Hypothesis Testing

FGCZ Protein Informatics Training

Witold Wolski wew@fgcz.ethz.ch
24 February, 2021

hypothesis testing

Goals

- What is a hypothesis and how it can be tested?
- What is a test statistic?
- How to generate the distribution of the test statistic if null true?
- One sample t-test?
- What happens if assumptions are not met?
- Central limit theorem
- Asymptotic tests
- Non parametric tests - randomization test
- Comparing parametric, non parametric and asymptotic tests
- Paired t-test
- Equivalence of t-test and linear models

Lady tasting tea

Dr. Muriel Bristol, a female colleague of Fisher claimed to be able to tell whether the tea or the milk was added first to a cup.

- The hypothesis was that the Lady had no such ability.
- The experiment was to prepare 8 cups of tea 4 with milk and 4 with tea first.
- The test statistic was a simple count of the number of successes in selecting the 4 cups out of 8.
- She got all correct. What was the probability of getting all correct?

Lady tasting tea

```r
truth <- c(0,1,0,1,1,0,0,1)
x <- combn(truth,4)
nrcor <- apply(x, 2, sum)
nulldistr <- table(nrcor)
nulldistr
plot(nulldistr, xlab="nr correct")
```

```
## nrcor
##  0  1  2  3  4
##  1 16 36 16  1
```

There are 70 combinations of the elments in x taken m at a time.

Count number of successes for each combination.

Count how often 0, 1, 2, 3, 4 successes.

Lady tasting tea

```
probs ← nulldistr / sum(nulldistr) # compute probabilities
probs ← round(probs, digits = 3)
probs
```

```
## nrcor
##     0     1     2     3     4
## 0.014 0.229 0.514 0.229 0.014
```

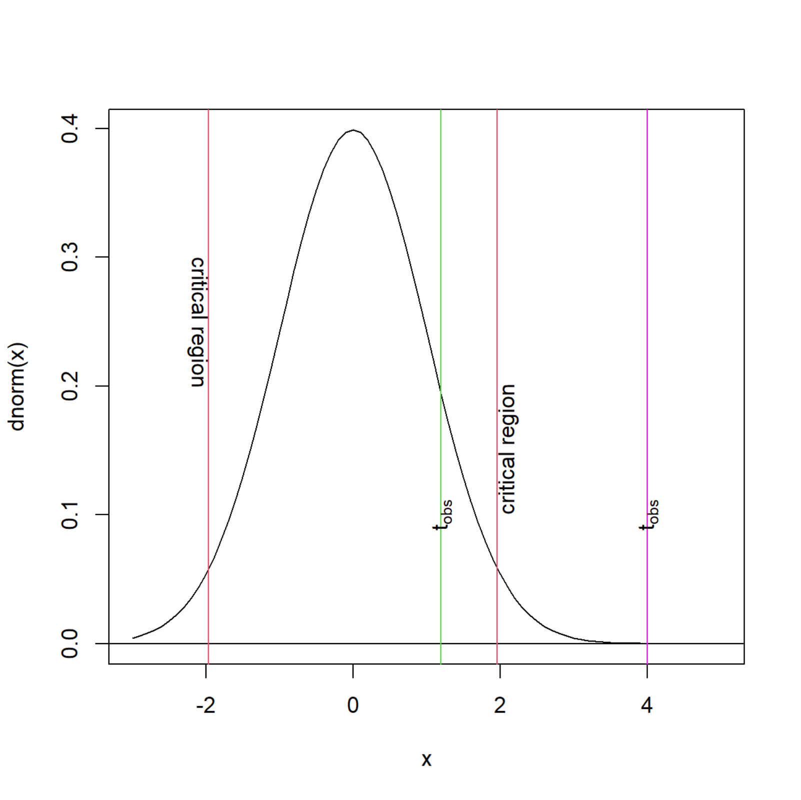
Hence, on $\alpha = 0.05$ reject hypothesis, that she can not recognize if milk or tea first, since getting 4 right is $P(x = 4) = 0.014$

If she would have 3 right would you accept the null hypothesis?

$P(x > 3) = 0.014 + 0.229 = 0.243$.

Hypothesis testing - Brief version

- **State research hypothesis**
- State Relevant Null and Alternative hypothesis.
- Define test (T) statistic.
- Determine distribution of the test statistic under null hypothesis.
- Define Critical region.
- Check if T_{obs} is within the critical region.
- Answer YES or NO.
- or *, ** or *** for 0.1 0.05 or 0.01

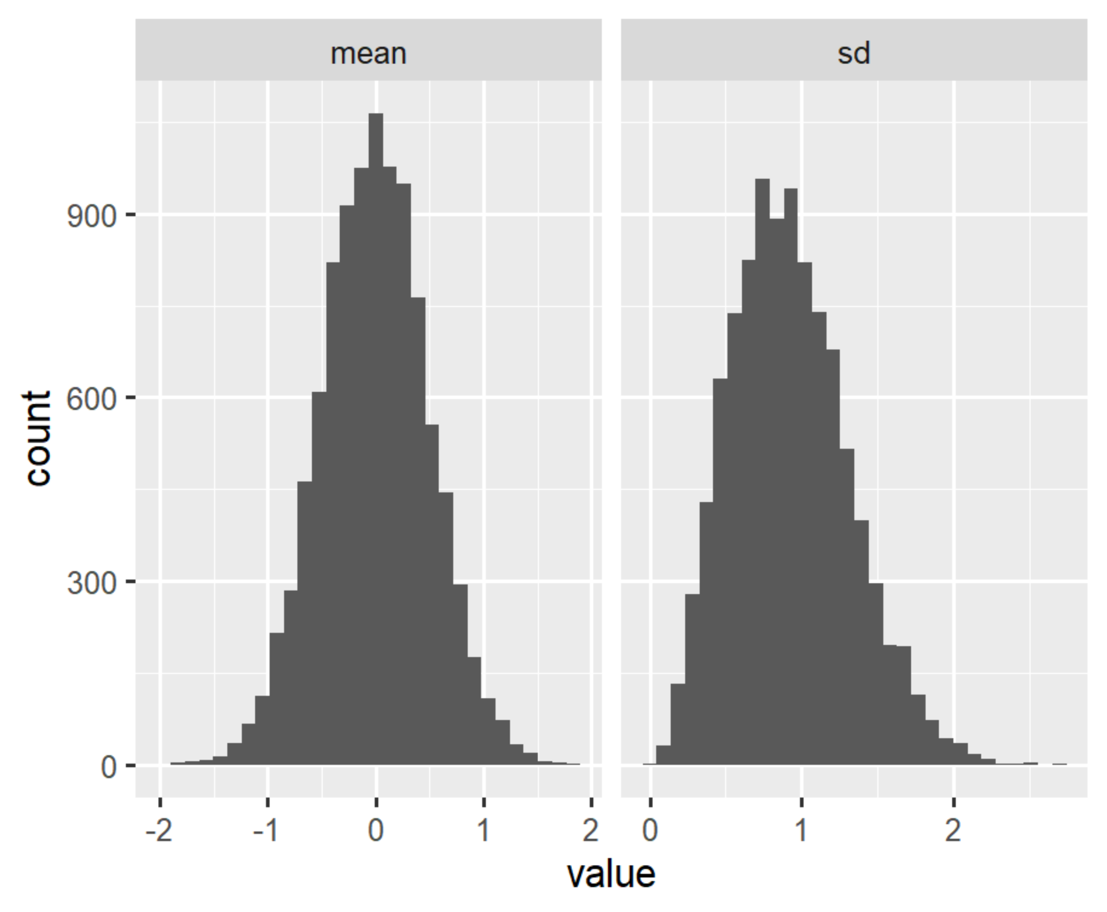


Testing if mean is equal to μ

- Hypothesis - mean of sample is different than some value μ.
- What is the null and what is the alternative?
  - Null is that the mean of observed data is equal to μ.
  - Alternative - it is NOT equal to μ.
- What is the distribution of the observations?
  - Observations are independent, identically distributed (iid)
  - $x \sim N(\mu, \sigma)$.
- State the relevant test statistic T?
  - A suitable test statistics $\bar{X} - \mu$.
- What is the distribution of T under the null hypothesis?
  - It will depend on samples size n and on the variance σ^2

Mean is equal to μ? Simulate data under null

```r
# Simulating data from Null
N ← 10000; N_obs ← 4;
mu ← 0; sigma ← 1
bb ← function(y){
  x ← rnorm( N_obs, mu, sigma )
  data.frame(mean = mean( x ),
             sd = sd(x))
}
res ← purrr::map_df(1:N, bb)
res %>% tidyr::gather() %>%
ggplot(aes(x = value)) +
  geom_histogram() +
  facet_grid(~key,scales="free_x")
```

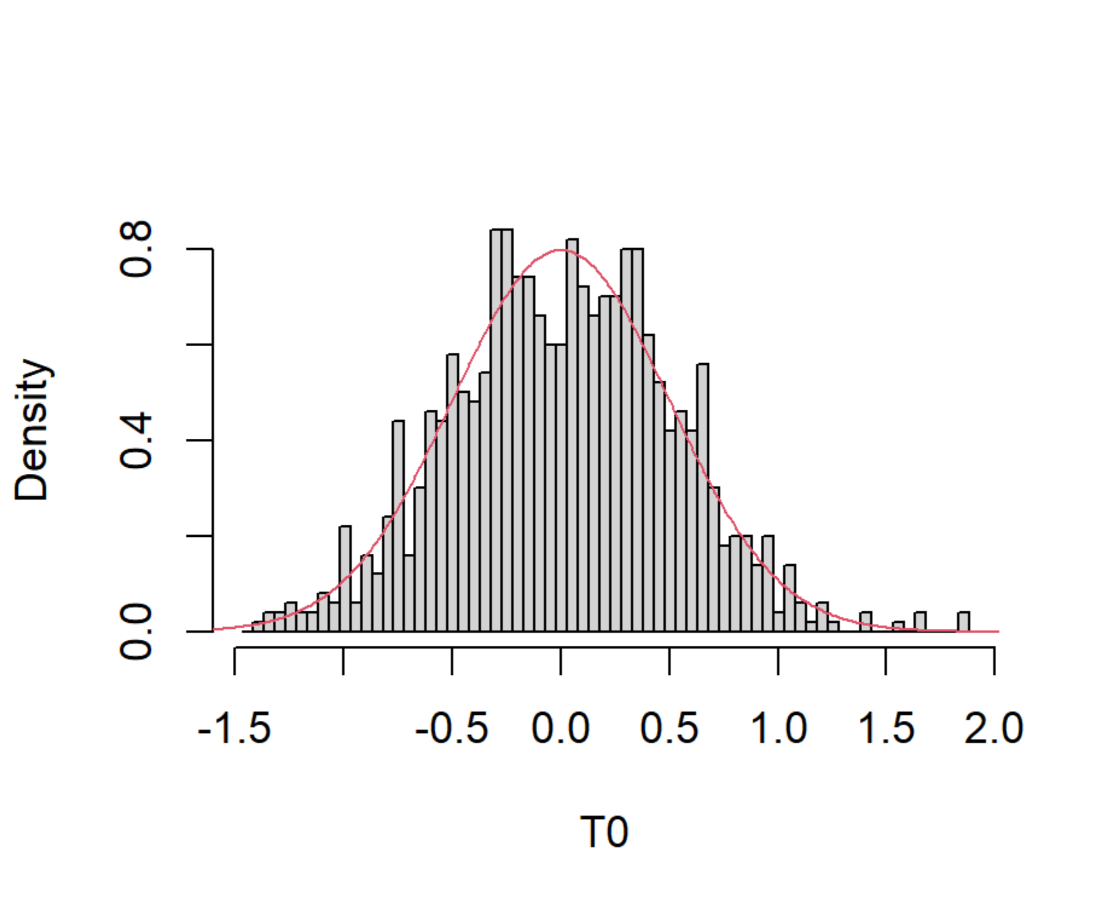


Mean = μ? What is the distribution of T under null?

if σ known

```
T0 ← mu - res$mean
hist(T0,
     breaks=getBreaks(T0, by=0.05),
     probability = T,
     main="")
x ← seq(-10,10,0.01)
lines(x,
      dnorm(x,
            sd= (sigma/sqrt(N_obs))),
      col=2)
```

$T|H_0 \sim N(0, \sigma/\sqrt{N_{obs}})$



Improved test statistic T*

The **t-statistic**

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

Z- transformed data $\sim WN(0,1)$.

The variance of the sampling distribution of the mean is the population variance divided by n (given iid data).

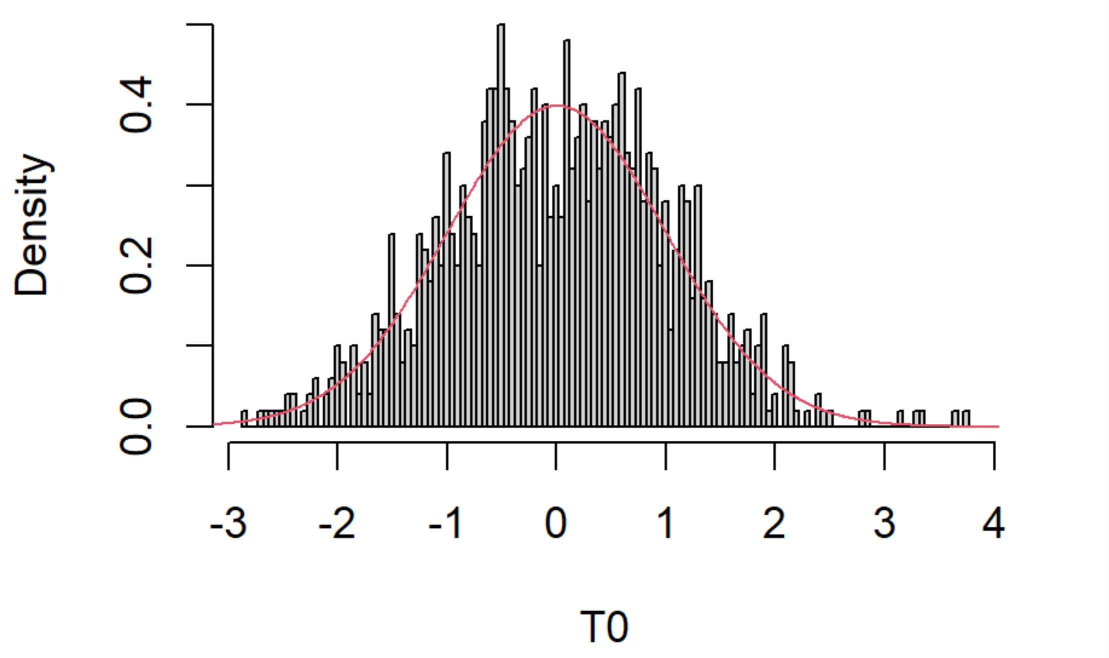
$$\sigma^2_{mean} = \sigma^2/n$$

Mean = μ? What is the distribution of T*?

if σ known

```
T0 <- (mu - res$mean)/
  (sigma/sqrt(N_obs))
hist(T0,
     breaks=getBreaks(T0, by=0.05),
     probability = T,
     main="")
x <- seq(-10,10,0.01)
lines(x,
      dnorm(x,
            sd= 1),
      col=2)
```

$T|H_0 \sim N(0,1)$

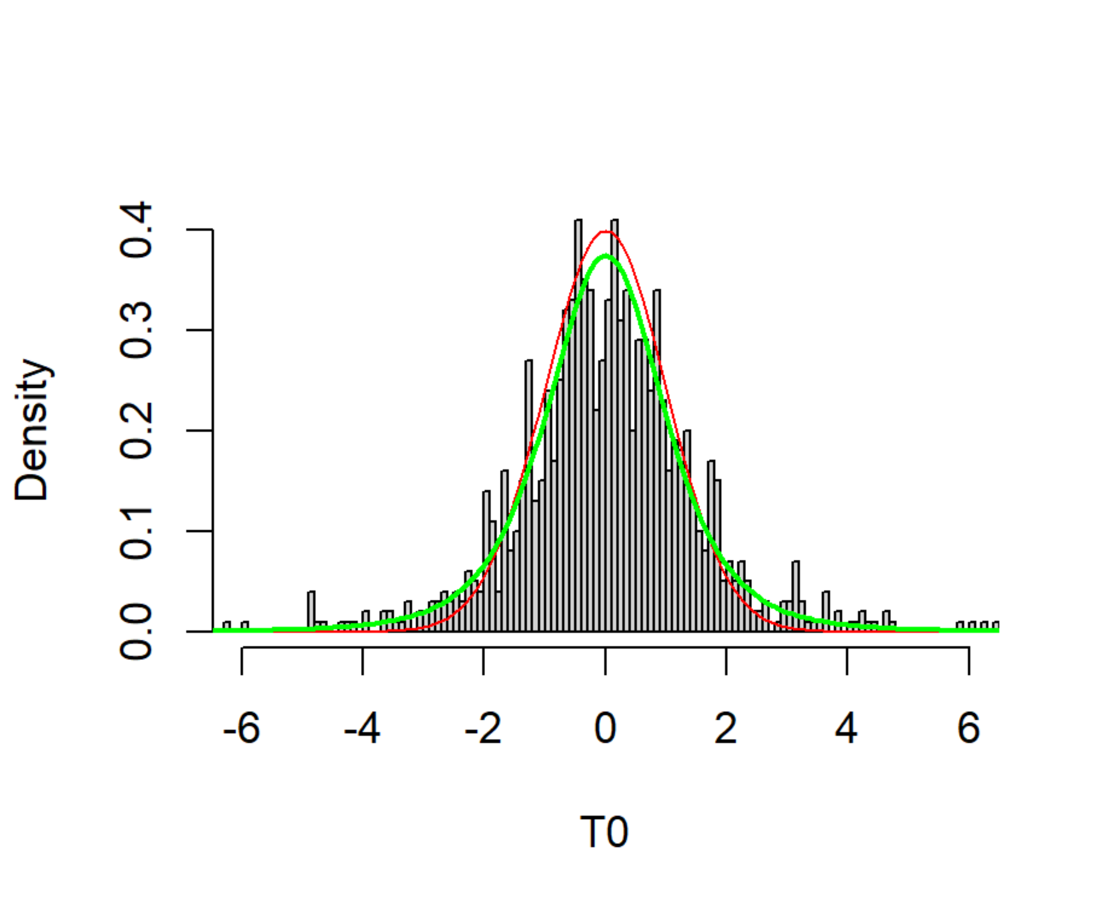


Mean = μ? Unknown Variance

if σ UNKNOWN

```
T0 ← (mu - res$mean) /
  (res$sd/sqrt(N_obs))
hist(T0,
     breaks=getBreaks(T0),
     probability = T, xlim=c(-6,6),
     ylim=c(0,0.4),main="")
x ← seq(-10,10,0.1)
lines(x, dnorm(x),
      col="red")
lines(x,dt(x,df = N_obs),
      type="l",col="green",lwd=2)
```

$$T|H_0 \sim T(\mu = 0, df = N_{obs})$$



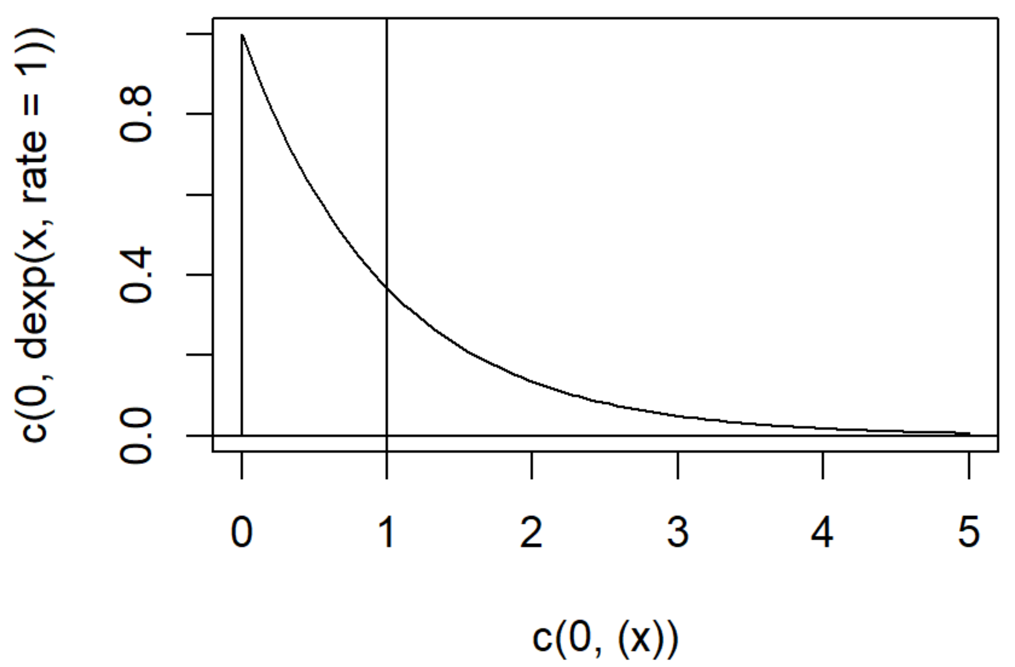
Mean = μ? if sampleing $x \sim Exp(1)$

Sampling from a skewed distribution.

$$x \sim Exp(1)$$

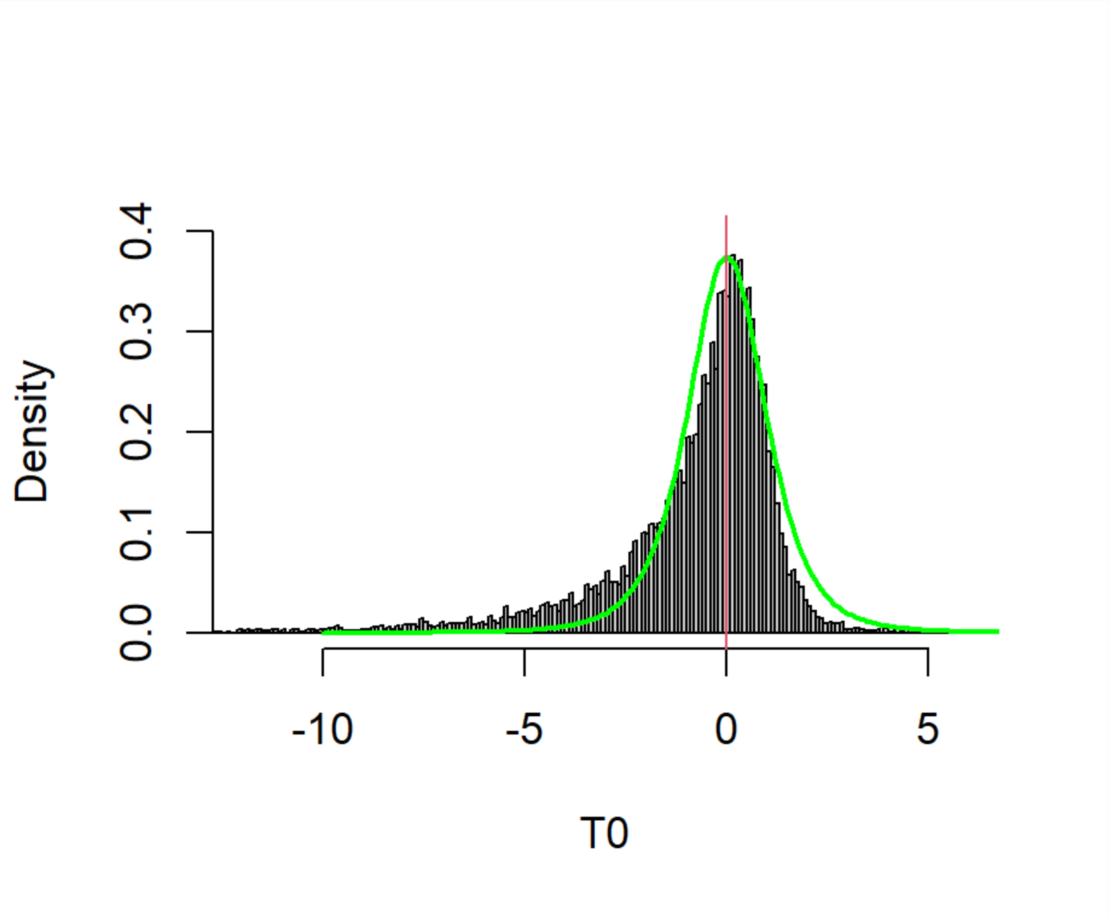
We know $\mu = \lambda = 1$

```
x ← seq(0,5,length = 100)
plot(c(0,(x)),
     c(0,dexp(x, rate = 1)),
     type = "l")
abline(h = 0)
abline(v = 1)
```



Mean = μ? $x \sim Exp(1)$ with $N_{obs} = 4$

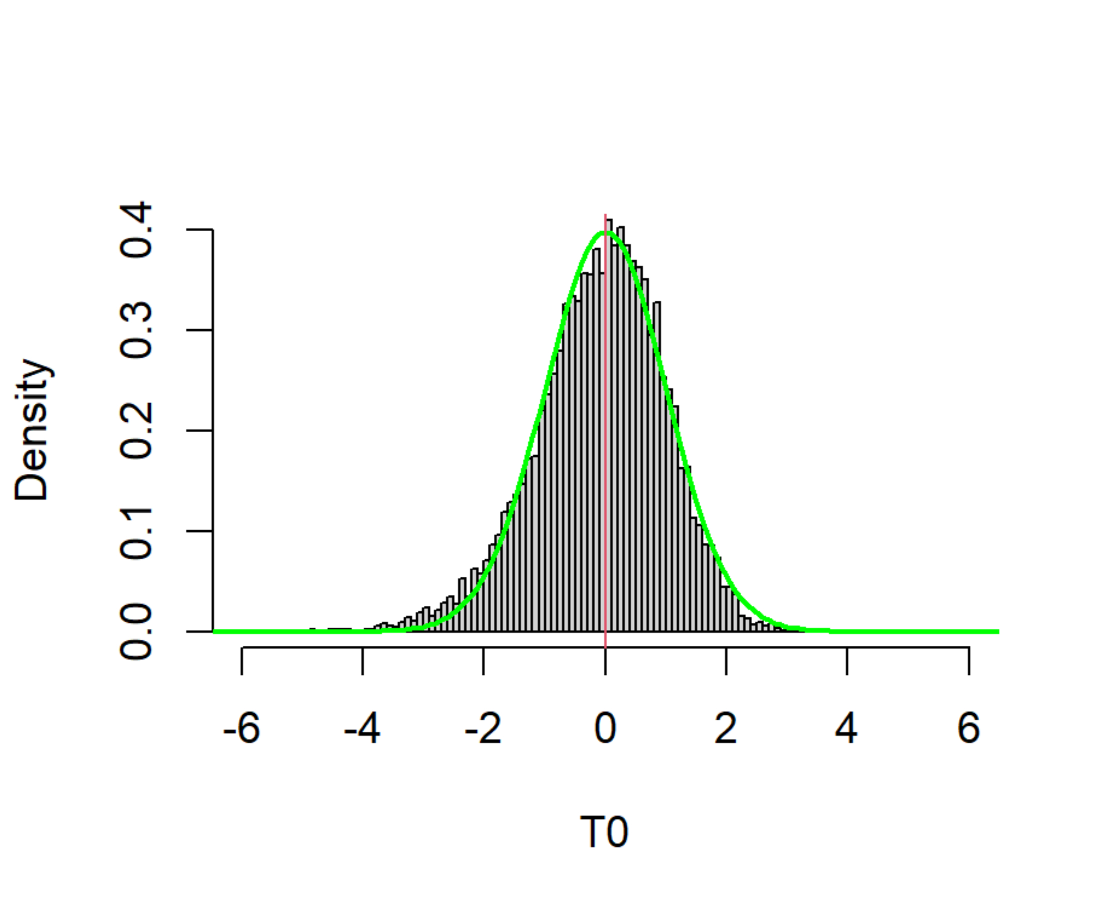
```r
N ← 10000;rate ← 1;
N_obs ← 4; mu ← 1;
bb_exp ← function(y){
  x ← rexp( N_obs, rate=rate )
  data.frame(mean = mean( x ),
             sd = sd(x))
  }
res ← purrr::map_df(1:N, bb_exp)
T0 ← (res$mean - mu)/
  (res$sd/sqrt(N_obs))
hist(T0, breaks=getBreaks(T0),
     probability = T, xlim=c(-12,6),
     ylim=c(0,0.4), main="")
x ← seq(-10,10,0.1)
lines(x,
      dt(x,df = N_obs),
      type="l",col="green",lwd=2)
abline(v = 0, col=2)
```



We simulate 4 datapoints from an exponential distribution. Observe how the Null distribution

Mean = μ? $x \sim Exp(1)$ but with $N_{obs} = 100$

```r
# Simulating data from Null
N <- 10000;rate <- 1
N_obs <- 100
bb_exp <- function(y){
  x <- rexp( N_obs, rate=rate )
  data.frame(mean = mean( x ), sd = sd(x))
}
res <- purrr::map_df(1:N, bb_exp)
T0 <- (res$mean - mu)/
  (res$sd/sqrt(N_obs))
hist(T0, breaks=getBreaks(T0),
     probability = T, xlim=c(-6,6),
     ylim=c(0,0.4), main="")
lines(x,dt(x,df = N_obs),type="l",
      col="green",lwd=2)
abline(v=0, col=2)
```



We simulate 100 datapoints from an exponential distribution. Observe how the Null distribution changed.

Central Limit Theorem

- In probability theory, the **central limit theorem** (CLT) establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed.
- The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.
- Some methods rely on **asymptotic properties** (e.g. `multcomp` p-value computation)

CLT in Proteomics?

- The error of transformed intensities in an LFQ experiment is normally distributed because it is the sum of biological, biochemical, and technical variability.
- Sample sizes are small. Therefore great care has to be taken to meet the requirement of normally distributed observations when using the t-test.

Types of tests

- parametric tests e.g. t-test
  - assume underlying statistical distributions in the data.
  - Therefore, several conditions of validity must be met so that the result of a parametric test is trustworthy
  - For example, Student's $t - test$ for two independent samples is reliable only if each sample follows a normal distribution and if sample variances are homogeneous.
- asymptotic tests
  - assume that methods which work for normal distributions work also elsewhere
- nonparametric tests e.g. randomization test
  - do not rely on any distribution. They can thus be applied even if parametric conditions of validity are not met.
  - robust to outliers
  - Parametric tests **often** have nonparametric equivalents.

Two sample t-test for equal means

- Null hypothesis - there is no such difference

- Test statistic:

$$T = \frac{Y_1 - Y_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

- Significance level α

- Reject the null hypothesis that the two means are equal if $|T| > t_{1-\alpha/2,v}$ with v degrees of freedom

$$v = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/(N_1 - 1) + (s_2^2/N_2)^2/(N_2 - 1)}$$

Two sample randomization tests for equal means

1. Suppose the 10 individuals in the study have been labelled

rowname	1	2	3	4	5
Diet.A	1	2	3	4	5
Diet.B	6	7	8	9	10

1. Randomly re-assign the 10 individuals to the two groups.
2. Re-calculate the test-statistic for this permuted data
3. Repeat 2 and 3 to obtain B sampled test-statistics, denoted $T_1, \ldots, T_B$.
4. For a two-sided test, the estimated p-value of the observed test statistic T_{obs} is:

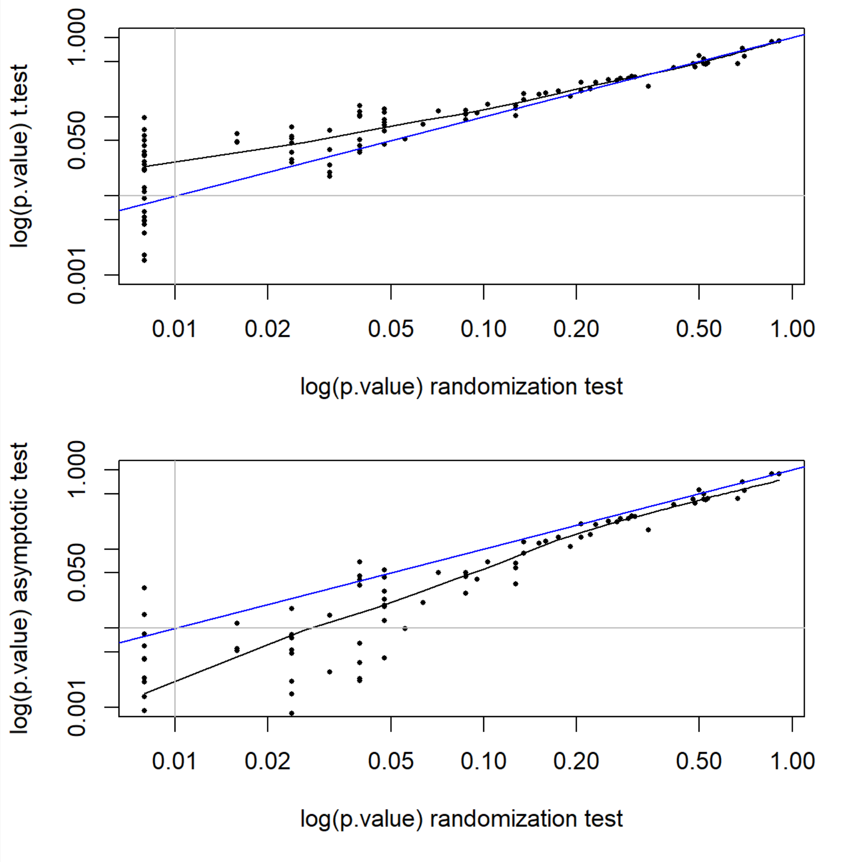
$$\frac{1}{B}\sum_{i=0}^{B} I_{T_i} >= |T_{obs}|$$

Step 1-3 generates a sample from the null distribution.

Compare Tests

- Simulate data from $x_1 \sim N(1,1)$ and $x_2 \sim N(10,10)$ (5 each)
- compute p-values using:
  - randomization test
  - t-test
  - asymptotic test (T under null $\sim N(\mu, \sigma)$)

	coin	t.test	asymp.test
Accept H0	74	90	57
Reject H0	26	10	43



Types of error

A **type I error** (false positive) occurs when
the null hypothesis (H0) is true, but is rejected.
The *type I error rate* or **significance level** (p-Value)
is the probability of rejecting the
null hypothesis given that it is true.

A **type II error** (false negative) occurs when
the null hypothesis is false,
but erroneously fails to be rejected.
The *the type II error rate* is denoted by the Greek letter β
and is related to the **power of a test** (which equals $1 - \beta$).

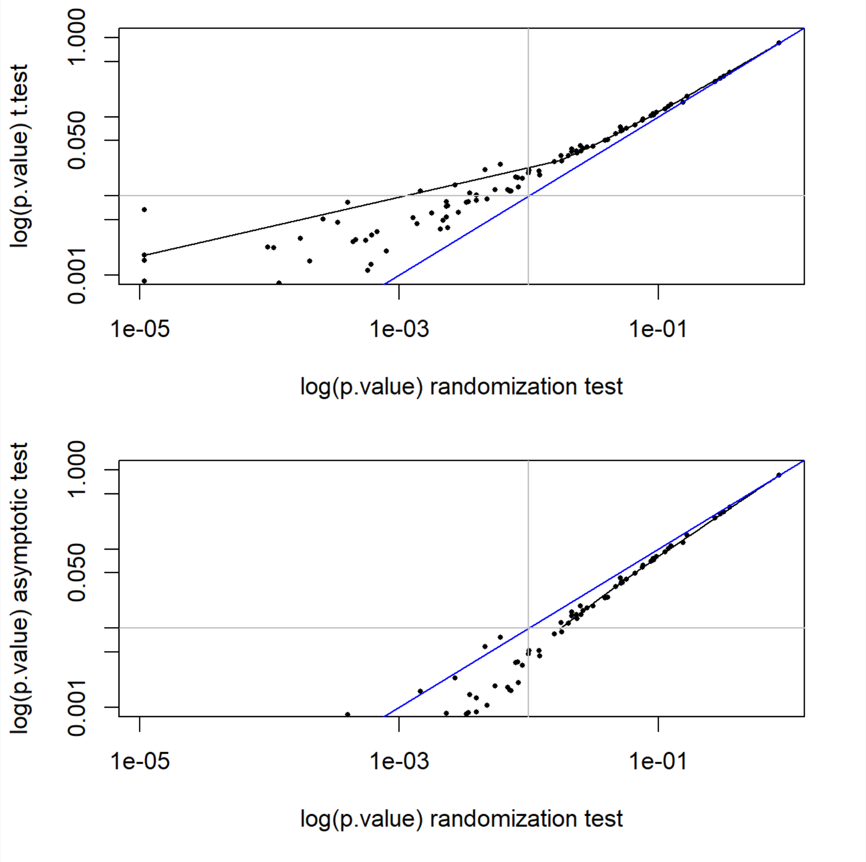
For a given test, the only way to reduce both error rates
is to **increase the sample size**, and this may not be feasible.

		reality	
		H_0 = true	H_0 = false
conclusion	H_0 is not rejected	OK	type II error
	H_0 is rejected	type I error	OK

Increasing sample size

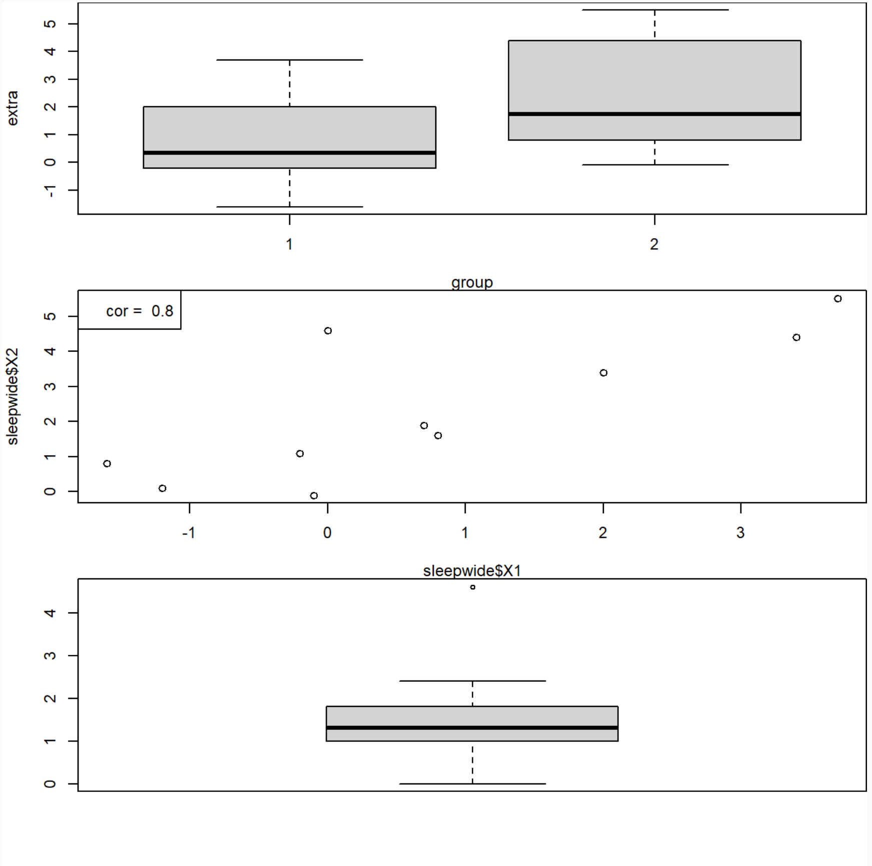
- Simulate data from $x_1 \sim N(1,1)$ and $x_2 \sim N(10,10)$ (20 each)
- compute p-values using:
  - randomization test
  - t-test
  - asymptotic test (T under null $\sim N(\mu, \sigma)$)

	coin	t.test	asymp.test
Accept H0	14	17	12
Reject H0	86	83	88



Repeated - correlated measurements

```
old_mar ← par()$mar
par(mfrow=c(3,1), mar=c(4,4,0,0))
plot(extra ~ group, data = sleep)
sleep %>% tidyr::spread(group, extra) →
  sleepwide
colnames(sleepwide) ← make.names(
  colnames(sleepwide)
  )
plot( sleepwide$X1, sleepwide$X2)
legend("topleft", legend =
       paste("cor = ",
       round(
         cor(sleepwide$X1, sleepwide$X2),
         digits=2)))
sleepwide ← sleepwide %>%
  dplyr::mutate(diff = X2-X1)
boxplot(sleepwide$diff)
par(mar = old_mar)
```



To see `sleep` dataset documentation run `?sleep` in R.

Repeated - correlated measurements

- test-statistics two groups

$$t_{unpaired} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

- test-statistics paired

$$t_{paired} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

with $\bar{d}$ the mean of the differences d_i with $i \in (1, \ldots, n)$, and $d_i = x_{2i} - x_{1i}$ (the correlated samples in condtion 1 and 2).

Repeated - correlated measurements

```
test.p.values <- data.frame(
unpaired.p =
  t.test(extra ~ group,
         data = sleep,
         paired = FALSE)$p.value,
paired.p =
  t.test(extra~group,
         data = sleep,
         paired = TRUE)$p.value,
diff.p =
  t.test(sleepwide$diff)$p.value
)
```

unpaired.p	paired.p	diff.p
0.079	0.0028	0.0028

Top code block - two sample t-test, middle code block - paired t-test, bottom code - one sample t.test on differences. Note that the paired t-test gives the same results as the one sample t.test of differences

Missing data

```r
sleepless ← datasets::sleep
sleepless$extra[c(1,4,6,12)] ← NA
sleepless$extra[1:4]
```

```
## [1]   NA -1.6 -0.2   NA
```

```r
tryCatch(
  t.test(extra ~ group, data = sleepless, paired =TRUE),
  error = function(e) e)
```

```
## <simpleError in complete.cases(x, y): not all arguments have the same length>
```

running the paired t-test with missing data fails.

Linear models

```r
lm1 <- lm(extra ~ group, data = sleep)
lm2 <- lm(extra ~ group + ID, data = sleep)
lmermod <-
  lmerTest::lmer(extra ~ group + (1|ID),
                 data = sleep)

x <- bind_rows(
broom::tidy(anova(lm1))[1,],
broom::tidy(anova(lm2))[1,],
broom::tidy(anova(lmermod))[1,],
)

xx <- add_column(x, model =
        c("lm_1","lm_2","lmer"),
        .before = 1) %>%
  dplyr::select(model, p.value) %>%
  mutate(p.value = signif(p.value, digits=2))
```

model	p.value
lm_1	0.0790
lm_2	0.0028
lmer	0.0028

performing the same t-tests but using linear models and mixed effect linear models produces same result.

Linear models - missing data

```
lm1 <- lm(extra ~ group,
          data = sleepless)
lm2 <- lm(extra ~ group + ID,
          data = sleepless)
lmermod <- lmerTest::lmer(
  extra ~ group + (1|ID),
  data = sleepless)
x <- bind_rows(
broom::tidy(anova(lm1))[1,],
broom::tidy(anova(lm2))[1,],
broom::tidy(anova(lmermod))[1,],
)

xx <- add_column(x,
        model = c("lm_1","lm_2","lmer"),
        .before = 1) %>%
  dplyr::select(model, p.value) %>%
  dplyr::mutate(p.value = signif(p.value, digits =
```

model	p.value
lm_1	0.077
lm_2	0.022
lmer	0.029

Linear models do work also with missing data.

Conclusion

- What is a hypothesis test
- How to report results of hypothesis tests?
  (you do not report p-values
  state if you reject null given your size of test α
)
- If assumptions in parametric tests are not met
  - null distribution is wrong =>
    p-value estimate is wrong
    Except?
- Understand CLT and what assymptotic properties are.
- Parametric tests do not make as many assumptions
  about the data.