Introduction to statistics: Calibrating errors and the t-test continued

Shravan Vasishth

Universität Potsdam
vasishth@uni-potsdam.de
http://www.ling.uni-potsdam.de/∼vasishth

June 16, 2019

Type I, Type II error, power

When we do a hypothesis test, the sample mean

1. will either fall in the rejection region → reject null
2. or it will not → fail to reject null

But the null hypothesis is either true or not true. *We don't know which of those two is the reality*.

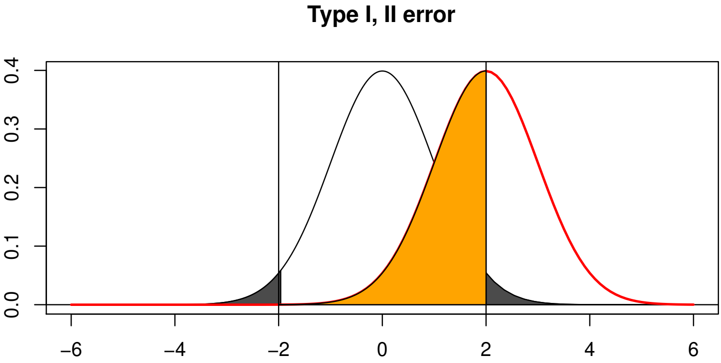
Type I, Type II error, power

Reality:	H_0 TRUE	H_0 FALSE
Decision: ‘reject’:	α **Type I error**	$1 - \beta$ **Power**
Decision: ‘fail to reject’:	$1 - \alpha$	β **Type II error**

For simplicity, assume that SE=1. This means that the t-score is really the sample mean.
Consider the situation where, in reality, the true $\mu = 2$ and the null hypothesis H_0 is that $\mu = 0$. Now the H_0 is false.

Type I, Type II error, Power

Type I error is conventionally held at 0.05. Power is 1-Type II error.



The typical statistical test

t-test

Given data:

```
## Sampling from Normal(0,1)
(sample<-rnorm(10))

##  [1]  0.607331 -0.615250  0.935344 -0.747748  0.507814 -
##  [8]  0.696097 -2.534253 -0.012147
```

The typical statistical test

t-test

If we do a t-test to test the null hypothesis that $\mu = 0$:

```
n<-length(sample)
x_bar<-mean(sample)
stddev<-sd(sample)
(t_value<- (x_bar - 0)/(stddev/sqrt(n)))

## [1] -0.97345
```

The typical statistical test

t-test

We can also compute the probability of getting the observed t-value that we got from the sample, or something more extreme, given the null hypothesis.
This can be computed, as done earlier, simply by calculating the area under the curve in the rejection region for the relevant t-distribution:

```
pt(-abs(t_value),df=n-1)

## [1] 0.17788
```

The typical statistical test

t-test

I just took the absolute t-value from the sample and took its negation in order to compute the probability on the left tail. I could have also written:

```
pt(abs(t_value),df=n-1,lower.tail=FALSE)

## [1] 0.17788
```

The typical statistical test

t-test

The convention is to compute the probability of getting the observed t-value that we got or something more extreme on either side of the t-distribution — we look at both sides of the t-distribution.

```
2*pt(-abs(t_value),df=n-1)

## [1] 0.35576
```

Conventionally, we reject the null if $p < 0.05$. This is because we set the Type I error at 0.05.

The typical statistical test

```
2*pt(-abs(t_value),df=n-1)

## [1] 0.35576
```

Conventionally, we reject the null if this probability < 0.05.
This probability is called the p-value.

The typical statistical test

t-test

You can use the built-in function in R to do such a t-test:

```
t.test(sample)

##
##  One Sample t-test
##
## data:  sample
## t = -0.973, df = 9, p-value = 0.36
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.14882  0.45756
## sample estimates:
## mean of x
##  -0.34563
```

Type I error vs p-values

Type I error is the False Discovery Rate: the probability of your incorrectly rejecting the null under repeated sampling:

```
nsim<-10000
n<-10
pvals<-rep(NA,nsim)
for(i in 1:nsim){
  x<-rnorm(n)
  pvals[i]<-t.test(x)$p.value
}
mean(pvals<0.05)

## [1] 0.0487
```

The p-value you get from one experiment will vary from experiment to experiment. Type I error is a value we fix in advance.

Type II error (1-power) vs p-values

1. Many studies in linguistics and psychology have very low power (sometimes as low as 0.06). (1-power) is Type II error.
2. This implies that if power is low and we get a so-called null result, i.e., fail to reject the null hypothesis (when $p > 0.05$), we can't really conclude anything.
3. If power were high, then a null result could be more meaningful and we might be justified in accepting the null.

But the situation with low power is not just that null results are inconclusive. Even "statistically significant" results are suspect with low power.

Type S- and M-error

Gelman and Carlin, Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors, Perspectives on Psychological Science November 2014 vol. 9 no. 6 641-651

If your true effect size is believed to be $D = 15$, then we can compute (apart from statistical power) these error rates, which are defined as follows:

1. **Type S error**: the probability that the sign of the effect is incorrect, given that the result is statistically significant.
2. **Type M error**: the expectation of the ratio of the absolute magnitude of the effect to the hypothesized true effect size, given that result is significant. Gelman and Carlin also call this the exaggeration ratio, which is perhaps more descriptive than "Type M error".

Type S- and M-error

Suppose a particular study has standard error 46, and sample size 37. And suppose that our true $\mu = 15$. Then, we can compute Type S error as follows:

```
## probable effect size, derived from past studies:
D<-15
## SE from the study of interest:
se<-46
stddev<-se*sqrt(37)
nsim<-10000
drep<-rep(NA,nsim)
for(i in 1:nsim){
samp<-rnorm(37,mean=D,sd=stddev)
drep[i]<-mean(samp)
}
```

Type S- and M-error

```
##power: the proportion of cases where
## we reject the null hypothesis correctly:
(pow<-mean(ifelse(abs(drep/se)>2,1,0)))

## [1] 0.0579
```

Type S- and M-error

```
## which cells in drep are significant at alpha=0.05?
signif<-which(abs(drep/se)>2)

## Type S error rate | signif:
(types_sig<-mean(drep[signif]<0))

## [1] 0.20207
```

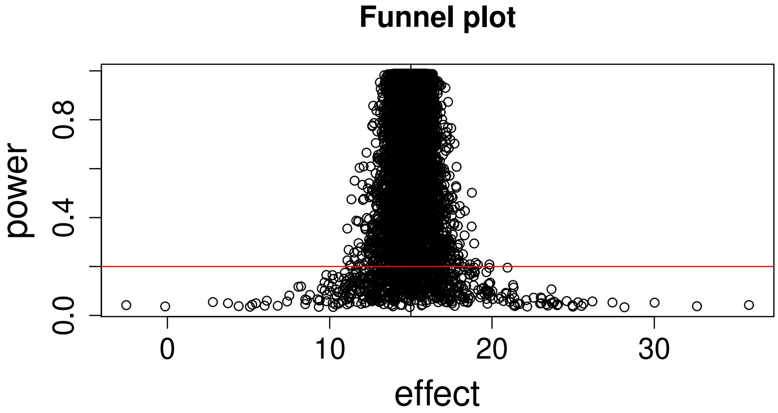
Type S- and M-error

```
## Type M error rate | signif:
(typem_sig<-mean(abs(drep[signif])/D))

## [1] 7.46
```

Type S- and M-error

Funnel plot



Type S- and M-errors

1. So, you can see that the Type S error and the exaggeration ratio, conditional on a result being significant, are pretty high.
2. The practical implication of this is that if most studies are low powered, then **it doesn't matter much whether you got a significant result or not**.
3. The main point here is: run **high powered** studies, and **replicate** the results. There's really nothing that can match consistent replication.

Stopping rules

Researchers often adopt the following type of data-gathering procedure:

1. The experimenter gathers n data points, then checks for significance ($p < 0.05$ or not).
2. If it's not significant, he gets more data (n more data points). Since time and money are limited, he might decide to stop anyway at sample size, say, some multiple of n.

Stopping rules

1. One can play with different scenarios here. A typical n might be 15.
2. This approach would give us a range of p-values under repeated sampling.
3. Theoretically, under the standard assumptions of frequentist methods, we expect a Type I error to be 0.05. This is the case in standard analyses (I also track the t-statistic, in order to compare it with my stopping rule code below).

Stopping rules

```
##Standard:
pvals<-NULL
tstat_standard<-NULL
n<-10
nsim<-10000
## assume a standard dev of 1:
stddev<-1
mn<-0
for(i in 1:nsim){
  samp<-rnorm(n,mean=mn,sd=stddev)
  pvals[i]<-t.test(samp)$p.value
  tstat_standard[i]<-t.test(samp)$statistic
}
```

Stopping rules

```
## Type I error rate: about 5% as theory says:
table(pvals<0.05)[2]/nsim

##   TRUE
## 0.0501
```

Stopping rules

But the situation quickly deteriorates as soon as adopt the strategy I outlined above. I will also track the distribution of the t-statistic below.

```
pvals<-NULL
tstat<-NULL
## how many subjects can I run?
upper_bound<-n*6
```

Stopping rules

```
for(i in 1:nsim){
  significant<-FALSE
  x<-rnorm(n,mean=mn,sd=stddev) ## take sample
while(!significant & length(x)<upper_bound){
  ## if not significant:
if(t.test(x)$p.value>0.05){
  x<-append(x,rnorm(n,mean=mn,sd=stddev)) ## get more data
} else {significant<-TRUE}   ## otherwise stop:
}
pvals[i]<-t.test(x)$p.value
tstat[i]<-t.test(x)$statistic
}
```

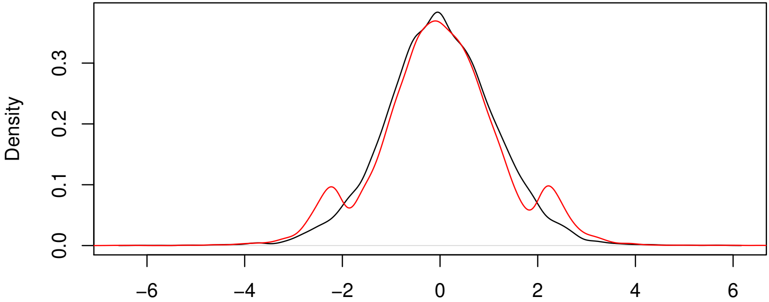
Stopping rules

```
## Type I error rate: much higher than 5%:
table(pvals<0.05)[2]/nsim

##  TRUE
## 0.153
```

Stopping rules

Now let's compare the distribution of the t-statistic in the standard case vs with the above stopping rule (red):



Stopping rules

1. We get fatter tails with the above stopping rule—a higher Type I error than 0.05.
2. The point is that one should fix one's sample size in advance based on a power analysis, not deploy a stopping rule like the one above; if we used such a stopping rule, we are much more likely to incorrectly declare a result as statistically significant.
3. Of course, if your goal is to get an article published no matter what, such stopping rules are a great way to have a successful career!

Summary

1. We learnt about the single sample t-test.
2. We learnt about Type I, II error (and power).
3. We learnt about Type M and Type S errors.