Displays for Statistics 5303

Lecture 14

October 4, 2002

Christopher Bingham, Instructor

612-625-7023 (St. Paul)
612-625-1024 (Minneapolis)

Class Web Page
http://www.stat.umn.edu/~kb/classes/5303

Sample size and Power

An important part of experimental design is deciding how big an experiment should be, that is, what **sample size** or sizes you should use.

Sometimes you very little choice
- Because of **limited time**
- Because of **limited funds** - perhaps you can afford only n = 4 replicates.

Even in this case, sample size and power calculations can still be useful. When you find out the smallest sample size that will meet your goals exceeds your resources, your best action may be to
- Try to get a larger grant (more $$)
- Experiment more by sleeping less
- Put more thought into how you can reduce variability and be able to reach your goals with a smaller sample.
- Change your goals: accept smaller power or a wider confidence interval
- Bail out and do something else

There are two basic sample size problems, one related to **accuracy of estimation** and one related to the **power of a significance test**.

Accuracy of estimation

You plan to estimate a parameter θ such as $\alpha_1 - \alpha_2$ with a confidence interval of the usual form

$$\hat{\theta} \pm t_{\alpha/2}\hat{SE}[\hat{\theta}] = (\hat{\theta} - t_{\alpha/2}\hat{SE}[\hat{\theta}], \hat{\theta} + t_{\alpha/2}\hat{SE}[\hat{\theta}])$$

You want the width of the interval to be no more than W, a number chosen by you. That is you want the smallest sample size n such that

$$\text{interval width} = 2\times t_{\alpha/2}\hat{SE}[\hat{\theta}] \leq W$$

Since standard errors decrease as n increases, you try to find n such that

$$2\times t_{\alpha/2}\hat{SE}[\hat{\theta}] \cong W$$

In terms of the margin of error $M = W/2$, this is

$$t_{\alpha/2}\hat{SE}[\hat{\theta}] \cong M$$

The margin of error is the ± part: You want the C.I. to be $\hat{\theta} \pm M$, so that, say, you have 95% confidence that the distance between $\hat{\theta}$ and θ is no more than M.

Often $\hat{SE}[\hat{\theta}] = C\sqrt{MS_E}/\sqrt{n}$ for some constant C such as $\sqrt{\sum w_i^2}$, so the equation is

$$t_{\alpha/2}\times\hat{SE} = t_{\alpha/2}\times C\sqrt{MS_E}/\sqrt{n} = M$$

This means n is given by the equation

$$n \cong t_{\alpha/2}^2\times C^2\times MS_E/M^2$$

There are two problems

1 You haven't done the experiment yet so *you don't know* MS_E; you somehow have to come up with a value for MS_E

2 $t_{\alpha/2}$ is really $t_{\alpha/2,df_{error}} = t_{\alpha/2,g(n-1)}$ which depends on n which you don't yet know, so you may need trial and error to get the result.

Problem 6.1 data

Six treatments determined by 4 levels of N and irrigation level (Y and N)

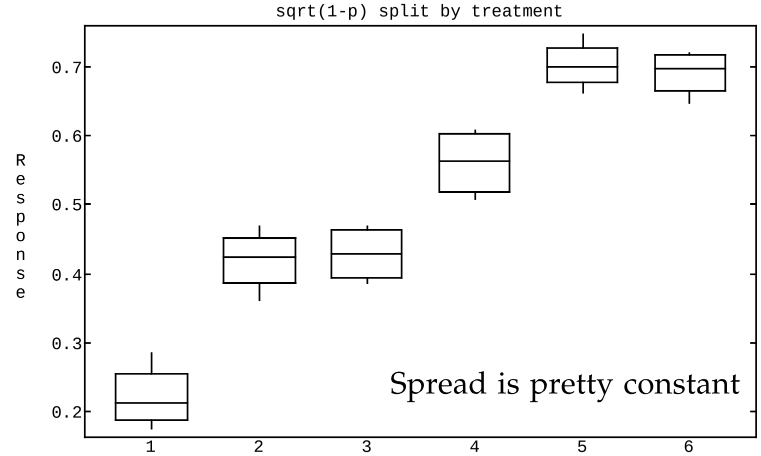
Treatment No	1	2	3	4	5	6
Nitrogen	1	1	2	3	4	4
Irrigation	N	Y	N	N	Y	N

```
Cmd> readdata("",treat,percentgood)
Read from file "TP1:Stat5303:Data:Ch06:pr6-1.dat"
Column 1 saved as REAL vector treat
Column 2 saved as REAL vector percentgood

Cmd> treat <- factor(treat)

Cmd> y <- sqrt(1 - percentgood/100)

Cmd> vboxplot(split(y,treat),xlab:"Treatment number",\
    ylab:"Response"title:"sqrt(1-p) split by treatment")
```




```
Cmd> tabs(y,treat,count:T) # sample sizes
(1)           4           4           4           4           4
(6)           4

Cmd> anova("y=treat",fstat:T)
Model used is y=treat
              DF          SS          MS           F      P-value
CONSTANT       1      6.1003      6.1003  3401.99685            0
treat          5     0.68745     0.13749    76.67585  1.7012e-11
ERROR1        18    0.032277   0.0017931
```

$Df_{error} = g(n-1) = 6\times 3 = 18.$

```
Cmd> w_N <- vector(-1,-1,0,0,1,1) # contrast weights

Cmd> result <- contrast(treat, w_N); result
component: estimate
(1)      0.75576
component: ss
(1)      0.57117
component: se
(1)     0.042346

Cmd> n <- 4; g <- 6

Cmd> t_025 <- invstu(1 - .05/2, g*(n-1)); t_025 #df=(4-1)*6=18
(1)       2.1009

Cmd> error_margin <- t_025*result$se; error_margin
(1)     0.088965   Margin of error for this C.I.
```

What sample size would you need for M = .025 with this MS_E = 0.0017931.

A contrast SE = $\sqrt{\{\sum w_i^2/n\}}\times\sqrt{MS_E}$

```
Cmd> ssw <- sum(w_N^2); ssw

Cmd> mse <- SS[3]/DF[3]; mse # same as in ANOVA table
      ERROR1
   0.0017931

Cmd> sqrt(mse*ssw/n) # standard error of contrast
(1)     0.042346   Same as computed by contrast()
```

```
Cmd> M <- .025 # target Margin of error = W/2

Cmd> t_025 <- 2 # starting value for ctricial value

Cmd> n <- t_025^2*ssw*mse/M^2; n # first try
(1)      45.904

Cmd> n <- 46  # round 45.9 up to 46

Cmd> t_025 <- invstu(1 - .025,g*(n-1)) # new critical value

Cmd> n <- t_025^2*ssw*mse/M^2; n 3 second try
(1)      44.483

Cmd> n <- 45  # round 44.5 up to 45

Cmd> n <- t_025^2*ssw*mse/M^2; n # 3rd try
(1)      44.483      Still rounds up to 45; stop
```

You can do the search more automatically. This code applies the "secant method" of solving an equation. Its an "educated" trial and error method.

```
Cmd> N <- vector(10,20) # two different trial values

Cmd> ME <- invstu(1-.025,g*(N-1))*sqrt(mse*ssw/N)#2 marg of err

Cmd> ME # margins of error for n = 10 and 20
(1)    0.053694    0.037515

Cmd> for(i,1,7){ # do 7 steps
        b <- (ME[2] - ME[1])/(N[2]-N[1]) # secant slope
        ME[1] <- ME[2]; N[1] <- N[2]
        N[2] <- N[2] + (M - ME[2])/b # update N[2]
        ME[2] <- invstu(1-.025,g*(N[2]-1))*sqrt(mse*ssw/N[2])
        vector(N[2],ME[2]) # new n and Margin of Error
     }
(1)      27.735    0.031758
(1)      36.817    0.027512
(1)      42.188    0.025682
(1)      44.189    0.025088
(1)      44.485    0.025003
(1)      44.497       0.025
(1)      44.497       0.025    n rounds up to 45
           n  Margin of error
```


This method coverges faster if you update $1/\sqrt{n}$ instead of n:

```
Cmd> invsqrt_n <- 1/sqrt(vector(10,20))# two trial 1/sqrt(n)

Cmd> N <- 1/invsqrt_n^2  # sample sizes

Cmd> ME <- invstu(1-.025,g*(N-1))*sqrt(mse*ssw/N)#error_margins

Cmd> for(i,1,5){ # do 5 steps
        b <- (ME[2] - ME[1])/(invsqrt_n[2]-invsqrt_n[1]) #slope
        ME[1] <- ME[2]; invsqrt_n[1] <- invsqrt_n[2]
        invsqrt_n[2] <- invsqrt_n[2] + (M - ME[2])/b #update
        N <- 1/invsqrt_n^2   # sample sizes
        ME[2] <- invstu(1-.025,g*(N[2]-1))*sqrt(mse*ssw/N[2])
        vector(N[2],ME[2]) # new n and Margin of Error
     }
(1)      43.305    0.025345
(1)      44.486    0.025003
(1)      44.497       0.025
(1)      44.497       0.025
(1)      44.497       0.025
```

It converged in only 3 steps, and even the first step was closer.

I'm not sure how important all this accuracy is.

In most cases, the value for MS_E you use is judy an educated guess and could be off by a factor of 2 or more. If you just use $z_{\alpha/2}$ you're usually aren't far off.

Find sample size for power goal

The objective is to achieve a specified power P for a significance test with **given type I error probability α.**

That is, given a desired power P find n such that

- a **significance test** with specific **significance level** α (type I error probability) has power P
- Power is computed *as if* a *particular alternative* to H_0 were true.

$$P = \text{Power} = P(\text{reject } H_0 \mid H_0 \text{ false})$$
$$= P(\text{no type II error}) = 1 - \beta$$

where β = P(type II error).

- High power means small type II error rate and vice versa.
- *Power depends on the particular alternative*. You may get a differrent value for different alternatives.

The power of a F or t test depends on

- the sample size (power increases with n)
- σ^2 (power increase as σ^2 decreases)
- how far away H_a is away from the H_0

Generally the distance that matters is relative to the value of σ. This means **you need a value both for σ^2** and for the distance.

Treatment effects $\alpha_1, \alpha_2, \ldots, \alpha_g$ are used in several formulas. These are always defined as

$$\alpha_i = \mu_i - \mu^*,$$

where

$$\mu^* = \sum_i n_i \mu_i / \sum n_i$$

They satisfy $\sum_i n_i \alpha_i = 0$

When the n_i's are equal, $\mu^* = \sum_i \mu_i / g$ and $\sum_i \alpha_i = 0$

Mathematics can show that, when at least one $\alpha_i \neq 0$, the F-statistic has the so called **non-central F** distribution.

The non-central F distribution depends on three quantities. Two are the same as for ordinary (Central) F:

- the **numerator d.f.** = g-1
- the **denominator d.f.** = df_{error} = g(n-1)
- the **non-centrality parameter**
  $\zeta = \sum_i n_i \alpha_i^2/\sigma^2$ (zeta)

When the n_i's are all equal to n,

$\zeta = n\sum_i \alpha_i^2/\sigma^2$.

Central F corresponds to $\zeta = 0$.

Since you reject H_0 for $F > F_{\alpha,\, df_{numerator},\, df_{error}}$,

$P = \text{power} = P(F_{non\text{-}central} > F_{\alpha,\, df_{numerator},\, df_{error}})$

δ and λ are sometimes used instead of ζ for the non-centrality parameter.

With $n_1 = \ldots = n_g = n$, the quantity

$$\zeta_1 = \sum_i \alpha_i^2/\sigma^2$$

measures the (squared) distance *relative* to σ^2 of the specific H_a from

$$H_0: \alpha_1 = \ldots = \alpha_g = 0.$$

We refer to ζ_1 as the

n = 1 non-centrality parameter.

- For fixed treatment effects $\{\alpha_i\}$, with at least one $\alpha_i \neq 0$, and fixed σ^2,
  ζ increases as n increases.
- For fixed n and σ^2,
  ζ_1 increases and so does ζ as the distance from H_0 to H_a increases, that is, as any or all of the treatment effects α_i increase
- For fixed n and $\{\alpha_i\}$, α_i not all zero, ζ_1 and ζ increase as σ^2 decreases

Since power is the probability of obtaining a large F-statistic when H_0 is false, you use the non-central F distribution to calculate power.

Example: $\alpha = .01$, $g = 6$, $n = 4$ and $\zeta_1 = .5$.

```
Cmd> g <- 6; n <- 4; df_error <- g*(n-1); df_error
(1)         18

Cmd> alpha <- .01

Cmd> F_alpha <- invF(1 - alpha, g-1, g*(n-1)); F_alpha
(1)      4.2479      Rejection cut-point for F-test

Cmd> zeta1 <- .5     n=1 non-centrality parameter
```

cumF() with 4 arguments computes non-central F:

```
Cmd> 1 - cumF(F_alpha,g-1,g*(n-1),n*zeta1)
(1)    0.034159
```

power() is a short cut for CRD.

```
Cmd> power(zeta1,g,alpha,n)
(1)    0.034159
```

In the equal n_i case, non-central F depends on $\zeta_1 = \sum\alpha_i^2/\sigma^2$ and you need to somehow come up with values for $\sum\alpha_i^2$ and σ^2 before you can find a sample size.