Statistical Inference

Sébastien Plat

Contents

Statistical Inference

Definition

Statistical inference is the process of drawing general conclusions about a population by:

- using noisy **statistical data** (samples)
- quantifying the **uncertainty** associated with those conclusions

Note: The uncertainty could arise from incomplete or bad data.

Terms

First, a **statistic** (singular) is a number computed from a **sample of data**. We use statistics to infer information about a population.

Second, a **random variable** is an **outcome from an experiment**. Deterministic processes, such as computing means or variances, applied to random variables, produce additional random variables which have their own distributions.

Random variables are said to be iid if they are **independent and identically distributed**.

- **Independent** means "statistically unrelated from one another"
- **Identically distributed** means that "all have been drawn from the same population distribution"

IID random variables are the default model for random samples; many of the important theories assume that variables are iid. We'll usually assume that our samples are random and that variables are iid.

Finally, there are two broad flavors of inference:

- **Frequency**, which uses "long run proportion of times an event occurs in iid repetitions."
- **Bayesian**, in which the probability estimate for a hypothesis is updated as additional evidence is acquired.

Both flavors require an understanding of probability: quantifying the likelihood of particular events occurring. For a given experiment, the probability of a particular outcome is the number of ways that outcome can occur divided by all the possible outcomes.

*Please refer to **SI01 - probability.md** to know more about probabilities.*

Exponential Distribution

*Please refer to **SI03 - common distributions.md** to know more about Binomial, Gaussian and Poisson distributions.*

Definition

According to Wikipedia:

> The exponential distribution [...] is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant **average rate** λ.

- Its mean is $1/\lambda$
- Its standard deviation is also $1/\lambda$

In our example, we will use λ=**0.2**.

Samples mean & variance

According to the Law of Large Numbers, samples mean & sample variance are **consistent estimators** of the populations mean & variance: they converge to the correct value as the number of samples increases.

Let's study the distribution of 1000 averages of 40 exponentials. Fig.1 and Fig.2 show the distribution of sample mean & sample variance, plus their mean and theoretical values:

Fig. 1: Sample Mean PDF

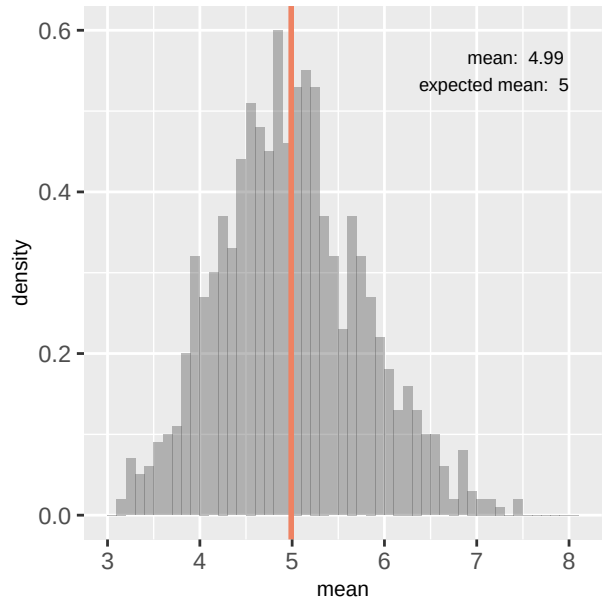
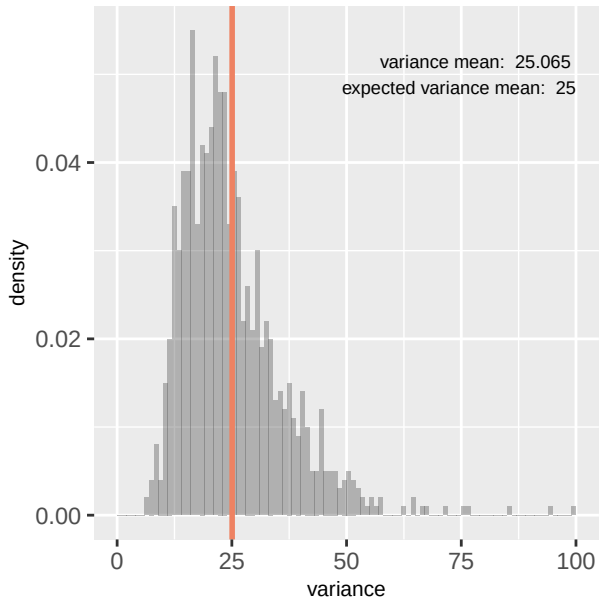


Fig. 2: Sample Variance PDF



We clearly see that **their mean is close to the value they estimate**.

*Please refer to **SI02 - random variables.md** to know more about probability distributions, mean and variance.*

Asymptotics & Central Limit Theorem

Definition

Asymptotics describes how statistics behave as sample sizes get very large and approach infinity. This is useful for making **statistical inferences** and approximations.

The **Central Limit Theorem** states that, according to the Law of Large Numbers:

> The **distribution $\bar{X}$ of sample means** of iid obeservations (from a population of mean = μ, sd = σ) will become **normal**, or nearly normal, as the sample size n increases. It will be approximated by:
>
> $\bar{X} \sim N\ (\mu, \sigma^2/n)$

An important point is that the CLT works **even if the original distribution is not normal**.

Example

Fig. 3 shows the PDF for our distribution (λ=0.2) vs a Normal with the same mean and standard deviation.

Fig.4 shows the empirical distribution of samples mean *(cf. Fig.1)* VS the CLT predicted one:

- $Est = 1/\lambda = 5$
- $SE_{Est} = 1/(\lambda \times \sqrt{n}) = 0.625$

Fig. 3: PDF − Exponential vs Normal

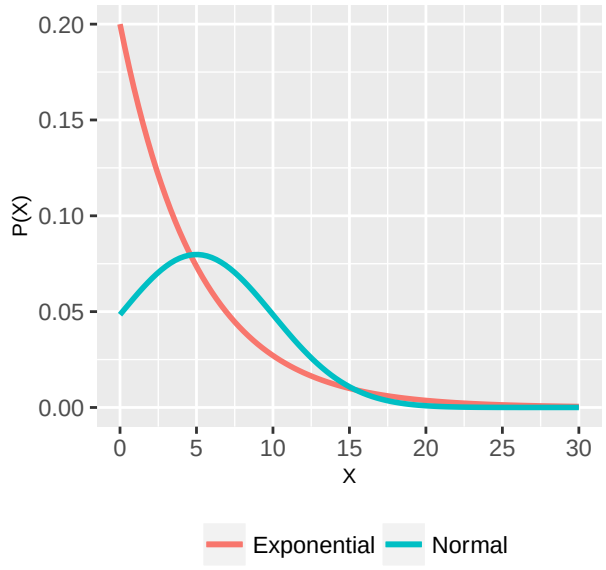
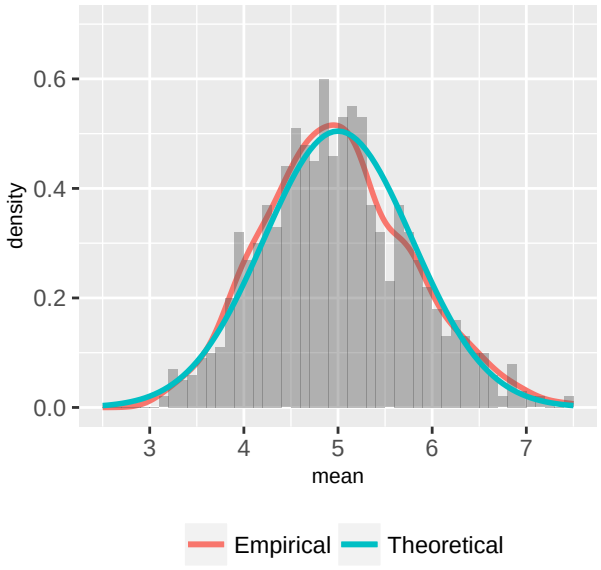


Fig. 4: PDF − Samples Mean vs Normal



We clearly see that:

- our exponential distribution is **not even close to being normal**
- **the empirical distribution is very close to being normal**, as predicted by the CLT

CLT Confidence Interval

Definition

As the distribution of sample means $\bar{X}$ is roughly normal (mean = μ and sd = $\sigma/\sqrt{n}$), we have for each of our samples:

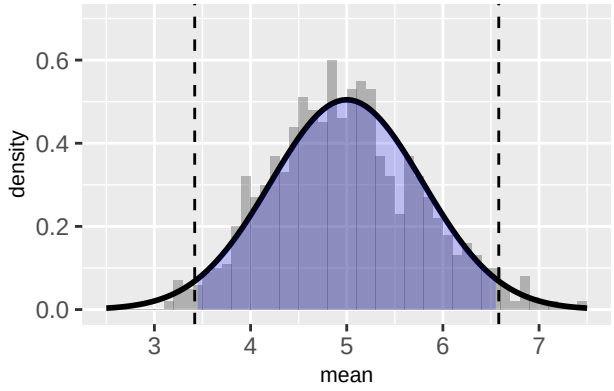
$P\ (\bar{X} < mean - 2sd) = P\ (\bar{X} < \mu - 2\sigma/\sqrt{n}) \simeq 0.025$

$P\ (\bar{X} > mean + 2sd) = P\ (\bar{X} > \mu + 2\sigma/\sqrt{n}) \simeq 0.025$

It means that:

$P\ (\ \bar{X} \in [\mu \pm 2\sigma/sqrtn]\ ) \simeq 0.95$

Fig. 5: Samples Mean 95% Interval



We can deduce from above that:

$P\ (\ \mu \in [\bar{X} \pm 2\sigma/\sqrt{n}]\ ) \simeq 0.95$

$\bar{X} \pm 2\sigma/\sqrt{n}$ is called the **95% Confidence Interval** for μ. It mean that for 95% of our samples, this interval will include μ. But **it does not mean that μ is actually in this interval**.

- CI get wider as the coverage increases
- CI get narrower as the sample size increases & with less variability

The confidence interval represents values for the population parameter for which the difference between the parameter and the observed estimate is not statistically significant at the 5% level.

It means that, if the true value of the parameter lies outside the 95% confidence interval once it has been calculated, then an event has occurred which had a probability of 5% (or less) of happening by chance.

Empirical estimation

The CLT states that: • $mean_{Est} \simeq \mu$ • $SD_{Est} \simeq \sigma$ • $\bar{X} \sim N\ (\mu, \sigma^2/n)$ • so the CI is:

$$mean_{Est} \pm ZQ_{1-\alpha/2} \times SE_{Est} = mean_{Est} \pm ZQ_{1-\alpha/2} \times \frac{SD_{Est}}{\sqrt{n}}$$

T Distribution

Definition

The CLT works only for large enough sample sizes. For smaller ones, the Gosset's t distribution is more relevant. It is assumed the population is an iid normal (or roughly symetrical & mound-shaped): the t-interval does not work well with skewed data.

In that case:

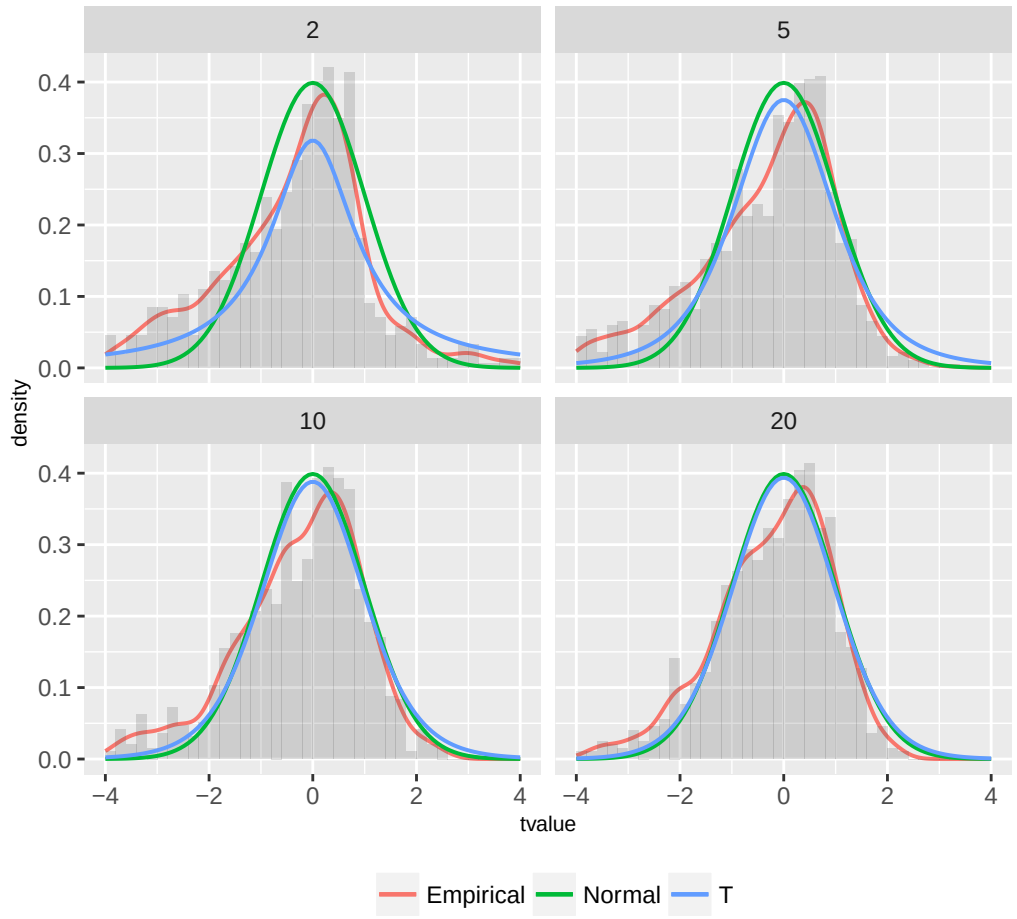
$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a t-distribution with n-1 degrees of freedom *(replacing S by σ would give exactly a standard normal)*. Its tails are **thicker than normal**, so its Confidence Interval is **wider** for the same Confidence Level. This is because estimating the population standard deviation introduces more uncertainty.

Example

Back to the Exponential Distribution, Fig.6 shows the experimental sample distribution vs T vs Normal for different sample sizes. The t-distribution gets close to normal even for relatively small sample sizes.

Fig. 6: PDF − Samples Mean vs Normal vs T



T Confidence Intervals

Definition

T Confidence Intervals are slightly different from CLT ones:

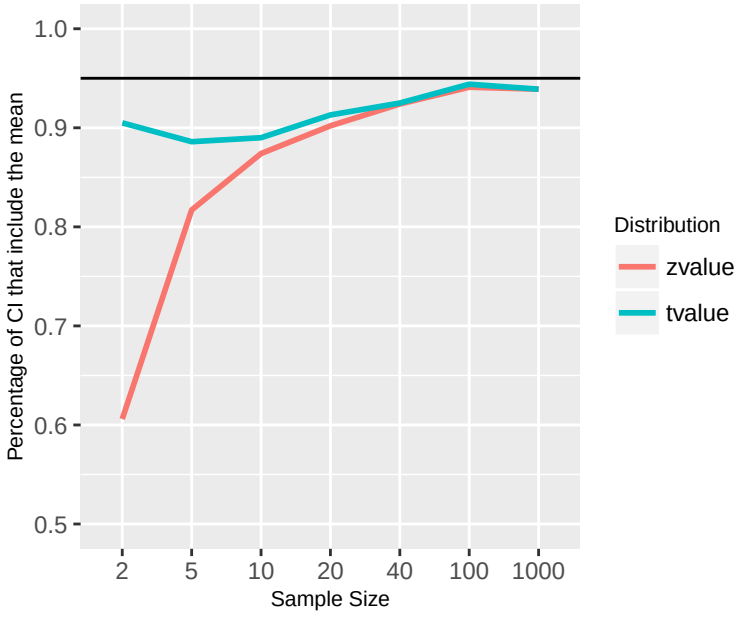
$$mean_{Est} \pm TQ_{1-\alpha/2,n-1} \times SE_{Est} = mean_{Est} \pm TQ_{1-\alpha/2,n-1} \times \frac{SD_{Est}}{\sqrt{n}}$$

Comparizon of CLT vs T Confidence Intervals

Back to the Exponential Distribution, Fig.7 shows how CLT and T Confidence Intervals perform for different sample sizes.

```
##  size             zConf                tConf
##     2   [ -2.093 , 12 ] [ -40.731 , 50.638 ]
##     5 [ 0.658 , 9.334 ]  [ -1.149 , 11.141 ]
##    10 [ 1.905 , 8.044 ]    [ 1.431 , 8.517 ]
##    20 [ 2.906 , 7.182 ]    [ 2.761 , 7.327 ]
##    40 [ 3.403 , 6.564 ]    [ 3.353 , 6.614 ]
##   100  [ 4.033 , 5.96 ]    [ 4.021 , 5.972 ]
##  1000 [ 4.678 , 5.326 ]    [ 4.677 , 5.326 ]
```

Fig. 7: 95% Confidence Interval − Normal vs T



The T-interval is, as expected, **much more reliable for small sample sizes**. The behaviour of the two methods converge when the sample size increases.

Comparing two populations

Definition

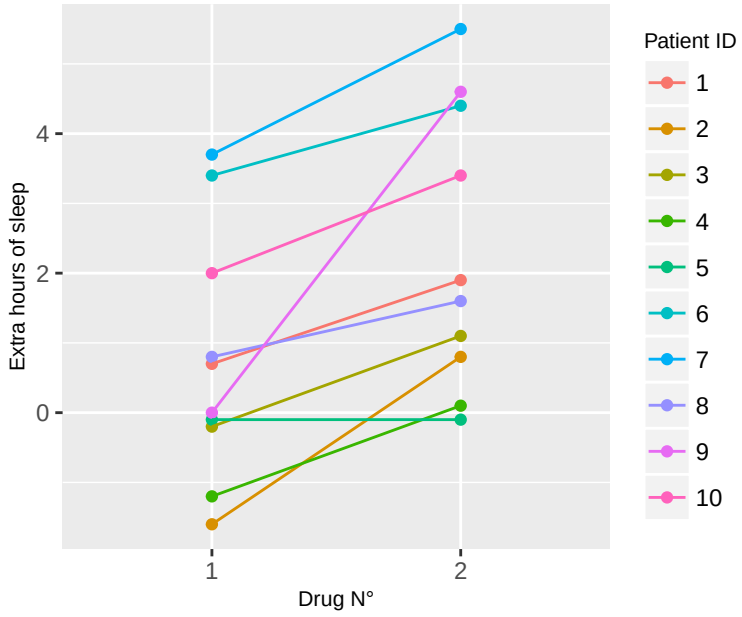
T-intervals are very useful to compare two populations.

> Confidence intervals of difference between populations that **do not contain 0** imply that there is a **statistically significant difference** between the populations.

Example

A classic example is the sleep data analyzed in Gosset's Biometrika paper. Fig.8 shows the increase in hours of sleep for 10 patients on two soporific drugs:

Fig. 8: Increase in hours of sleep − drug N°1 vs N°2



It seems that drug N°2 is more efficient than drug N°1. To confirm this hypothesis, we can take a t-test to calculate the T Confidence Interval of their difference.

```
g1 <- sleep$extra[sleep$group==1]; g2 <- sleep$extra[sleep$group==2]
ttest <-t.test(g2, g1, paired = TRUE)
result <- as.data.frame (cbind(round(ttest$conf, 3)[1],
                               round(ttest$conf, 3)[2],
                               round(ttest$p.value, 5)))
names(result) <- c("Lconf", "Uconf", "p.value")
print(result)
```

```
##   Lconf Uconf p.value
## 1   0.7  2.46 0.00283
```

The T Confidence Interval does not include 0 and P < 0.005, so **drug N°2 is statistically more efficient than drug N°1**: with 95% probability, the average difference of effects for an individual patient is between .7 and 2.46 additional hours of sleep.

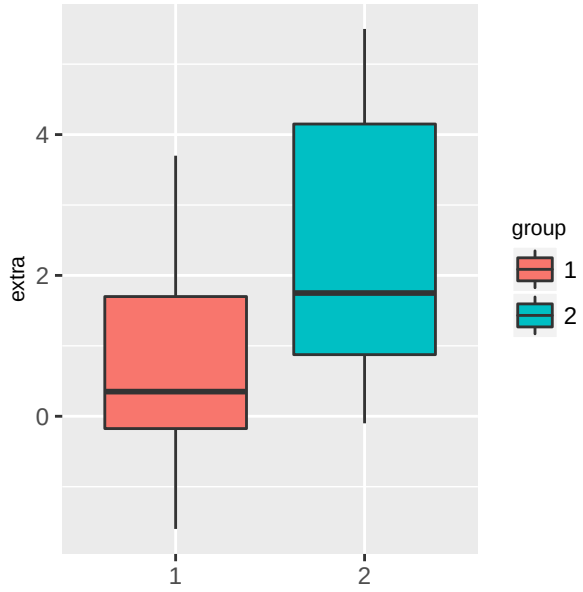
Generalization

A generalization of the t-test can be used for comparing independant groups (different sample sizes, etc.) with or without equal variance. Performing it on Gosset's sleep data gives the following results:

```
##   Paired EqVar  Lconf Uconf p.value
## 1      1     1  0.700 2.460 0.00283
## 2      0     1 -0.204 3.364 0.07919
## 3      0     0 -0.205 3.365 0.07939
```

By omitting the information that the two populations are paired, the results become less clear (but equal & unequal variance give a very similar CI). We can easily see why by studing the two distributions, as shown Fig.9.

Fig. 9: Increase in hours of sleep drug N°1 vs N°2 as independent groups



Hypothesis Testing

Principle

Hypothesis testing is the use of statistics to determine the **probability that a given hypothesis is true**. It is used to make decisions about populations using observed data.

The usual process of hypothesis testing consists of four steps:

1. **Formulate** the null hypothesis H_0 and the alternative hypothesis H_a. Commonly:
   - H_0: the observations are the result of pure chance
   - H_a: the observations show a real effect combined with a component of chance variation
   - **Statistical evidence** is required to reject H_0 in favor of the alternative hypothesis
2. Identify a **test statistic** that can be used to assess the truth of H_0.
3. **Compute** the P-value. The **smaller** the P-value, the **stronger** the evidence **against** H_0.
4. **Compare** the P-value to an acceptable significance value α. If $P \leq \alpha$:
   - the observed effect is **statistically significant**
   - the null hypothesis is ruled out, and the **alternative hypothesis** is **valid**

Outcomes

There are four possible outcomes of our statistical decision process:

Truth	Decide	Result
H_0	H_0	Correctly accept null
H_0	H_a	Type I error α
H_a	H_a	Correctly reject null
H_a	H_0	Type II error β

- Type I error: REJECTS a TRUE null hypothesis H_0
- Type II error: ACCEPTS a FALSE null hypothesis H_0

Since there's some element of **uncertainty** in questions concerning populations, we deal with **probabilities**. In our hypothesis testing we'll set the probability of **making errors small**.

The probabilities of making these two kinds of errors are related. If you decrease the probability of making a Type I error (rejecting a true hypothesis), you increase the probability of making a Type II error (accepting a false one) and vice versa.

P-value and Alpha

> The **P-value** is the probability that a test statistic at least as significant as the one observed would be obtained if H_0 were true.

We reject H_0 when P $< \alpha$. It means that α is the **Type I error rate** or **level** 3= Probability of rejecting H_0 when it is correct.

Example: sample mean

Let's suppose we have a sample of mean $= \bar{X}$ and standard deviation S. Our hypothesis H_0 is that the mean of the population from which our sample is drawn is μ_0:

$H_0 : \mu = \mu_0$

Under H_0, the sample mean distribution $Est \sim N\ (\mu_0, S/\sqrt{n})$. We want to measure how far from μ_0 our sample mean is: if it is too far away to be statistically likely, we can reasonably reject H_0. Otherwise, we will **fail to reject** H_0.

To challenge H_0, we will consider the Test Statistic:

$$TS = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

This reduces the Hypothesis Testing to the following table, where the TS is called the **Z-score**:

Alternate Hyp.	Reject H_0 if
$\mu < \mu_0$	$TS \leq Z_\alpha = -Z_{1-\alpha}$
$\mu \neq \mu_0$	$\lvert TS \rvert \geq Z_{1-\alpha/2}$
$\mu > \mu_0$	$TS \geq Z_{1-\alpha}$

For small sample sizes, the T-test is performed the same way:

Alternate Hyp.	Reject H_0 if
$\mu < \mu_0$	$TS \leq t_{\alpha,n-1} = -t_{1-\alpha,n-1}$
$\mu \neq \mu_0$	$\lvert TS \rvert \geq t_{1-\alpha/2,n-1}$
$\mu > \mu_0$	$TS \geq t_{1-\alpha,n-1}$

Link with Confidence Interval

When we test $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$, we fail to reject H_0 for all values $\bar{X}$ where $TS \leq Z_{1-\alpha/2}$. This set is a $(1-\alpha)100\%$ confidence interval for μ.

The same works in reverse; if a $(1-\alpha)100\%$ interval contains μ_0, then we **fail to reject** H_0.

Power

Definition

A type II error is **failing to reject H_0 when it's false**. Its probability is usually called β.

Its opposite is the **probability of rejecting H_0 when it is false**. It is called **power**: $power = 1 - \beta$

Reminder: α is the **probability of rejecting H_0 when it is correct**.

Power comes into play when you're designing an experiment, and in particular, if you're trying to determine **if a null result** (failing to reject a null hypothesis) is **meaningful**.

For instance, you might have to determine if your sample size was big enough to yield a meaningful, rather than random, result.

Link with Type I Error

Let's consider the hypothesis $H_a : \mu = \mu_a > \mu_0$. In that case **(same for t-tests)**:

$$\alpha = P\left(TS > Z_{1-\alpha}\ ;\ \mu = \mu_0\right) \quad \bullet \quad Power = P\left(TS > Z_{1-\alpha}\ ;\ \mu = \mu_a\right)$$

Power increases as:

- α increases
- n gets larger
- μ_a gets further away from μ_0
- S decreases

If $H_a : mu \neq mu_0$ we would calculate the one sided power using alpha / 2 in the direction of mu_a (This is only approximately right, it excludes the probability of getting a large test statistic in the opposite direction of the truth). As α is bigger than $\alpha/2$, power of a one-sided test is greater than the power of the associated two sided test.

Calculating Power

The quantities mu_0 and α are specified by the test designer. The other four quantities (β, σ, n, and μ_a), are all unknown. Knowing three of these, we can solve for the missing fourth: usually $power = 1 - \beta$ or the sample size n.

Note that:

- only the TS $\sqrt{n} * (\mu_a - \mu_0)/\sigma$ is needed
- $(\mu_a - \mu_0)/\sigma$ is called the **effect size**: the difference in the means in standard deviation units.
- the effect size is **unit free**, so it can be interpreted in different settings.

We can calculate the Power in R:

```
power.t.test (n = sampleSize, delta = mu_a - mu_0, sd = sigma, ...)$power # power for a fixed setting
power.t.test (power = requiredPower, delta = mu_a - mu_0, sd = sigma, ...)$n # n for specific power
```

Example

Fig.10-11 show an example of $H_a : \mu_a < \mu_0$ and an example of $H_a : \mu_a > \mu_0$. The vertical black bar shows the measured mu value below or above which we can statistically reject H_0 ($\alpha = 0.05$), and the corresponding power (that depends on the value of μ_a).

Fig. 10: alpha vs power; muA = mu0 − 4

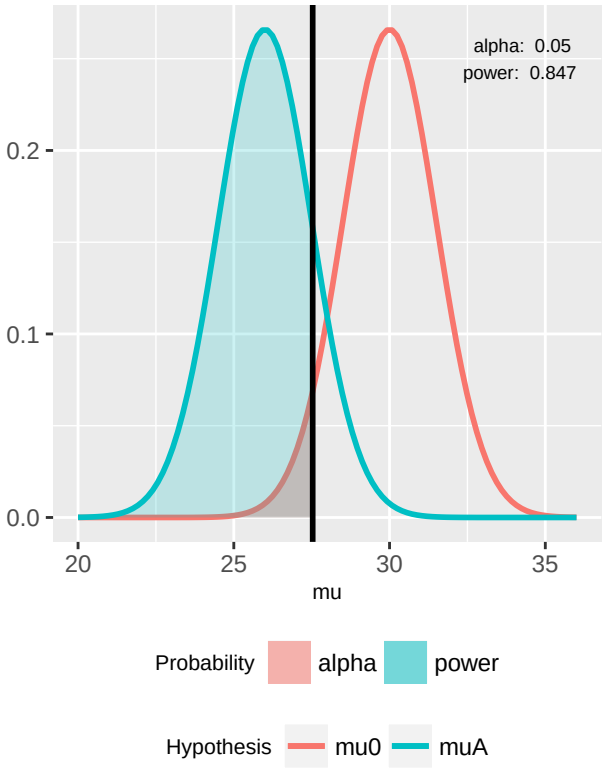
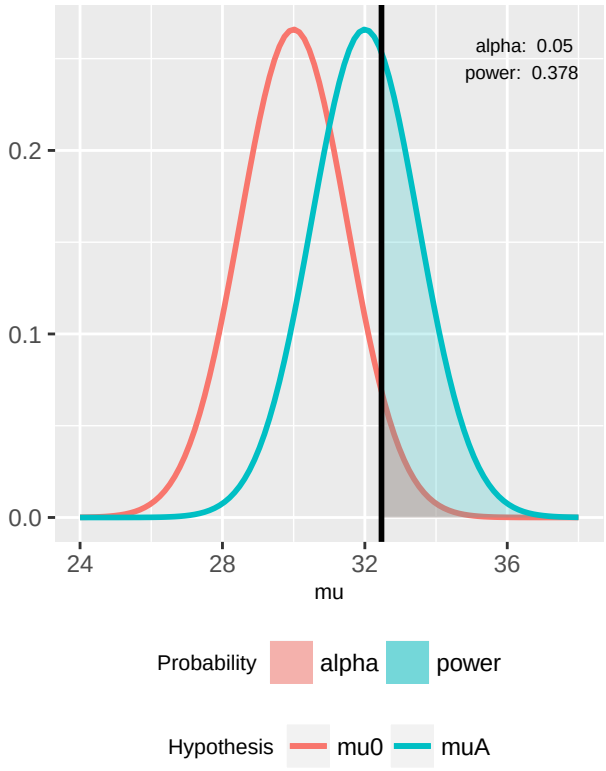


Fig. 11: alpha vs power; mu1 = mu0 + 2



Multiple Testings

Multiple testing is particularly relevant now in this age of BIG data. Statisticians are tasked with questions such as:

- Which variables matter among the thousands measured?
- How do you relate unrelated information?

When doing a large number of tests, even small Type I/Type II Error Rates can lead to many erroneous results. Technics exist to limit this.

Type of errors

When performing m tests on H_0 vs H_a, we have the following outcomes:

Test Result	Chosen Hyp	H_0 is true	H_a is true	N° of results	Accuracy Rate
$P > \alpha$ *(negative)*	H_0	U	T $[\beta]$	$m - R$	Specificity: $U/(m-R)$
$P < \alpha$ *(positive)*	H_a	V $[\alpha]$	S	R	Sensitivity: S/R
–	N° of tests	m_0	$m - m_0$	m	

- Type I error *(false positive)*: V results. Say H_0 is false when it is not
- Type II error *(false negative)*: T results. Say H_0 if true when it is not

Note: we reject H_0 when outcomes are considered significant (ie. positive), ie $P < \alpha$.

Error rates

False positive rate - The rate at which false results (falsely rejecting H_0) are called significant: $E\left[\frac{V}{m_0}\right]$

Family wise error rate (FWER) - The probability of at least one false positive $\Pr(V \geq 1)$

False discovery rate (FDR) - The rate at which claims of significance are false: $E\left[\frac{V}{R}\right]$

Control FWER: Bonferroni correction

Idea: $Pr(V \geq 1) < \alpha$

- Set $\alpha_{fwer} = \alpha/m$
- Call any $P_{(i)} \leq \alpha_{fwer}$ significant

Pros: conservative **Cons**: May be very conservative

Controlling FDR: Benjamini-Hochberg method (BH)

Idea: $E\left[\frac{V}{R}\right] < \alpha$

- Order the P-values from smallest to largest $P_{(1)}, ..., P_{(m)}$
- Call any $P_{(i)} \leq \alpha \times \frac{i}{m}$ significant

Pros: less conservative (maybe much less) **Cons**: more false positives, may behave strangely under dependence

Example

```
# 50% true positive: y = f(x)
set.seed(1010093)
pValues <- rep(NA, 1000)
for (i in 1:1000) {
  x <- rnorm(20)
  # First 500 beta=0, last 500 beta=2  (regression slope)
  if (i <= 500) { y <- rnorm(20) }
  else { y <- rnorm(20, mean = 2 * x) }
  pValues[i] <- summary(lm(y ~ x))$coeff[2, 4]
}

# not zero = significant; We get 24 False positive - around 5%
trueStatus <- rep(c("zero", "not zero"), each = 500)
table(pValues < 0.05, trueStatus)
```

```
##        trueStatus
##         not zero zero
##   FALSE        0  476
##   TRUE       500   24
```

```
# Controls FWER - no more false positive, but the threshold is so low we now have 23 false negative
table(p.adjust(pValues, method = "bonferroni") < 0.05, trueStatus)
```

```
##        trueStatus
##         not zero zero
##   FALSE       23  500
##   TRUE       477    0
```

```
# Controls FDR - less false positive, no false negative
table(p.adjust(pValues, method = "BH") < 0.05, trueStatus)
```

```
##        trueStatus
##         not zero zero
##   FALSE        0  487
##   TRUE       500   13
```

Resampling

Bootstrapping

Bootstrapping is a technique that uses simulation and computation to infer **distributional properties** you might not otherwise be able to determine.

Its basic principle is to use OBSERVED data as a **substitute for the population**. We can simulate observations by doing **random sampling with replacement**. From this distribution (constructed from the observed data), we can **estimate the distribution** of the statistic we're interested in. This lets us better understand the underlying population (from which we didn't have enough data).

We can easily do it in R *(see also the R package bootstrap)*:

```
# create a matrix of N new samples of the same size as obsData
newMatrix <- matrix (sample (obsData, obsSize*N,replace=TRUE), N, obsSize)

# calculate the statistic value of each new sample (here, the median)
statVector <- apply (newMatrix, 1, median)

# calculate a CI for the resulting statistic
quantile(statVector, c(.025,.975))
```

Permutation testing

Permutation testins is also based on resampling a single dataset, but it measures whether or not outcomes are independent of group identity (ie. if group labels are exchangable). The resampling simply permute group labels associated with outcomes.

The idea is to see if the dataset value is statistically likely when the outcomes are independent of group identity. To do so, we can estimate the statistic distribution by resampling, and calculate P.