Statistical Inference

Exponential Distribution

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.

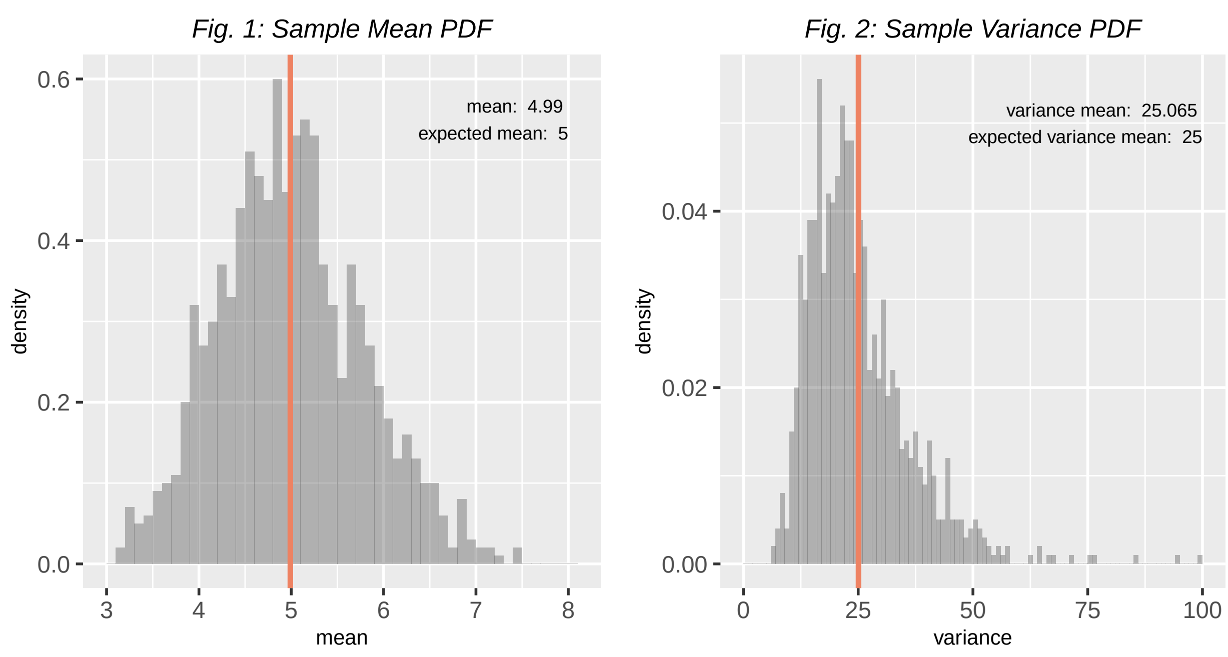
- Its mean is 1/ λ
- Its standard deviation is also 1/ λ

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

Samples mean & variance

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:



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

Central Limit Theorem

Definition

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

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

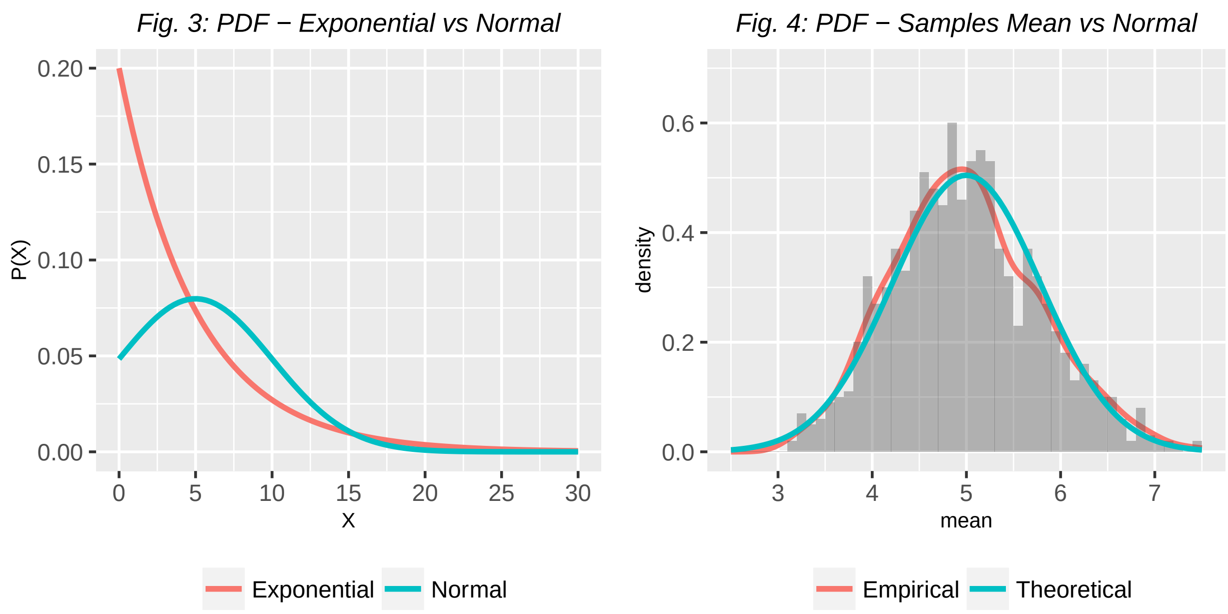
An important point is that the CLT works **even if the original distribution is not normal**.

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:

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



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

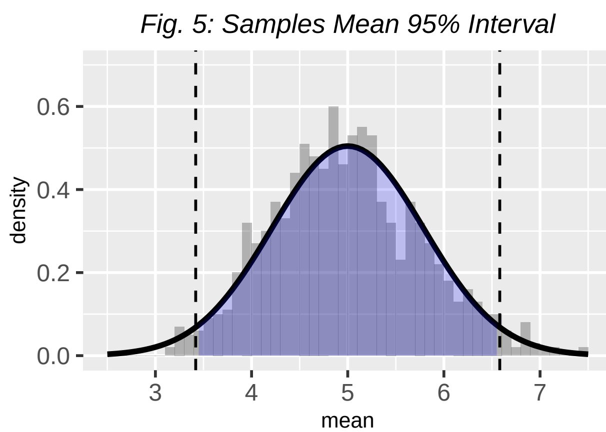
As the distribution of sample means $\bar{X}$ is roughly normal (mean = μ and sd = $\sigma/\sqrt{n}$), we have for each of our samples:

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

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

It means that:

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



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:

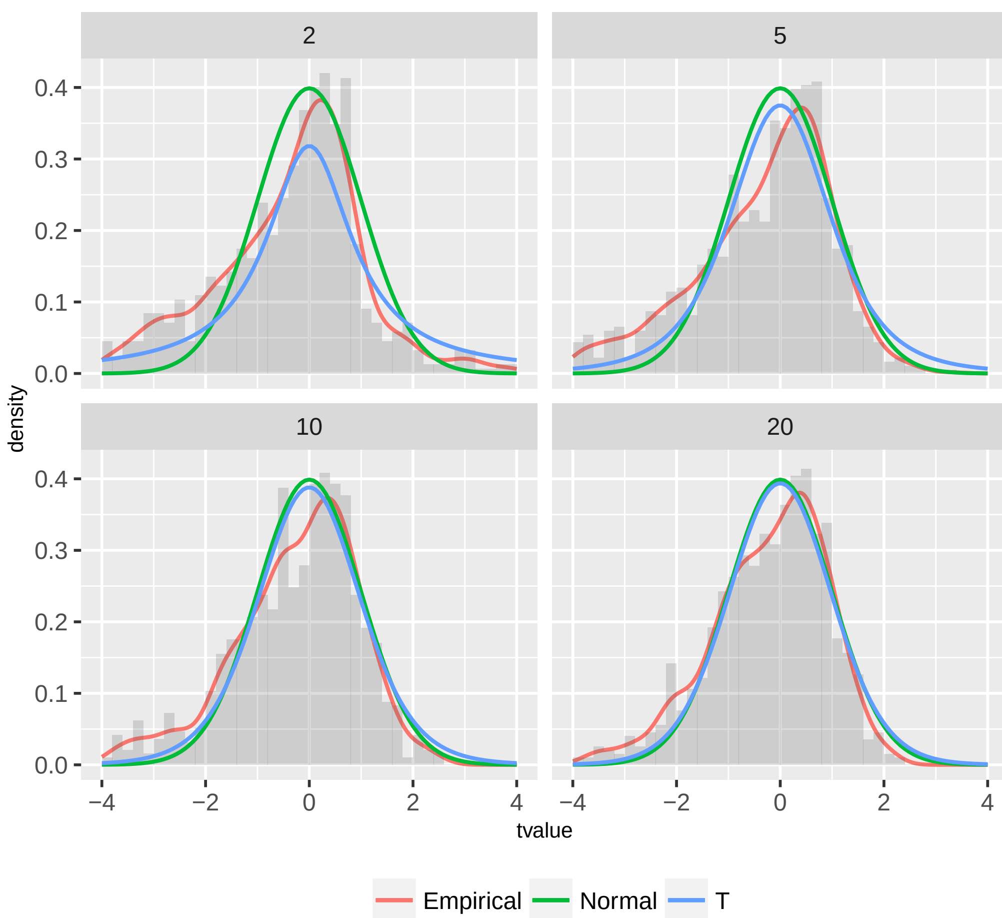
$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

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.

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:

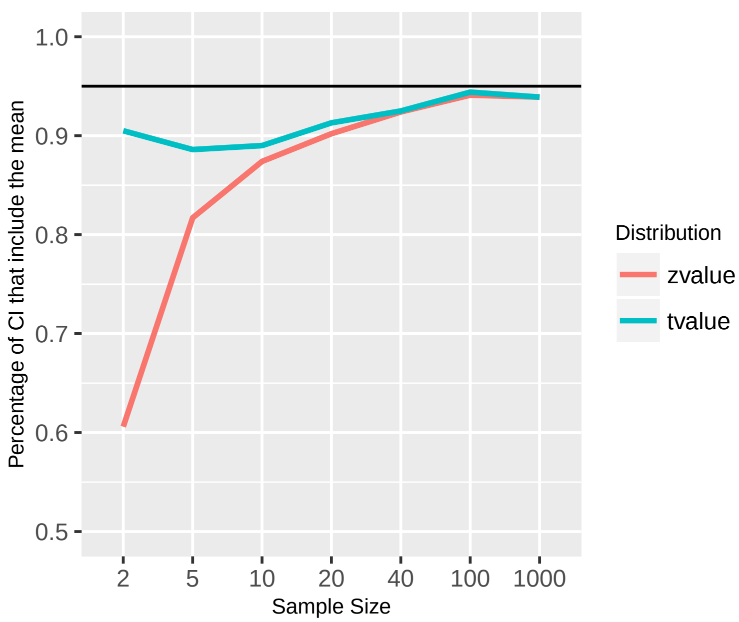
$$mean_{Est} \pm TQ_{1-\alpha/2,n-1} \times SE_{Est} = mean_{Est} \pm TQ_{1-\alpha/2,n-1} \times \frac{SD_{Est}}{\sqrt{n}}$$

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

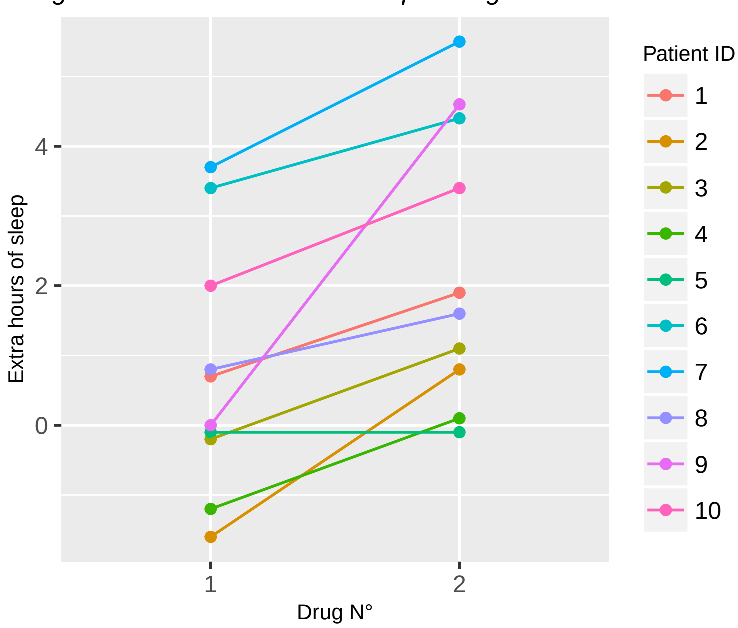
Definition

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:



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
```

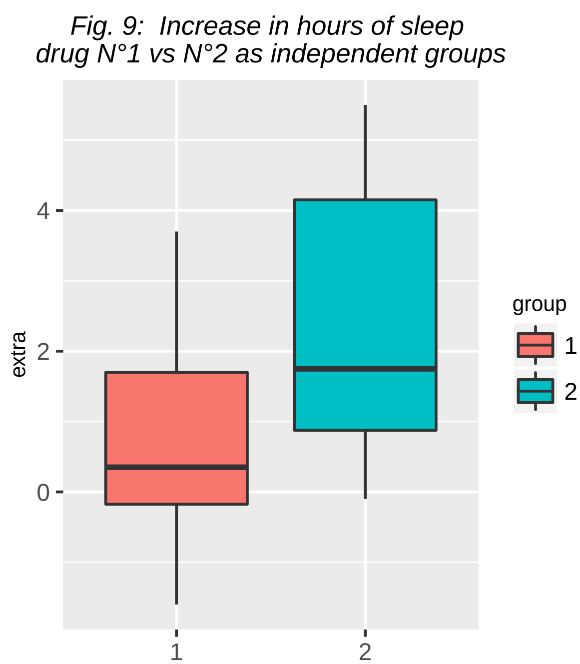
The T Confidence Interval does not include 0 and P $<$ 0.005, so **drug N°2 is statistically more efficient than drug N°1**.

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.



Hypothesis Testing

Principle

Hypothesis testing is the use of statistics to determine the **probability that a given hypothesis is true**.

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
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

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.

It means that α is the **Type I error rate** = 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

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**.

Example

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

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

Power increases as:

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

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

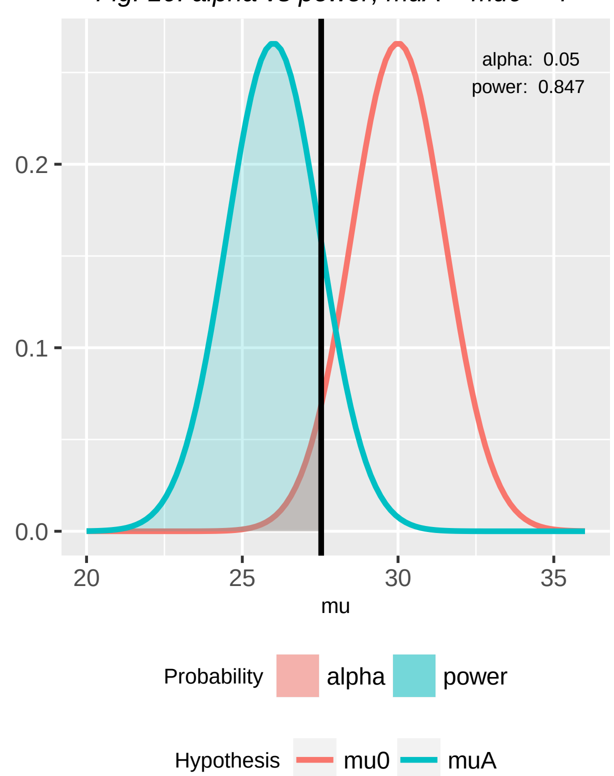


Fig. 11: alpha vs power; mu1 = mu0 + 2