0534       1.75       1.96
```

Note: These are 95%-CIs are based on the normal approximation which can be poor in small samples, e.g. in age groups 6-7 and 10-11 years.

Differences in means

Let's compare the means between the sexes in the largest age group (8-9 years):

```
comp_mean←tbl_fev1 %>%
  filter(age_cat == "[8,10)") %>%
  select(mean, SE, CI_norm_lb, CI_norm_ub)
print(comp_mean,digits=3)
```

```
# A tibble: 2 × 5
# Groups:   sex [2]
  sex    mean     SE CI_norm_lb CI_norm_ub
  <fct> <dbl>  <dbl>      <dbl>      <dbl>
1 f      1.53 0.0161       1.50       1.56
2 m      1.67 0.0186       1.63       1.71
```

Mean FEV_1 is higher in boys than girls by 138 ml:

```
D←comp_mean$mean[comp_mean$sex=="m"]-comp_mean$mean[comp_mean$sex=="f"]
print(D, digits=4)
```

```
[1] 0.1378
```

Differences in means

If the samples are *independent* the standard error of the difference in means $D = \overline{x}_2 - \overline{x}_1$:

$$\hat{SE}_D = \sqrt{\hat{SE}_{\overline{x}}^2 + \hat{SE}_{\overline{x}_2}^2}$$

True parameter	Estimate	z-statistic
μ_1	$\overline{x}_1$	$z_1 = \frac{\overline{x}_1 - \mu_1}{SE_1}$
μ_2	$\overline{x}_2$	$z_1 = \frac{\overline{x}_2 - \mu_2}{SE_1}$
$\Delta = \mu_2 - \mu_1$	$D = \overline{x}_2 - \overline{x}_2$	$z_D = \frac{D - \Delta}{SE_D}$

By the CLT, the large sample sampling distribution of z_D is approximately standard normal.

95%-CI for difference in means

We can thus use the same recipe for constructing a 95%-CI for the difference in means:

$[D - 1.96 \cdot \hat{SE}_D, D + 1.96 \cdot \hat{SE}_D]$

```
  D = comp_mean$mean[comp_mean$sex=="m"]-comp_mean$mean[comp_mean$sex=="f"]
  SE_D←sqrt(sum(comp_mean$SE^2))
  CI_lb←D-qnorm(0.975)*SE_D
  CI_ub←D+qnorm(0.975)*SE_D
sprintf("D =  %1.3f, SE_D = %1.3f, 95%%-CI: [%1.3f, %1.3f]", D, SE_D, CI_lb,
```

```
[1] "D =  0.138, SE_D = 0.025, 95%-CI: [0.090, 0.186]"
```

Interpretation: “We are 95% confident that mean FEV_1 in boys is between 90 ml and 186 ml higher than in girls”

Another approach

We formulate following *null hypothesis* and *alternative hypothesis*:

$H_0 : \ \mu_1 = \mu_2$ or equivalently $\Delta = \mu_2 - \mu_1 = 0$

$H_A : \ \mu_1 \neq \mu_2$

Large deviations of $D = \overline{x}_2 - \overline{x}_1$ from 0 on either side provide evidence against H_0.

Note: H_0 typically states the absence of a difference between groups or of an association.

Another approach - Testing

Under $H_0 : \Delta = 0$, the statistic $z_D = \frac{D}{SE_D}$ approximately follows the standard normal distribution.

The combined shaded area beyond the $|z_D|$ on either side of 0 gives us the two-sided p-value.

The P-value

P-value: The conditional probability (under repeated sampling) of obtaining a test statistic *as extreme or more extreme* than the one observed given H_0 is true.

Calculate the two sided p-value:

```
D = comp_mean$mean[comp_mean$sex=="m"]-comp_mean$mean[comp_mean$sex=="f"]
SE_D←sqrt(sum(comp_mean$SE^2))
z←D/sqrt(sum(comp_mean$SE^2))
p←2*pnorm(-abs(z))
sprintf("D =  %1.3f, SE_D = %1.3f, z_D = %1.3f, P = %1.3e", D, SE_D, z, p)
```

```
[1] "D =  0.138, SE_D = 0.025, z_D = 5.597, P = 2.184e-08"
```

$P = 2.2 \cdot 10^{-8}$ is an extremely small value. There is strong evidence against H_0.

Interpreting the P-value

From: Sterne et al. BMJ 2001;322:226

„A p-value of < 0.05 was considered statistically significant …“

Significance tests

Significance tests define a threshold for $|z_D|$ above which H_0 is rejected. We can err in two ways:

- Type I error: We (falsely) reject H_0 when H_0 is true
- Type II error: We (falsely) maintain H_0 when H_0 is false

To determine the threshold we fix the probability of Type I error given H_0 to a small pre-specified value α, the *significance level*.
Typically α is set to 0.05, which amounts to rejecting H_0 whenever $|z_D| \geq 1.96$ or, equivalently, $P \leq 0.05$.
The difference is then said to be *"statistically significant"*

A word of caution

The P-value is frequently (if not usually) misinterpreted:

- **A dirty dozen: twelve p-value misconceptions**

  $P \leq 0.05$ has been immensely abused in the literature and often equated with proving an effect. Here some of the criticisms:

- **The ASA statement on p-values**
- **Scientists rise up against statistical significance**
- **Sifting the evidence — what's wrong with significance tests?**

.

from Sterne & Davy Smith. BMJ 2001;322:226-31

The t-distribution, t-tests

The normality assumption

Normality assumption: We assume the underlying distribution (population) of our measurements X is normal with unknown mean μ and standard deviation σ. Under normality, the sampling distribution of the mean of $\overline{x}$ is *exactly* normal with mean μ and standard deviation $SE = \frac{\sigma}{\sqrt{n}}$ (no approximation involved).

Remember: The standard error is the standard deviation of a sampling distribution.

The normality assumption

Example: Assume that the body height of men is normally distributed with $\mu = 170\ \mathrm{cm}$ and $\sigma = 10\ \mathrm{cm}$ (Made up)

Caveat: σ is unknown. We can only estimate SE using $\hat{SE} = \frac{s}{\sqrt{n}}$.

The normality assumption

In small samples, $z = \frac{\overline{x} - \mu}{\hat{SE}}$ is somewhat wider than the standard normal distribution.

Student’s t-distribution

Under the normality assumption, the statistic $t = \frac{\overline{x} - \mu}{\hat{SE}}$ follows the t**-distribution** with $n - 1$ degrees of freedom (one df lost because the mean was used to obtain s).

Quantiles of the t-distribution

The t-distribution approaches the standard normal distribution as the df increase:

df	5	50	100	500	1000
$t_{\mathrm{df},0.975}$	2.5706	2.0086	1.984	1.9647	1.9623

Better 95%-CIs for the mean

If normality holds, we can calculate exact confidence intervals even from small samples:

$[\overline{x} - t_{n-1,0.975}\hat{SE},\ \overline{x} + t_{n-1,0.975}\hat{SE}]$

where $t_{n-1,0.975}$ is the 0.975-quantile of the t-distribution with n-1 df.

Note: In large samples this is equivalent to the normal approximation and hence the normality assumption is superfluous.

Better 95%-CIs for the mean

```r
tbl_fev1$SE <-tbl_fev1$sd/sqrt(tbl_fev1$n)
tbl_fev1$CI_t_lb<-tbl_fev1$mean - qt(0.975,tbl_fev1$n-1)*tbl_fev1$SE
tbl_fev1$CI_t_ub<-tbl_fev1$mean + qt(0.975,tbl_fev1$n-1)*tbl_fev1$SE
print(tbl_fev1, digits=3)
```

```
# A tibble: 6 × 10
# Groups:   sex [2]
  sex   age_cat     n  mean    sd     SE CI_norm_lb CI_norm_ub CI_t_lb
CI_t_ub
  <fct> <fct>   <int> <dbl> <dbl>  <dbl>      <dbl>      <dbl>   <dbl>
<dbl>
1 f     [6,8)      29  1.30 0.232 0.0431       1.21       1.38    1.21
1.38
2 f     [8,10)    275  1.53 0.267 0.0161       1.50       1.56    1.50
1.56
3 f     [10,12)    31  1.83 0.300 0.0538       1.73       1.94    1.72
1.94
4 m     [6,8)      24  1.35 0.257 0.0525       1.25       1.45    1.24
1.46
5 m     [8,10)    254  1.67 0.296 0.0186       1.63       1.71    1.63
1.71
6 m     [10,12)    23  1.85 0.256 0.0534       1.75       1.96    1.74
1.96
```

Note: The CIs based on the t-distribution are wider, particularly in the smaller age groups.

Two-sample t-test

Recall our testing problem for the difference in means between two groups:

$H_0 : \ \mu_1 = \mu_2$ or equivalently $\Delta = \mu_2 - \mu_1 = 0$

$H_A : \ \mu_1 \neq \mu_2$

Two-sample t-test

If normality holds in both populations, we can use the t-distributed statistics to test H_0. We distinguish two situations:

Assumptions	test statistic	df of t-distribution	R t.test option
Equal variances	$t = \frac{D}{s_p \cdot \sqrt{1/n_1 + 1/n_2}}$	$n - 2$	`var.equal = TRUE`
Unequal variances	$t = \frac{D}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	complicated formula	`var.equal = FALSE`

Here s_p is a pooled estimator of the common standard deviation. For more details see the **Wikipedia article**

The second version is known as Welch's test. It is the default in R (we recommend to keep it).

Two-sided p-values are obtained by summing the tail areas of the corresponding t-distribution below $-|t|$ and above $|t|$ (as for the z-test using the standard normal distribution).

Two-sample t-test

Let's test for differences in mean FEV_1 between sexes in the smaller age groups

```
1 ind←data$age_cat=="[6,8)"
2 t.test(fev1~sex, data=data, subset=ind)
```

```
    Welch Two Sample t-test

data:  fev1 by sex
t = -0.7894, df = 46.944, p-value = 0.4338
alternative hypothesis: true difference in means between group f and group m
is not equal to 0
95 percent confidence interval:
 -0.19012150  0.08296633
sample estimates:
mean in group f mean in group m
       1.295172        1.348750
```

Interpretation: There is no evidence of a difference in mean FEV_1 between sexes in the youngest age group (P=0.43).

Two-sample t-test

Here a summary output (last columns) of the two-sample t-test for all age groups:

```
   age_cat n_f n_m mean_f mean_m      t        P
1   [6,8)  29  24   1.30   1.35 -0.789 4.34e-01
2  [8,10) 275 254   1.53   1.67 -5.597 3.57e-08
3 [10,12)  31  23   1.83   1.85 -0.277 7.83e-01
```

As a reminder: the P-value for the largest age group (8-9 years) using the normal approximation was P=2.184e-08.

One-sample t-test

Assume wanted to test whether mean FEV_1 in healthy 8-9 year old girls is 1.40 l, a reference value from the literature.

To test $H_0 : \mu = 1.40$ we compare $t = \frac{\overline{x} - 1.40}{\hat{SE}}$ to the t-distribution with $n - 1$ df.

```
ind←data$age_cat=="[6,8)" & data$sex=="f"
t.test(data$fev1[ind], mu=1.4)
```

```
    One Sample t-test

data:  data$fev1[ind]
t = -2.4337, df = 28, p-value = 0.02158
alternative hypothesis: true mean is not equal to 1.4
95 percent confidence interval:
 1.206940 1.383404
sample estimates:
mean of x
 1.295172
```

Paired-samples t-test

Situations with dependent samples $x_{1,i}$, $x_{2,i}$ $(i = 1, \ldots, n)$:

- Repeated measurements on the same subjects (e.g. pre- post treatment, using different devices)
- Measurements on persons matched on certain characteristics (e.g. same family)

Paired-samples t-test: To test $H_0 : \mu_1 = \mu_2$ we run a one sample t-test on the sample $d_i = x_{1,i} - x_{2,i}$ (pairwise differnces) testing for $H_0 : \delta = 0$ where δ is the population mean of the pairwise differences.

Comparing multiple groups - ANOVA

Analysis of variance (ANOVA) is a method for comparing means (despite the name) across multiple groups.

- It is a generalization of the t-test to multiple, say k, groups/sample.
- It assumes normality withing groups equal variance across groups.
- When $k = 2$ it is equivalent to the t-test (with equal variances)

FEV_1 by asthma history

```
ggplot(data, aes(x=asthma_hist, y=fev1)) +
  geom_boxplot() +
  ylab("FEV1") + xlab("Asthma history") +
  facet_wrap(~sex)
```

FEV_1 by asthma history

Let's focus on the girls and let μ_1, μ_2, μ_3 be the population means of FEV_1 in the following groups:

- Group 1: "never" (no asthma history)
- Group 2: "previous asthma"
- Group 3: "current asthma"

We want to test
$H_0 : \mu_1 = \mu_2 = \mu_3$
Note: H_0 is false even if only one mean deviates the others.

FEV_1 by asthma history

Sample means by group:

```
dat←data %>% filter(sex=="f")
tbl2_fev1 ← dat %>% group_by(asthma_hist) %>% summarise(n=n(), mean=mean(fev
tbl2_fev1
```

```
# A tibble: 3 × 4
  asthma_hist          n  mean    sd
  <fct>            <int> <dbl> <dbl>
1 current asthma      37  1.42 0.281
2 never              238  1.56 0.301
3 previous asthma     60  1.52 0.234
```

And overall:

```
total ← dat %>% summarise(n=n(), mean=mean(fev1), sd=sd(fev1))
total
```

```
# A tibble: 1 × 3
      n  mean    sd
```

```
  <int> <dbl> <dbl>
1   335  1.54 0.291
```

ANOVA

ANOVA compares following sums of squares SS:

Between groups:

$SS_B = \sum_{j=1}^{3} \sum_{i \in group_j} (\hat{\mu}_j - \hat{\mu})^2 = 0.647$

Withing groups:

$SS_W = \sum_{j=1}^{3} \sum_{i \in group_j} (x_i - \hat{\mu}_j)^2 = 27.575$

Here the $\hat{\mu}_j$ stand for the group-wise sample means, $\hat{\mu}$ stand for the overall sample mean, and the x_i for individual FEV_1 values.

ANOVA

A large value of SS_B provide evidence against H_0.
We use the F-statistic to test H_0:

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

In our case $F = 3.896$
This value must be compared against the F distribution $k - 1 = 2$ und $n - k = 332$ degrees of freedom.

ANOVA

The p-values is the tail area of the of F distribution above the test statistic.

```
p←pf(F,2,332,lower.tail = FALSE)
p
```

```
# [1] 0.02125017
```

This yields $p = 0.02125$
We thus have some evidence against H_0 suggesting that the FEV_1 depends on asthma history.

ANOVA in R

All in one using the `anova` function:

```
1 anova(lm(fev1~asthma_hist, data=dat))
```

```
Analysis of Variance Table

Response: fev1
             Df  Sum Sq Mean Sq F value  Pr(>F)
asthma_hist   2  0.6473 0.32363  3.8964 0.02125 *
Residuals   332 27.5755 0.08306
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: ANOVA can be treated as a special case of the linear regression model. Here, the result of a linear model, executed with `lm`, is passed on to the `anova` function.

Summary

- Measures of central tendency include the mean and median. The median is robust againts outliers.
- Measures of spread include variance, standard deviation, IQR and range.
- The sample mean is an unbiased estimator.
- For large n (sample size), the means from independent samples will vary normally around the true mean.
- Increasing n by 100 will reduce the standard error of the sample mean by 10 (i.e. divide by 10)
- Confidence intervals based on the t-distribution and t-tests for comparisons of means (one-sample and two-sample situation) are appropriate
  - if the normality assumption holds (regardless of sample size) or
  - more generally, when samples sizes are large.
- The F-test (in ANOVA) generalizes the two-sample t-test to the comparison of multiple groups

- While Welch's t-test (the default in R) allows for unequal variances between groups, ANOVA strictly assumes equal variances across groups.