DATA2002

Testing means

Garth Tarr

THE UNIVERSITY OF SYDNEY

One sample t-test

Two-sample t-test

Paired samples t-test

General t-test background

The t-distribution: what is it?

- Some basic probability facts about samples from *normal populations* will prove useful.

1. The sample mean from a normal sample is itself normally distributed.
2. The sample variance from a normal sample has a *scaled* χ^2 (chi-squared) distribution.
3. The sample mean and sample variance from a normal sample are *statistically independent*.

- If $Z \sim N(0, 1)$ is independent of a χ^2_d random variable, the quantity

$$\frac{Z}{\sqrt{\chi^2_d / d}} \sim t_d \,,$$

  a t-distribution with d degrees of freedom.

The t-statistic

- Combining these facts means that if the population mean is μ, the sample mean and variance are $\bar{X}$ and S^2, the ratio

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{S/\sigma} \sim t_{n-1}\,.$$

  - the numerator is $N(0,1)$;
  - the denominator is $\sqrt{\chi^2_{n-1}/(n-1)}$, independently of the numerator.
- Indeed, in many statistical applications we have a model whereby a certain statistic has this general form:
  - some estimator of some parameter is normally distributed;
  - a standard error based on the data has a distribution like $\sqrt{\chi^2_d/d}$ times the true SD of the estimator (for some d) and is *independent* of the estimator;
  - then the ratio $\frac{\text{estimator}-\text{true value}}{\text{standard error}} \sim t_d$.

One sample t-test

Some of the selection at Noosa Chocolate Factory.

Crime & Justice

Noosa Chocolate Factory fined $12,000 for improper weighing

Kay Dibben, The Courier-Mail
June 2, 2017 3:41pm

The scales of justice: National Measurement Institute audits fruit and vegetable weights

Share on Facebook | Share on Twitter

ABC Rural By Nikolai Beilharz

Updated 21 March 2017 at 3:04 pm

First posted 21 March 2017 at 3:01 pm

Fruit and vegetable packers will come under extra scrutiny as they are audited to check that a bag of produce actually has as much weight in it as it says on the label.

The weights of packaged fruit and vegetables from farms are to be audited. (ABC News: Gary Rivett:)

The National Measurement Institute, a division of the Department of Industry, Innovation and Science, said its inspectors would visit 1,400 traders ranging from producers to wholesalers and retailers to check over 1,500 measuring instruments to ensure that they were accurate.

They will also inspect over 11,000 lines of packaged fruit and vegetable goods as part of the audit.

JANUARY 13 2016

Thousands of food products found to be underweight by the measurement authority

Esther Han

Birthday sponge cakes sold at Woolworths and lamb chops prepared at Coles were among thousands of products found to be weighing less than what was promised on packaging, according to the National Measurement Institute.

In 2014-15, Australia's peak weights and measures body issued 3962 non-compliance notices to traders, up 13 per cent on the previous year's figure. It sent 139 warning letters and imposed 98 fines totalling $92,650.

Beer contents

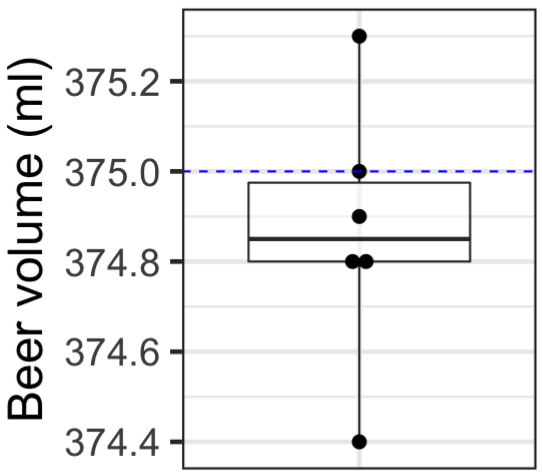
Beer contents in a pack of six bottles (in millilitres) are:

```r
y = c(374.8, 375.0, 375.3, 374.8, 374.4, 374.9)
```

Is the mean beer content less than the 375 ml claimed on the label?

```r
library("ggplot2")
df = data.frame(y)
set.seed(124)
p1 = ggplot(df, aes(x = "", y = y)) +
  geom_boxplot(alpha = 0.5, coef = 10) +
  geom_dotplot(binaxis = 'y',
               stackdir = 'center') +
  geom_hline(yintercept = 375,
             colour = "blue",
             linetype = "dashed") +
  labs(y = "Beer volume (ml)", x = "") +
  theme_bw(base_size = 24) +
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank())
```

```r
p1
```



Note `geom_boxplot()` requires something for the x-axis, this is usually a factor variable used to compare groups, but in this case, we only have one group. We can provide anything, here we just specified a blank group on the x axis with `x = ""`.

Hypothesis testing

- **Hypotheses:** $H_0\colon\ \theta = \theta_0$ vs $H_1\colon\ \theta > \theta_0$ or $\theta < \theta_0$ or $\theta \neq \theta_0$
- **Assumptions:** $X_1, X_2, \ldots, X_n \sim F_\theta$
- **Test statistic:** $T = f(X_1, X_2, \ldots, X_n)$.
- **Observed test statistic:** $t_0 = f(x_1, x_2, \ldots, x_n)$.
- **Significance:** p-value $= P(T \geq t_0)$ or $P(T \leq t_0)$ or $2P(T \geq |t_0|)$
- **Decision:** If the p-value is less than α, there is evidence against H_0.

Hypothesis

- The statement against which you search for evidence is called the null hypothesis, and is denoted by H_0. It is generally a "no difference" statement.
- The statement you claim is called the alternative hypothesis, and is denoted by H_1.

Typical Hypotheses:

H_0: $\theta = \theta_0$

VS

H_1: $\theta > \theta_0$ (upper-side alternative)

H_1: $\theta < \theta_0$ (lower-side alternative)

H_1: $\theta \neq \theta_0$ (two-sided alternative)

Assumptions

- Each observation $X_1, X_2, \ldots, X_n$ is chosen at random from a population.
- We say that such random variables are *iid* (independently and identically distributed).
- Each test we consider will have its own assumptions.

Test statistic

- Since observations X_i vary from sample to sample we can never be sure whether H_0 is true or not.
- We use a test statistic $T = f(X_1, \ldots, X_n)$ to test if the data are consistent with H_0 such that the distribution of T is known assuming H_0 is true.

The **observed test statistic**, t_0, is where we plug our observed data into the formula for the test statistic.

Large (positive or negative depending on H_1) observed test statistic values is taken as evidence of poor agreement with H_0.

Significance

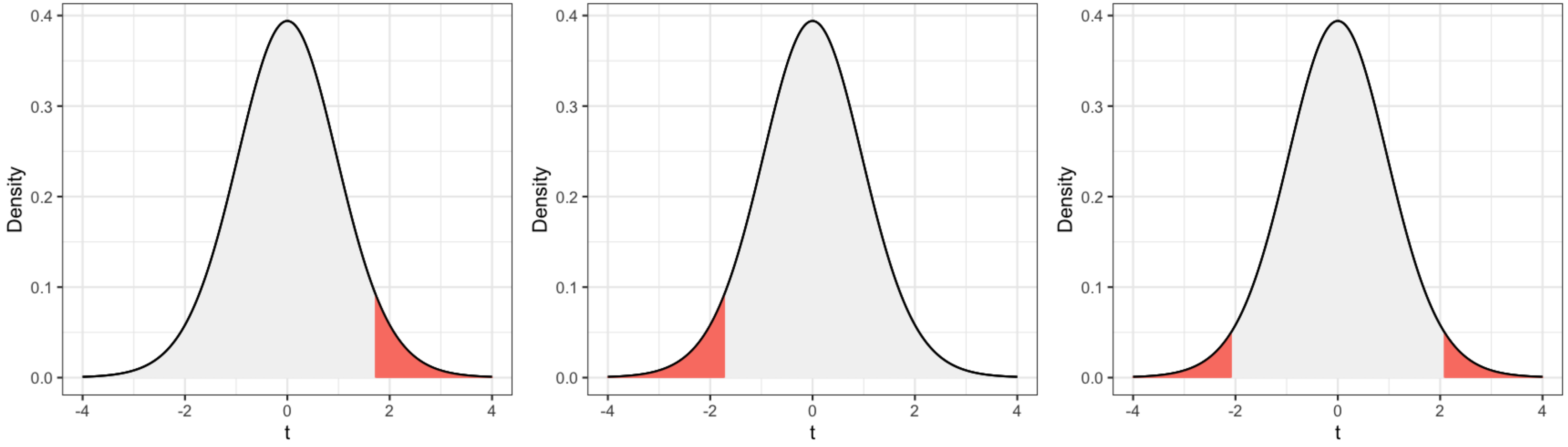
The p-value is defined as the probability of getting a test statistic, T, *as or more extreme* than the value we observed, t_0, *assuming* that H_0 is true.

Typical p-value statements:

For $H_1\colon\ \theta > \theta_0$, p-value = $P(T \geq t_0)$

For $H_1\colon\ \theta < \theta_0$, p-value = $(T \leq -t_0)$

For $H_1\colon\ \theta \neq \theta_0$, p-value = $2P(T \geq |t_0|)$



Decision

An observed *large* positive or negative value of t_0 and hence small p-value is taken as evidence of poor agreement with H_0.

- If the p-value is small, then either H_0 is true and the poor agreement is due to an unlikely event, or H_0 is false. Therefore..
- the smaller the p-value, the stronger the evidence against H_0 in favour of H_1.
- A large p-value does not mean that there is evidence that H_0 is true
- The level of significance, α, is the strength of evidence needed to reject H_0 (often $\alpha = 0.05$).

One sample t-test

Suppose we have a sample $X_1, X_2, \ldots, X_n$ of the size n drawn from a normal population with an unknown variance σ^2. Let $x_1, x_2, \ldots, x_n$ be the observed values. We want to test the population mean μ.

- **Hypothesis:** $H_0\colon\ \mu = \mu_0$ vs $H_1\colon\ \mu > \mu_0,\ \mu < \mu_0$ or $\mu \neq \mu_0$
- **Assumptions:** X_i are iid rv and follow $N(\mu, \sigma^2)$.
- **Test statistic:** $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$. Under $H_0, T \sim t_{n-1}$.
- **Observed test statistic:** $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- **p-value:** $P(t_{n-1} \geq t_0)$, $P(t_{n-1} \leq t_0)$ or $2P(t_{n-1} \geq |t_0|)$
- **Decision:** Reject H_0 in favour of H_1 if the p-value is less than α.

Beer contents

Beer contents in a pack of six bottles (in millilitres) are:

$$374.8, 375.0, 375.3, 374.8, 374.4, 374.9$$

Is the mean beer content less than the 375 ml claimed on the label?

```r
x = c(374.8, 375.0, 375.3, 374.8, 374.4, 374.9)
mean(x)
```

```
## [1] 374.8667
```

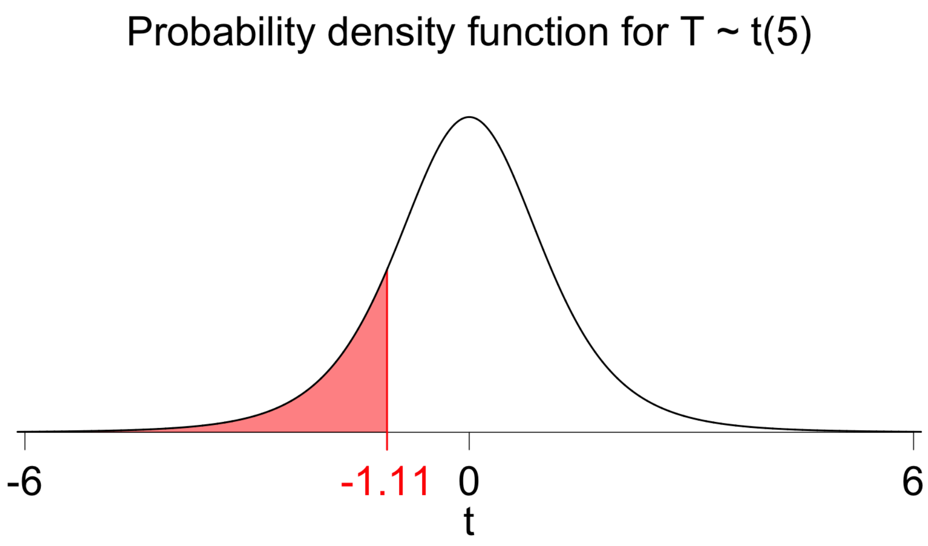
```r
sd(x)
```

```
## [1] 0.294392
```

Beer contents

Workflow for a one sample t-test. Let μ be the population mean beer content (in millilitres).

- **Hypothesis:** $H_0\colon\ \mu = 375$ vs $H_1\colon\ \mu < 375$
- **Assumptions:** X_i are *iid* rv and follow $N(\mu, \sigma^2)$.
- **Test statistic:** $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$. Under H_0, $T \sim t_{n-1}$.
- **Observed test statistic:** $t_0 = \dfrac{374.87 - 375}{0.29/\sqrt{6}} = -1.11$
- **p-value:** $P(t_5 \leq -1.11) = 0.16$.
- **Decision:** The data is consistent with the null hypothesis H_0.



```r
t.test(x, mu = 375, alternative = "less")
```

```
## 
##     One Sample t-test
## 
## data:  x
## t = -1.1094, df = 5, p-value = 0.1589
## alternative hypothesis: true mean is less than 375
## 95 percent confidence interval:
##      -Inf 375.1088
## sample estimates:
## mean of x 
##  374.8667
```

```r
n = length(x)
t0 = (mean(x) - 375)/(sd(x)/sqrt(n))
t0
```

```
## [1] -1.1094
```

```r
pval = pt(t0, n - 1)
pval
```

```
## [1] 0.1588721
```

Two-sample t-test

What if you have two samples?

There are times that we want to test if the population means of two samples are different.

Here we are left with two possible scenarios

- Two independent samples
- Two related samples (dependent samples or repeated measures)

> Blood samples are taken from 11 smokers and 11 non-smokers to measure aggregation of blood platelets. Are these **dependent** or **independent** samples?

Smoking Can Cause Platelets Aggregation in Blood

Platelets are important components in our blood. They are responsible for keeping the blood within the vessels. Research shows that Smoking can increase the count of the platelets in the blood. This can lead to clotting of the blood, medically called Thrombus. This article tries gains further insight.

What is Platelets Aggregation ?

- They are cells but without nucleus. They are produced in the bone marrow and live for 8 to 12 days.
- They block bleeding of blood, repair, replace and regenerate tissues in the body.
- Platelet count in normal healthy person is between 150,000 – 450,000 per microlitre of blood.
- When the platelets become too low, it causes excessive bleeding.
- If the count of platelets are high, it is called platelet aggregation causing clotting of blood.
- 1,727 people per year die on average because of Thrombus in U.S.

Smokers and blood platelet aggregation

Blood samples are taken from 11 smokers and 11 non-smokers to measure aggregation of blood platelets.

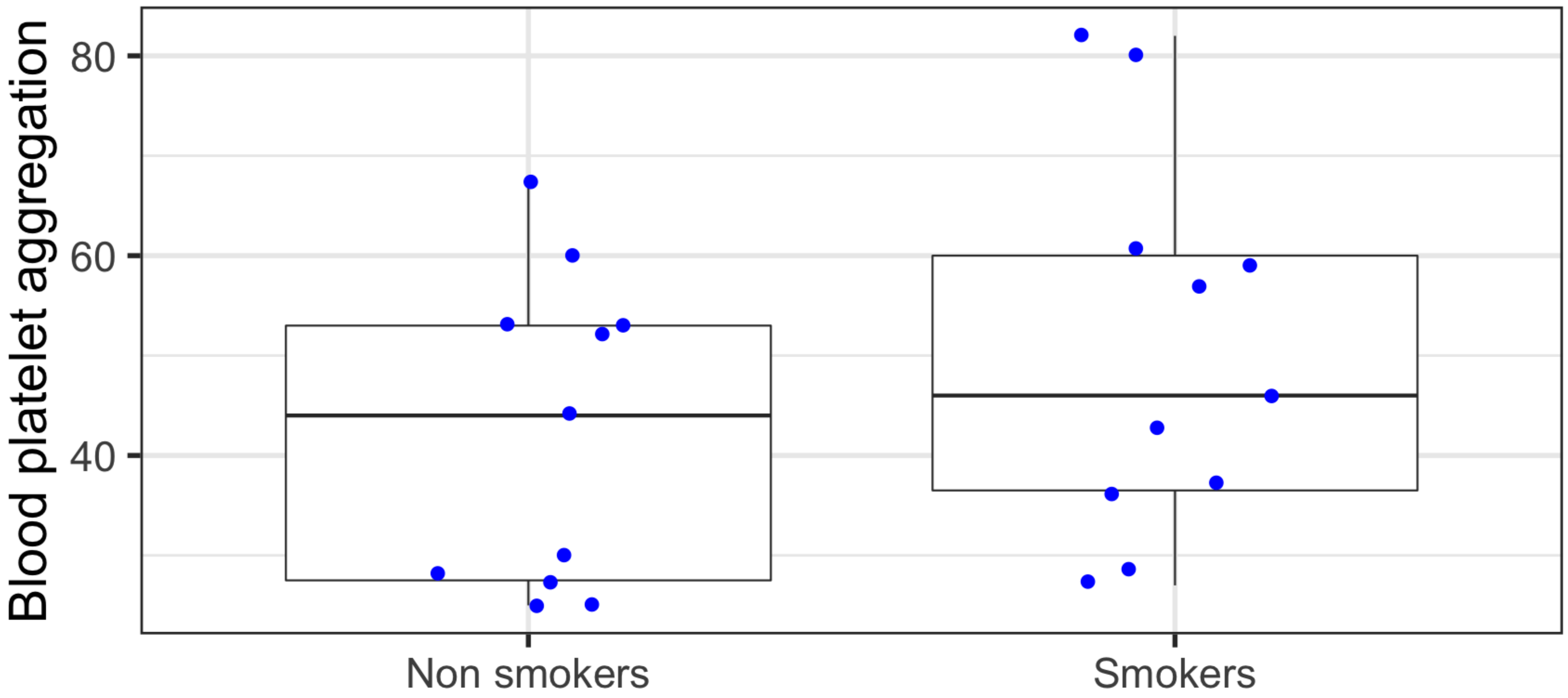
```r
non_smokers = c(25, 25, 27, 44, 30, 67,
                53, 53, 52, 60, 28)
smokers =  c(27, 29, 37, 36, 46, 82,
             57, 80, 61, 59, 43)
dat = data.frame(
  platelets = c(non_smokers, smokers),
  status = c(rep("Non smokers",
                 length(non_smokers)),
             rep("Smokers",
                 length(smokers)))
)
library(dplyr)
sum = dat %>%
  group_by(status) %>%
  summarise(Mean = mean(platelets),
            SD = sd(platelets),
            n = n())
```

```r
knitr::kable(sum, format = "html", digits = 1)
```

status	Mean	SD	n
Non smokers	42.2	15.6	11
Smokers	50.6	18.9	11

Is the aggregation of blood platelets affected by smoking?

```r
library("ggplot2")
ggplot(dat, aes(x = status, y = platelets)) +
  geom_boxplot() +
  geom_jitter(width=0.15, size = 3, colour = "blue") +
  theme_bw(base_size = 28) +
  labs(x = "", y = "Blood platelet aggregation")
```



Two-sample t-test

1. **Hypotheses:** $H_0\colon\ \mu_x = \mu_y$ vs $H_1\colon\ \mu_x > \mu_y$ or $\mu_x < \mu_y$ or $\mu_x \neq \mu_y$
2. **Assumptions:** $X_1, \dots, X_{n_x}$ are *iid* $N(\mu_X, \sigma^2)$, $Y_1, \dots, Y_{n_y}$ are *iid* $N(\mu_Y, \sigma^2)$ and X_i's are independent of Y_i's.
3. **Test statistic:** $T = \dfrac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $S_p^2 = \dfrac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$. Under H_0, $T \sim t_{n_x+n_y-2}$
4. **Observed test statistic:** $t_0 = \dfrac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $s_p^2 = \dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$.
5. **p-value:** $P(t_{n_x+n_y-2} \geq t_0)$ or $P(t_{n_x+n_y-2} \leq t_0)$ or $2P(t_{n_x+n_y-2} \geq |t_0|)$.
6. **Decision:** If the p-value is less than α, there is evidence against H_0. If the p-value is greater than α, the data are consistent with H_0.

Let X_i be the blood platelet aggregation levels for the i^{th} non-smoker and Y_j the levels for the j^{th} smoker. Let μ_S and μ_N be the population mean platelet aggregation levels for smokers and non-smokers respectively.

1. **Hypotheses:** $H_0\colon\ \mu_S = \mu_N$ vs $H_1\colon\ \mu_S \neq \mu_N$
2. **Assumptions:** $X_1, \ldots, X_{n_x}$ are *iid* $N(\mu_X, \sigma^2)$, $Y_1, \ldots, Y_{n_y}$ are *iid* $N(\mu_Y, \sigma^2)$ and X_i's are independent of Y_i's.
3. **Test statistic:** $T = \dfrac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $S_p^2 = \dfrac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$. Under H_0, $T \sim t_{n_x+n_y-2}$
4. **Observed test statistic:** $t_0 = \dfrac{50.6 - 42.2}{17.3\sqrt{\frac{1}{11} + \frac{1}{11}}} = 1.14$ where $s_p^2 = \frac{(11-1)18.9^2+(11-1)15.6^2}{11+11-2} = 17.3.$
5. **p-value:** $2P(t_{20} \geq |1.14|) = 0.27$
6. **Decision:** The data are consistent with H_0. There does not appear to be evidence that blood platelet aggregation levels are different in smokers.

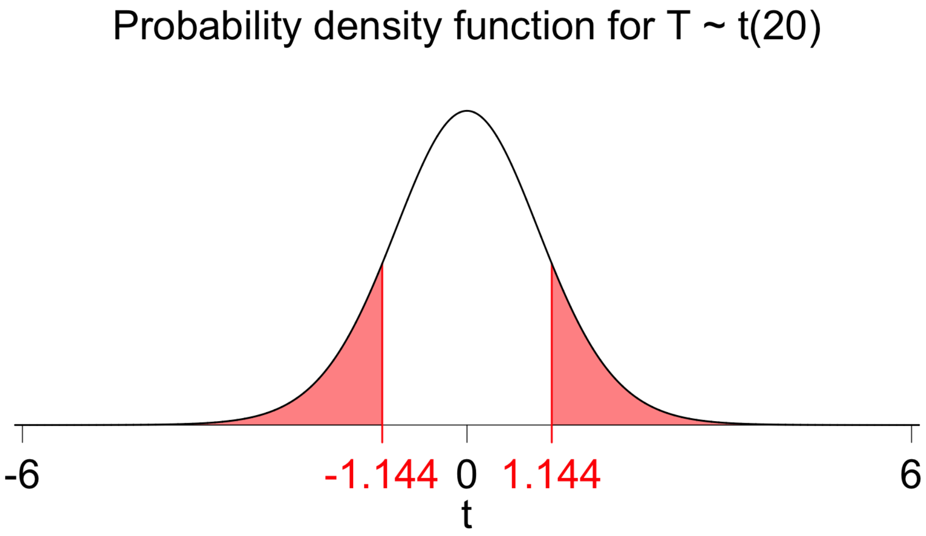
```r
t.test(smokers, non_smokers, alternative = "two
```

```
##
##     Two Sample t-test
##
## data:  smokers and non_smokers
## t = 1.144, df = 20, p-value = 0.2661
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.961816 23.870907
## sample estimates:
## mean of x mean of y
##  50.63636  42.18182
```

```r
nS = length(smokers)
nN = length(non_smokers)
sS = sd(smokers)
sN = sd(non_smokers)
sP = sqrt(((nS - 1) * sS^2 + (nN - 1) * sN^2)/
            (nS + nN - 2))
xbarS = mean(smokers)
xbarN = mean(non_smokers)
deg_free = nS+nN-2
```

```r
t0 = (xbarS - xbarN)/(sP * sqrt(1/nS + 1/nN))
p_val = 2 * (1 - pt(abs(t0), deg_free))
c(t0, p_val)
```

```
## [1] 1.1439712 0.2661433
```



The Equal Variance Assumption

- Lurking among all the assumptions a few slides back was:
- $X_1, \ldots, X_{n_x}$ are *iid* $N(\mu_X, \sigma^2)$,
- $Y_1, \ldots, Y_{n_y}$ are *iid* $N(\mu_Y, \sigma^2)$ and
- $X_i's$ are independent of $Y_i's$.
- In particular, we assume that the two underlying normal populations *have the same variance*.
- In this example, does this seem reasonable?

```
c(sS,sN)
```

```
## [1] 18.89589 15.61293
```

- These are a little different: are they so different that the "equal underlying population variances" assumption is not reasonable?
- We have options: the **Welch Test**.

The Welch Two-Sample t-test

- Welch developed an alternative test which does **not** assume equal population variances.
- In that case, if
- the X_i's are $N(\mu_X, \sigma_X^2)$ and
- the Y_i's are $N(\mu_Y, \sigma_Y^2)$ then then the variance of the sample mean difference is

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}.$$

- The standard error is obtained by plugging in the two sample variances and taking the square root (**note**: we do not need to compute a "pooled" estimate of the common variance!).
- This gives the *Welch statistic*

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_x} + \frac{S_Y^2}{n_y}}}.$$

Welch statistic **not** a proper t-statistic

- Technically, this statistic is not a "usual" t-statistic since the denominator is not a scaled χ^2 independent of the numerator.
- However, the whole statistic still has an *approximate* t-distribution:
- the degrees of freedom is not necessarily a whole number, and is estimated from the data.
- **Thankfully** R can implement this with no drama (we leave out the `var.equal = TRUE` argument):

```
t.test(smokers, non_smokers, alternative = "two.sided")
```

```
## 
##  Welch Two Sample t-test
## 
## data:  smokers and non_smokers
## t = 1.144, df = 19.313, p-value = 0.2666
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.997031 23.906122
## sample estimates:
## mean of x mean of y 
##  50.63636  42.18182
```

Paired samples t-test

Smoking

Blood samples from 11 individuals **before and after** they smoked a cigarette are used to measure aggregation of blood platelets.

```r
before = c(25, 25, 27, 44, 30, 67, 53, 53, 52,
after =  c(27, 29, 37, 36, 46, 82, 57, 80, 61,
df = data.frame(before, after,
                difference = after-before)
df
```

```
##    before after difference
## 1      25    27          2
## 2      25    29          4
## 3      27    37         10
## 4      44    36         -8
## 5      30    46         16
## 6      67    82         15
## 7      53    57          4
## 8      53    80         27
## 9      52    61          9
## 10     60    59         -1
## 11     28    43         15
```

Is the aggregation affected by smoking?

```r
apply(df, 2, mean) %>% round(2)
```

```
before      after difference
 42.18      50.64       8.45
```

```r
apply(df, 2, sd) %>% round(2)
```

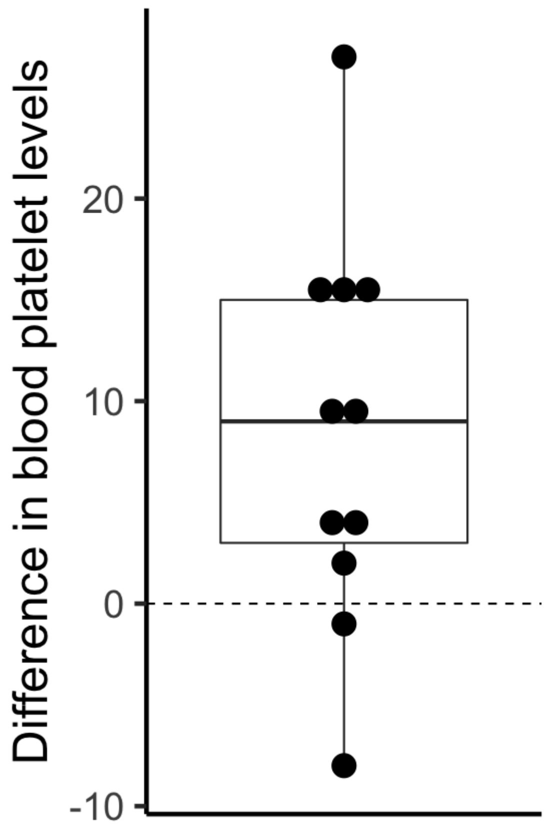
```
before      after difference
 15.61      18.90       9.65
```

```r
df %>%
  summarise(across(.cols = c(before, after),
                   .fns = list(Mean = mean,
                               SD = sd,
                               n = length))) %>%
  pivot_longer(cols = everything(),
               names_sep = "_",
               names_to = c("time", ".value"))
```

```
## # A tibble: 2 × 4
##   time    Mean    SD     n
##   <chr>  <dbl> <dbl> <int>
## 1 before  42.2  15.6    11
## 2 after   50.6  18.9    11
```

```r
p = ggplot(df, aes(x="", y=difference)) +
  geom_boxplot() +
  geom_dotplot(binaxis = "y", stackdir = "center") +
  theme_classic(base_size = 24) +
  geom_hline(yintercept = 0, linetype='dashed') +
  labs(y = 'Difference in blood platelet levels')+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

```r
p
```

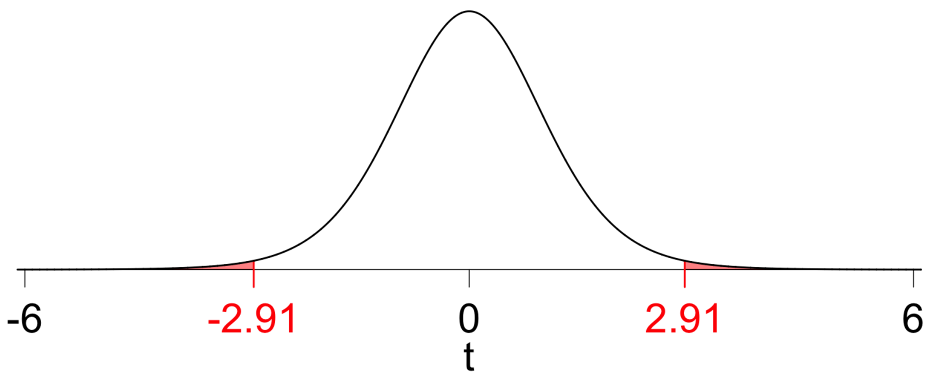


Let X_i and Y_i be the blood platelet aggregation levels for the i^{th} person before and after smoking, respectively. Define the change in person i's platelet aggregation levels as $D_i = Y_i - X_i$ and the population mean change in platelet aggregation levels as μ_d.

- **Hypothesis:** $H_0\colon\ \mu_d = 0$ vs $H_1\colon\ \mu_d \neq 0$
- **Assumptions:** D_i are *iid* $N(\mu_d, \sigma^2)$.
- **Test statistic:** $T = \dfrac{\bar{D}}{S_d/\sqrt{n}}$. Under H_0, $T \sim t_{n-1}$
- **Observed test statistic:** $t_0 = \dfrac{8.45}{9.65/\sqrt{11}} = 2.91$
- **p-value:** $2P(t_{10} \geq |2.91|) = 0.016$.

- **Decision:** As the p-value is small, there is evidence against the null hypothesis. There is evidence that blood platelet aggregation levels change after smoking.

Probability density function for T ~ t(10)



References

For further details see Larsen and Marx (2012), sections 6.1, 6.2, 7.1-7.4 and 9.2.

Larsen, R. J. and M. L. Marx (2012). *An Introduction to Mathematical Statistics and its Applications*. 5th ed. Boston, MA: Prentice Hall. ISBN: 978-0-321-69394-5.