Hypothesis testing

- Non-normal distributions
- Effect size
- Power of test
- Sample size
- Philosophy of hypothesis testing

Chi-square test

Chi-square test

- A fundamental problem in genetics is determining whether the experimentally determined data fits the results expected from theory

- How can you tell if an observed set of offspring counts is legitimately the result of a given underlying simple ratio?

$$Aa \times Aa$$

AA	1
Aa	2
aa	1

- For example, you do a cross and see 290 purple flowers and 110 white flowers in the offspring. This is pretty close to a 3/4 : 1/4 ratio, but how do you formally define "pretty close"? What about 250:150?

Goodness of fit

- Mendel has no way to solve this problem. Shortly after the rediscovery of his work in 1900, Karl Pearson and R. A. Fisher developed the "chi-square" test for this purpose.

- The chi-square test is a "goodness of fit" test: it answers the question of how well do observed data fit expectations.

- Theory for how the offspring will be distributed (null hypothesis): for a self-pollination heterozygote, the offspring will appear in a ratio of 3/4 dominant to 1/4 recessive.

Formula

- First determine the number of each phenotype that have been observed and how many would be expected given basic genetic theory.

- Then calculate the chi-square statistic using the formula.

$$X^2 = \sum \frac{(obs - \exp)^2}{\exp}$$

	exp	obs
purple	300	290
white	100	110

- The "X" is the Greek letter chi; the "∑" is a sigma, it means to sum the following terms for all phenotypes. "obs" is the number of individuals of the given phenotype observed; "exp" is the number of that phenotype expected from the null hypothesis.

- Note that you must use the **number** of individuals, the counts, and NOT proportions, ratios, or frequencies.

Example

As an example, you count offspring, and get 290 purple and 110 white flowers. This is a total of 400 (290 + 110) offspring.

We expect a 3/4 : 1/4 ratio, i.e.

we expect 400 x 3/4 = 300 purple, and 400 x 1/4 = 100 white.

Purple, obs = 290 and exp = 300.

White, obs = 110 and exp = 100.

$$X^2 = \sum \frac{(obs - \exp)^2}{\exp}$$

Then plugging into the formula:

$X^2 = (290 - 300)^2 / 300 + (110 - 100)^2 / 100$

= 0.333 + 1.000

= 1.333.

1- pchisq(1.333, df = 1) # 0.248

6

Chi-square distribution

- Although the chi-square distribution can be derived through math theory, we can also get it experimentally

- Do the same self-pollination of a Aa heterozygote 1000 times, which should give the 3/4 : 1/4 ratio. For each experiment we calculate the chi-square value, them plot them all on a graph

- The x-axis is the chi-square value calculated from the formula. The y-axis is the number of individual experiments that got that chi-square value.
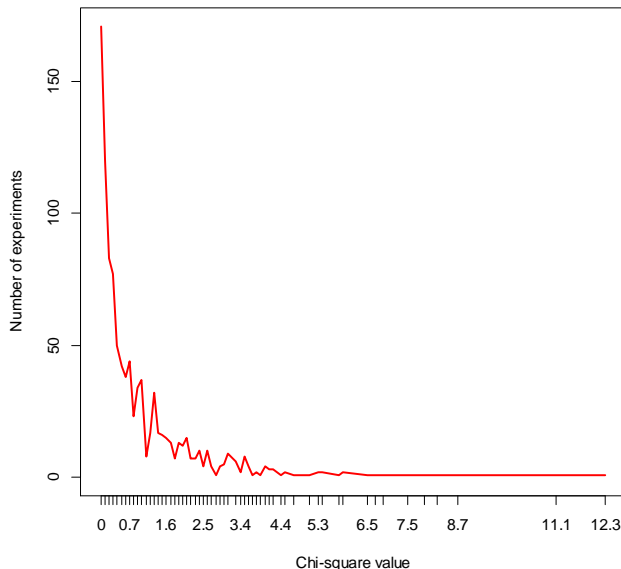
```
plot(table(round(rchisq(1000, df = 1),1)),
     type = 'l', col = 'red',
     xlab = 'Chi-square value',
     ylab = 'Number of experiments')
```

Chi-square distribution

- Most experiments give a small chi-square value (the hump in the graph).

- Note that all the values are greater than 0: that's because we squared the (obs - exp) term

- Sometimes you get really wild results, with obs very different from exp: the long tail on the graph. Really odd things occasionally do happen by chance alone (for instance, you might win the lottery).

$$X^2 = \sum \frac{(obs - \exp)^2}{\exp}$$
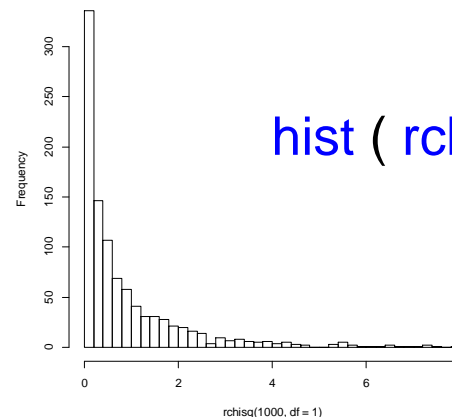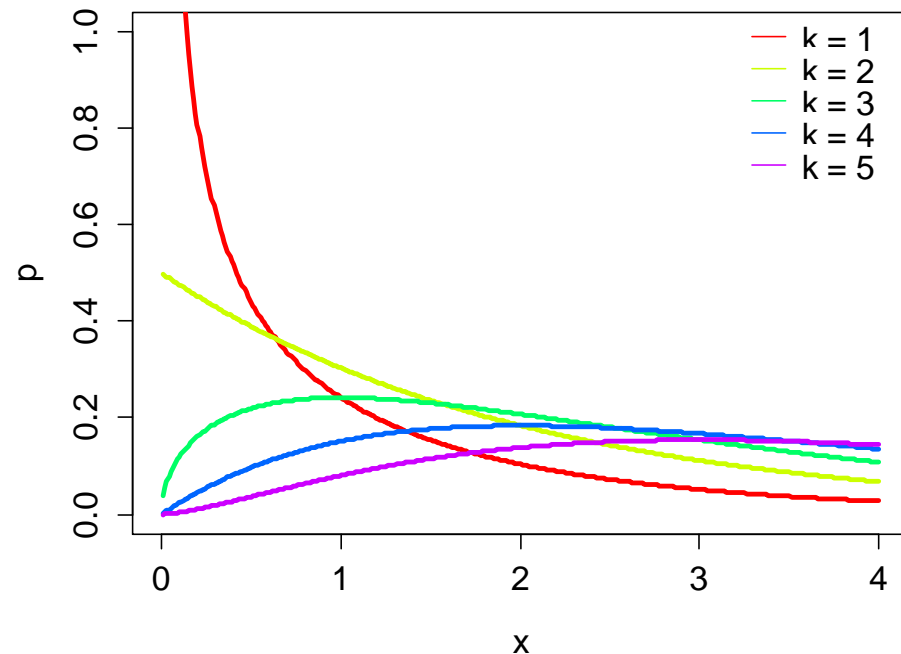
hist (rchisq (1000, df =1), nclass=30)

8

PDF of Chi-square distribution

$$P = f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \, x^{\frac{k}{2}-1} \, e^{-\frac{x}{2}}$$

k: degrees of freedom
$\Gamma(n)$ = (n-1)!   (Gamma(n))


```
x <- seq(0.01,4, by=0.01)

plot(x, dchisq(x,1), type="l", xlab='x',
    ylab="p", xlim=c(0,4), ylim=c(0,1))

for(i in 1:5){
    lines(x, dchisq(x,i), col=rainbow(5)[i], lwd=3)
    legend(3.2, .1.1-i/15, paste('k =', i, sep=' '),
        lty = 1, col = rainbow(5)[i],
        box.lty=0, cex=1)
}
```

9

The critical question

- How do you judge an odd result (e.g., 290:110) is right or wrong?

  Most of the time the results fit expectations pretty well, but occasionally very skewed distributions of data occur even though you performed the experiment correctly, based on the correct theory.

- The simple answer is: you can never tell for certain that a given result is "wrong", that the result you got was completely impossible based on the theory you used. All we can do is to determine whether a given result is likely or unlikely.

The critical question

- Key point: there are two ways of getting a high chi-square value: an unusual result from the correct theory, or a result from another theory.

- Using the example here, how can you tell if your 290: 110 offspring ratio really fits a 3/4 : 1/4 ratio (as expected from selfing a heterozygote) or whether it was the result of a 1/2 : 1/2 ratio?

- You can't be certain, but you can at least determine whether your result is reasonable.

Reasonable result

- What is a "reasonable" result is subjective.

- For most work, a result is said to not differ significantly from expectations if it could happen at least 1 time in 20.

- We use "fail to reject" instead of "accept".

- "1 time in 20" can be written as a probability value p = 0.05

Degrees of freedom

- A critical factor in using the chi-square test is the "degrees of freedom", which is essentially the number of independent random variables involved.

- Degrees of freedom is simply the number of classes of offspring minus 1.

- For our example, there are 2 classes of offspring: purple and white. Thus, degrees of freedom (d.f.) = 2 -1 = 1.

Critical Chi-square

- Critical values for chi-square are found on tables, sorted by degrees of freedom and probability levels. Or you can have it from statistical software. Be sure to use p = 0.05.

<p align="center">**qchisq(0.95, df = 1)**</p>

- If your calculated chi-square value is greater than the critical value from the table, you "reject the null hypothesis".

- If your chi-square value is less than the critical value, you "fail to reject" the null hypothesis.

Chi-square table

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401

Chi square test

- In our example of 290 purple to 110 white, we calculated a chi-square value of 1.333, with 1 degree of freedom.

- Looking at the table, 1 d.f. is the first row, and p = 0.05 is the sixth column. Here we find the critical chi-square value, 3.841.

- Since our calculated chi-square, 1.333, is less than the critical value, 3.841, we "fail to reject" the null hypothesis. Thus, an observed ratio of 290 purple to 110 white is a good fit to a 3/4 to 1/4 ratio.

chisq.test (c (290,110), p = c (0.75,0.25))   P = 0.248

Another example: from Mendel

phenotype	observed	expected proportion	expected number
round yellow	315	9/16	312.75
round green	101	3/16	104.25
wrinkled yellow	108	3/16	104.25
wrinkled green	32	1/16	34.75
total	556	1	556

Finding the expected numbers

- You are given the observed numbers, and you determine the expected proportions from a Punnett square.

- To get the expected numbers of offspring, first add up the observed offspring to get the total number of offspring. In this case, 315 + 101 + 108 + 32 = 556.

Punnett Square of Dihybrid Cross
Gametes from RrYy parent

- Then multiply total offspring by the expected proportion:
  -expected round yellow   = 9/16 * 556 = 312.75
  -expected round green   = 3/16 * 556 = 104.25
  -expected wrinkled yellow = 3/16 * 556 = 104.25
  -expected wrinkled green   = 1/16 * 556 = 34.75

- Note that these add up to 556, the observed total offspring.

Calculating the Chi-square value

$$X^2 = \sum \frac{(obs - \exp)^2}{\exp}$$

$X^2 = (315 - 312.75)^2 / 312.75$

$\qquad + (101 - 104.25)^2 / 104.25$

$\qquad + (108 - 104.25)^2 / 104.25$

$\qquad + (32 - 34.75)^2 / 34.75$

$\qquad = 0.016 + 0.101 + 0.135 + 0.218$

$\qquad = 0.470.$

chisq.test (c(315,101,108,32), p = c(9/16,3/16,3/16,1/16))

D.F. and critical value

- Degrees of freedom is 1 less than the number of classes of offspring. Here, 4 - 1 = 3.

- For 3 d.f. and p = 0.05, the critical chi-square value is 7.815.

- Since the observed chi-square (0.470) is less than the critical value, we fail to reject the null hypothesis. We can not reject Mendel's conclusion that the observed results for a 9/16 : 3/16 : 3/16 : 1/16 ratio so far.

- It should be mentioned that all of Mendel's numbers are unreasonably accurate.
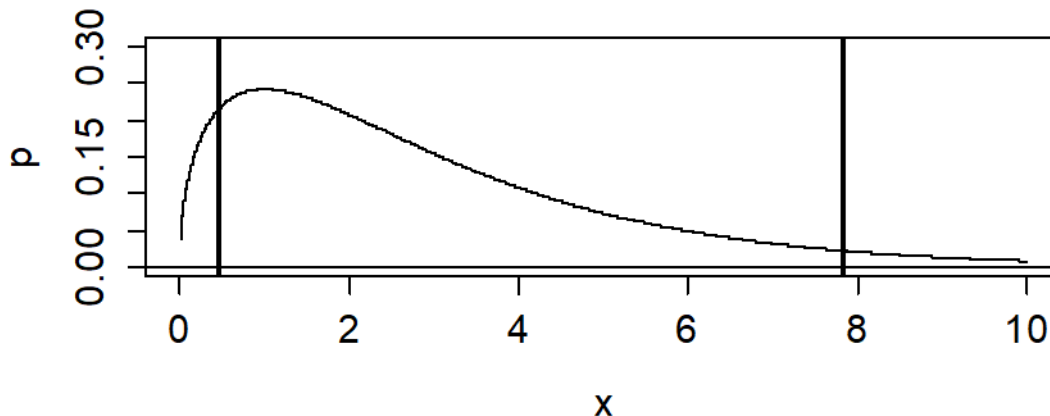
```
x <- seq(0.01,10, by=0.01)
plot(x, dchisq(x,3), type="l", xlab='x',
        ylab="p", xlim=c(0,10), ylim=c(0,.3))
abline(0,0)
abline(v=0.47, lwd=3) # Mendel's result
abline(v=qchisq(.95, 3), lwd=3) # 5%
```


20

Chi square assumptions

The chi square test can only be used on data that has the following characteristics:

- The data must be in the form of frequencies (count)

- The frequency data must have a precise numerical value and must be organised into categories or groups.

- The expected frequency in any one cell of the table must be greater than 5.

- The total number of observations must be greater than 20.

Fisher's exact test

- A statistical significance test used in the analysis of contingency tables

- Be appropriate for small sample sizes (valid for large sample sizes as well)

- One of a class of exact tests, because the <u>significance of the deviation from a null hypothesis</u> (e.g., P-value) can be calculated exactly, rather than relying on an approximation that becomes exact when the sample size grows to infinity

Fisher's exact test

	Men	Women	Row Total
Yoga	2	8	10
Non-yoga	10	4	14
Column Total	12	12	24

- There are10 of these 24 students are doing Yoga regularly, and that 12 of the 24 are female.

- **Assuming the null hypothesis that men and women are equally likely Yoga**, what is the probability that these 10 Yoga lovers would be so unevenly distributed between the women and the men?

  ✓ If we choose 10 Yoga lovers at random, what is the probability that 8 or more of them would be among the 12 women, and only 2 or fewer from among the 12 men?

Fisher's exact test

	Men	Women	*Row Total*
Yoga	*a*	*b*	*a + b*
Non-yoga	*c*	*d*	*c + d*
Column Total	*a + c*	*b + d*	*a + b + c + d* *(=n)*

Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\ b!\ c!\ d!\ n!}$$

where () is the binomial coefficient and the symbol ! indicates the factorial operator.
The formula above gives the exact hypergeometric probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that men and women are equally likely to do Yoga.

R for Fisher's exact test

```
TeaTasting  <-  matrix(c(8, 2, 3, 7),  nrow = 2,
                 dimnames = list(Guess = c("Milk", "Tea"),
                                 Truth = c("Milk", "Tea")))
fisher.test(TeaTasting, alternative = "greater")
```

	Truth	
Guess	Milk	Tea
Milk	8	3
Tea	2	7

```
Fisher's Exact Test for Count Data
data: TeaTasting
p-value = 0.03489
alternative hypothesis: true odds ratio
is greater than 1
95 percent confidence interval:
1.155327 Inf
sample estimates: odds ratio 8.153063
```

```
## A 4 x 4 table  Agresti (2002, p. 57) Job Satisfaction
Job <- matrix(c(1,2,1,0, 3,3,6,1, 10,10,14,9, 6,7,12,11), 4, 4,
         dimnames = list(income = c("< 15k", "15-25k", "25-40k", "> 40k"),
                         satisfaction = c("VeryD", "LittleD", "ModerateS", "VeryS")))
fisher.test(Job)
fisher.test(Job, simulate.p.value = TRUE, B = 1e5)
```

25

Test homogeneity of variance

Test homogeneity of variance – one sample

$$H_0 : \sigma^2 = \sigma_0^2$$

Test the null that the population variance has some specific value. Pick an alpha value. Then:

$$\chi^2_{(N-1)} = \frac{(N-1)s^2}{\sigma_0^2}$$

Plug hypothesized population variance and sample variance into equation along with sample size we used to estimate variance. Compare to chi-square distribution.

var.test()
bartlett.test()
fligner.test()

27

Example of chi square test of variance (one tailed test)

Test about variance of height of people in inches. Sample 30 people at random and measure the heights.

$H_0 : \sigma^2 \geq 6.25; H_1 : \sigma^2 < 6.25.$

$N = 30; s^2 = 4.55$

$$\chi^2_{(N-1)} = \frac{(N-1)s^2}{\sigma^2_0}$$

$$\chi^2_{29} = \frac{(29)(4.55)}{6.25} = 21.11$$

$p = 0.855$

1 tailed test on small side. Set alpha=.01.

For p=.01, the value of chi-square is 14.257. Cannot reject null.   qchisq(0.01, df = 29)

14.257
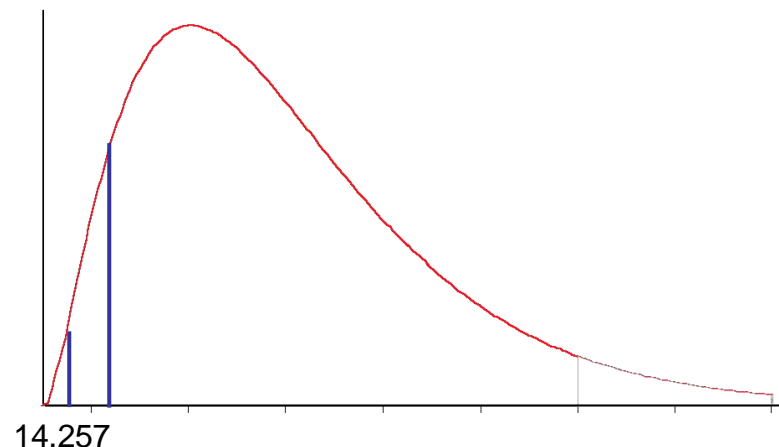
Example of chi square test of variance (two tailed test)

$$H_0 : \sigma^2 = 6.25; H_1 : \sigma^2 \neq 6.25.$$

$$N = 30; s^2 = 4.55$$

$$\chi^2_{(N-1)} = \frac{(N-1)s^2}{\sigma_0^2} \qquad \chi^2_{29} = \frac{(29)(4.55)}{6.25} = 21.11$$

2 tailed with alpha=.01.

Now chi-square with v=29 and p=.005 is 13.121 and also with p=.995 the result is 52.336. N. S. either way.
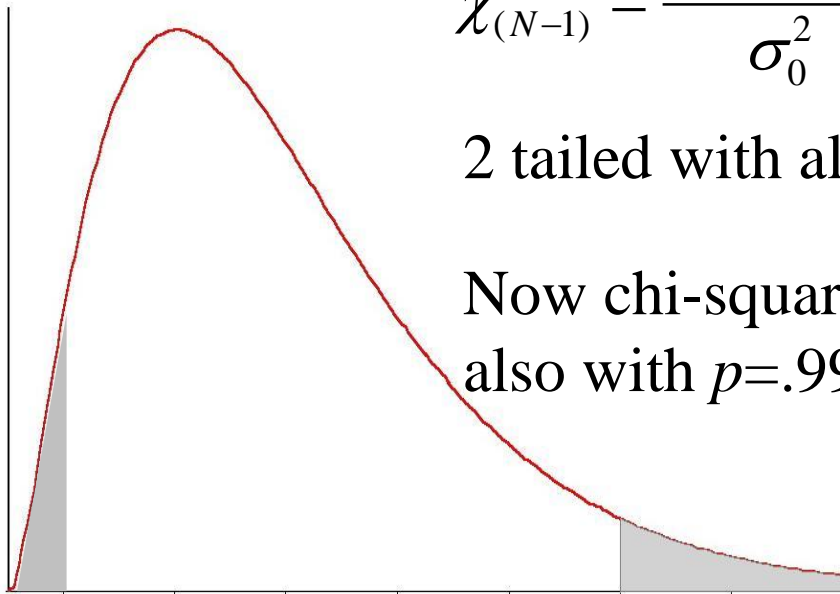
qchisq(0.995, df = 29)
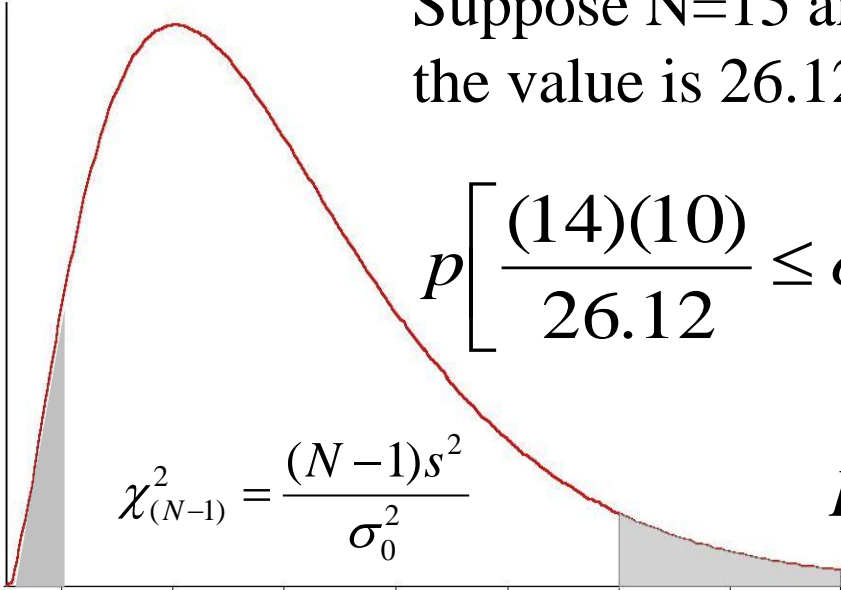
Confidence intervals for the variance

We use s^2 to estimate σ^2. It can be shown that:

$$p\left[\frac{(N-1)s^2}{\chi^2_{(N-1;.025)}} \le \sigma^2 \le \frac{(N-1)s^2}{\chi^2_{(N-1;.975)}}\right] = .95$$

Suppose N=15 and s^2 is 10. Then df=14 and for p=.975 the value is 26.12. For p=.025 the value is 5.63.

$$p\left[\frac{(14)(10)}{26.12} \le \sigma^2 \le \frac{(14)(10)}{5.63}\right] = .95$$

$$\chi^2_{(N-1)} = \frac{(N-1)s^2}{\sigma_0^2}$$

$$p\left[5.36 \le \sigma^2 \le 24.87\right] = .95$$

Test homogeneity of variance – F test for two independent samples

$$F = \frac{s_1^2}{s_2^2}$$

ZAR p138   **Test homogeneity of variance – two independent samples**

EXAMPLE 8.8 The data are the numbers of moths caught during the night by eleven traps of one style and eight traps of a second style.

The two-tailed variance ratio test for the hypotheses H_o: $\sigma_1^2 = \sigma_2^2$ and H_A: $\sigma_1^2 \neq \sigma_2^2$.

Trap type 1	*Trap type 2*
41	52
34	57
33	62
36	55
40	64
25	57
31	56
37	55
34	
30	
38	

$$F = \frac{s_1^2}{s_2^2} = \frac{21.87}{15.36} = 1.42$$

$\alpha = 0.05 \qquad F_{0.05(2),10,7} = 4.76$

Do not reject H_0

$n_1 = 11 \qquad n_2 = 8$

$\nu_1 = 10 \qquad \nu_2 = 7$

$$s_p^2 = \frac{218.73\text{moths}^2 + 107.50\text{moths}^2}{10 + 7} = 19.19\text{moths}^2$$

$SS_1 = 218.73\text{moths}^2 \quad SS_2 = 107.50\text{moths}^2$

$s_1^2 = 21.87\text{moths}^2 \qquad s_2^2 = 15.36 \text{ moths}^2$

Check variance

- bartlett.test(), sensitive to outliers
- var.test(), Fisher's F test, sensitive to outliers
- fligner.test(), performs a Fligner-Killeen (median) test of the null that the variances in each of the groups (samples) are the same. It is a non-parametric test which uses the ranks of the absolute values.
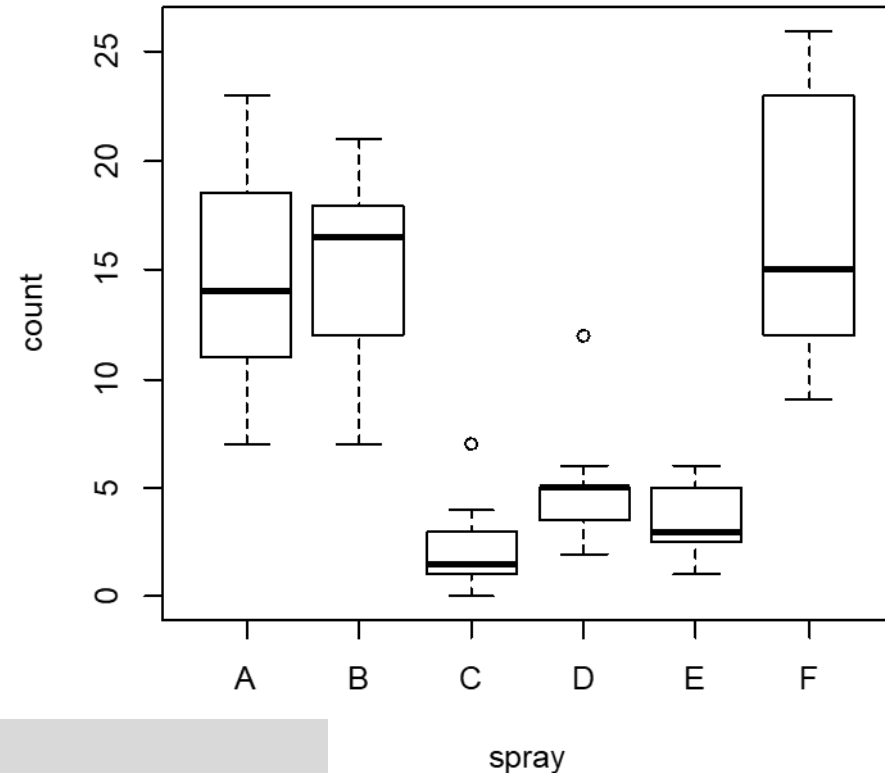

```
plot(count ~ spray, data = InsectSprays)

bartlett.test(InsectSprays$count, InsectSprays$spray)
var.test    (InsectSprays$count, InsectSprays$spray)
fligner.test (InsectSprays$count, InsectSprays$spray)
fligner.test (count ~ spray, data = InsectSprays)
```

33

How much a *p* value can tell?

Effect size (Cohen's d) $d = \dfrac{\bar{x} - \mu}{s_{pooled}}$

- The standardized difference

- Doesn't depend on the size of the sample

- Difference between the mean of your sample and the mean of the population if the null were true, divided by the standard deviation of the population

Effect size $\quad d = \dfrac{\bar{x} - \mu}{s_{pooled}}$
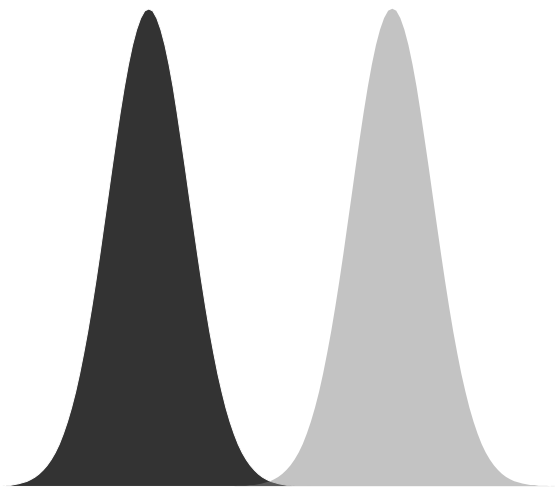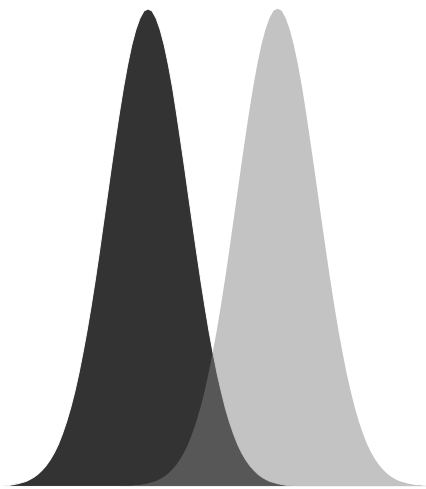
Big effect size　　　　　**Small effect size**　　　**Small effect size**

R script

```
effect.size <- function(data.1, data.2){
    d <- (mean(data.1) - mean(data.2))/
    sqrt(((length(data.1) - 1) * var(data.1) +
    (length(data.2) - 1) * var(data.2))/
    (length(data.1) + length(data.2)-2))
    names(d) <- "effect size d"
    return(d)
}


effect.size(rnorm(30),rnorm(50,2,1))
```

effect size d
   -2.151299

```
# demonstrate difference
plot.design(breaks ~ wool + tension, data = warpbreaks)
```


Factors

37
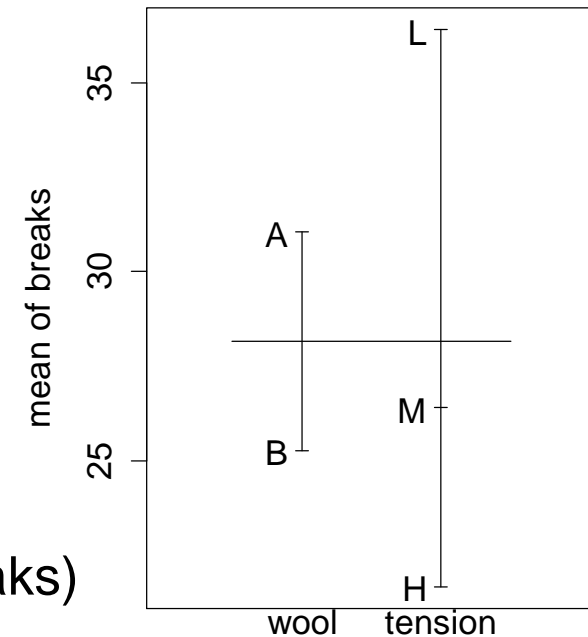
A standard hypothesis Testing (the 5 steps): a whole picture?

No!

Type I and Type II Errors

		NULL HYPOTHESIS	
		TRUE	**FALSE**
D E C I S I O N	**Reject the null hypothesis**	Type **I** error α **Rejecting a true null hypothesis**	**CORRECT** $1 - \beta$
	Fail to reject the null hypothesis	**CORRECT** $1 - \alpha$	Type **II** error β **Failing to reject a false null hypothesis**

Type I and Type II Errors

@游识猷 V 🎖️🎈

统计中的Type I error 和Type II error😄 来自推特aLittleMedic

↑ 收起　｜　🔍 查看大图　｜　⟳ 向左旋转　｜　C 向右旋转

Type I error
(false positive)
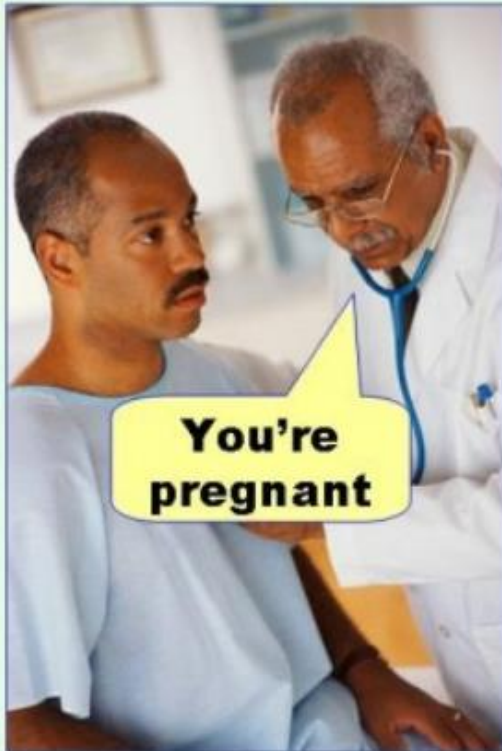
Type II error
(false negative)

You're pregnant

You're not pregnant

@游识猷

2014-5-10 13:17 来自 微博 weibo.com　　　　📤 1081　｜　💬 3　｜　👍 86
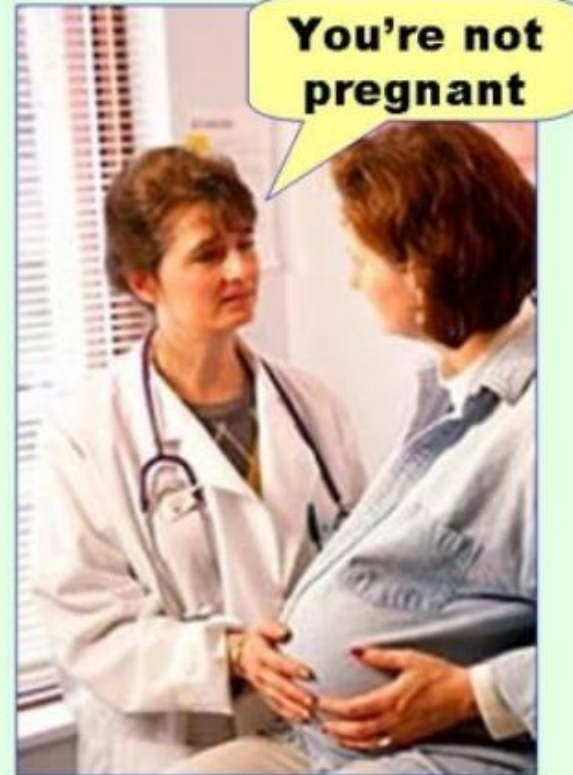
Null hypothesis:
everyone is normal
not pregnant

40

Controlling Type I and Type II Errors

- α, β, and *n* are interrelated. If one is kept constant, then an increase in one of the remaining two will cause a decrease in the other.

- For any fixed α, an increase in the sample size *n* will cause a decrease in β.

- For any fixed sample size *n* , a decrease in α will cause an increase in β.

- Conversely, an increase in α will cause a decrease in β .

- To decrease both α and β, increase the sample size *n.*

Power of test

Power (defined by Neyman and Pearson)

- Type I error: alpha (α). We say different, but really same.

- Type II errors: beta β. We say same, but really different. **Power is 1- β**.

- It is desirable to have both a small alpha (few Type I errors) and good power (few Type II errors), but it is a trade-off.

- Need a specific effect size to calculate β.

Power

Height of grade 4 students: 138 cm

Height of grade 5 students: 142 cm

- Suppose: $H_0 : \mu = 138; \quad H_1 : \mu = 142; \quad \sigma = 20; \quad N = 100$
- Rejection region is set for alpha =.05.

$$\sigma_M = \frac{20}{\sqrt{100}} = 2$$

$$Bound = 138 + 1.65\sigma_M = 141.3$$

$\alpha = p(reject\ H_0 | \mu = 138)$
$\alpha = p(reject\ H_0 | H_0) = .05$

$\beta = p(not\ reject\ H_0 | \mu = 142)$
$\beta = p(not\ reject\ H_0 | H_1)$

44
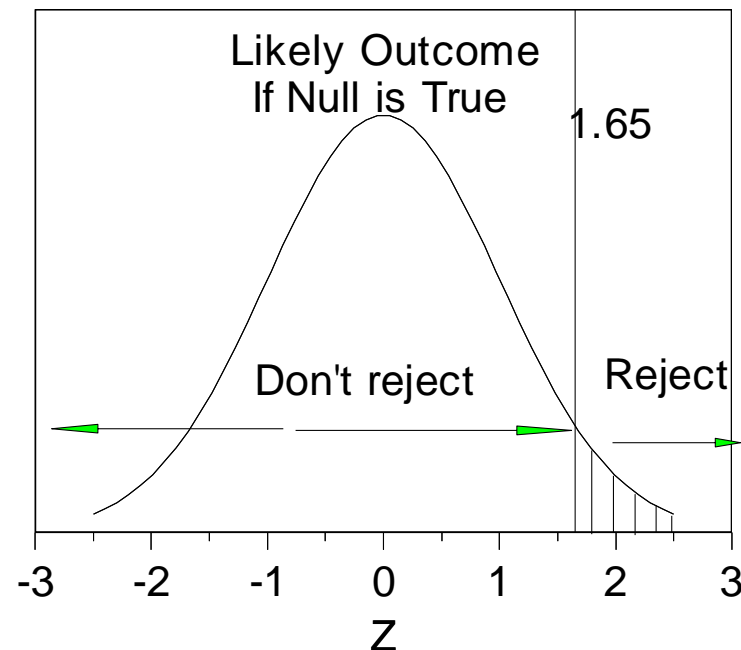
Power

The bound is a bit below the mean of the second distribution (142). It is z=(141.3-142)/2 = -.35. The area corresponding to z is 36%. This means that Beta is .36 and power is .64.

Four things affect power:
1. H_1, the alternative hypothesis.
2. The value and placement of rejection region.
3. Sample size.
4. Population variance.

141.3

Power (1-beta)

Beta

138   142

45
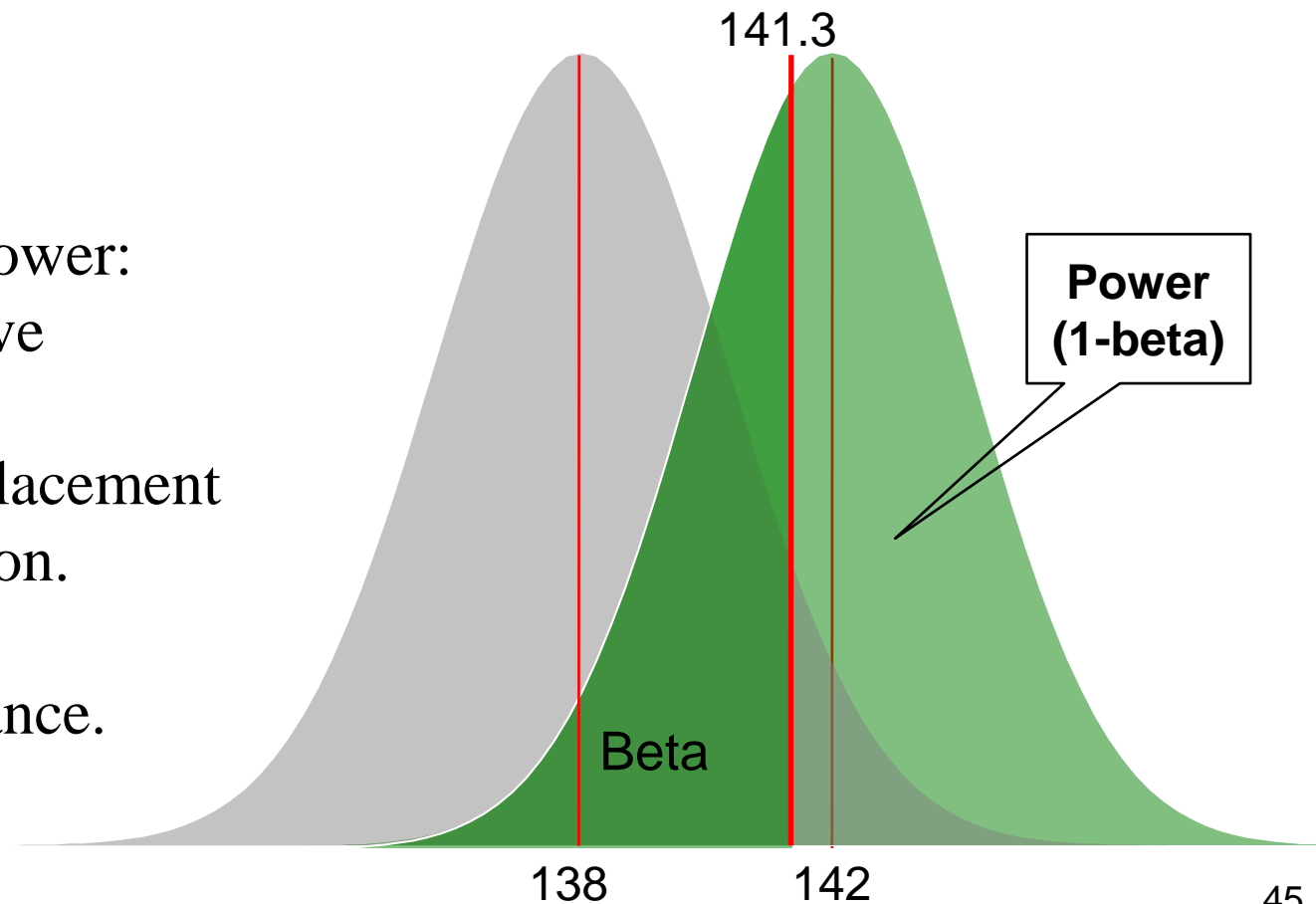
Power

The larger the difference in means, the greater the power.
This illustrates the choice of H_1.
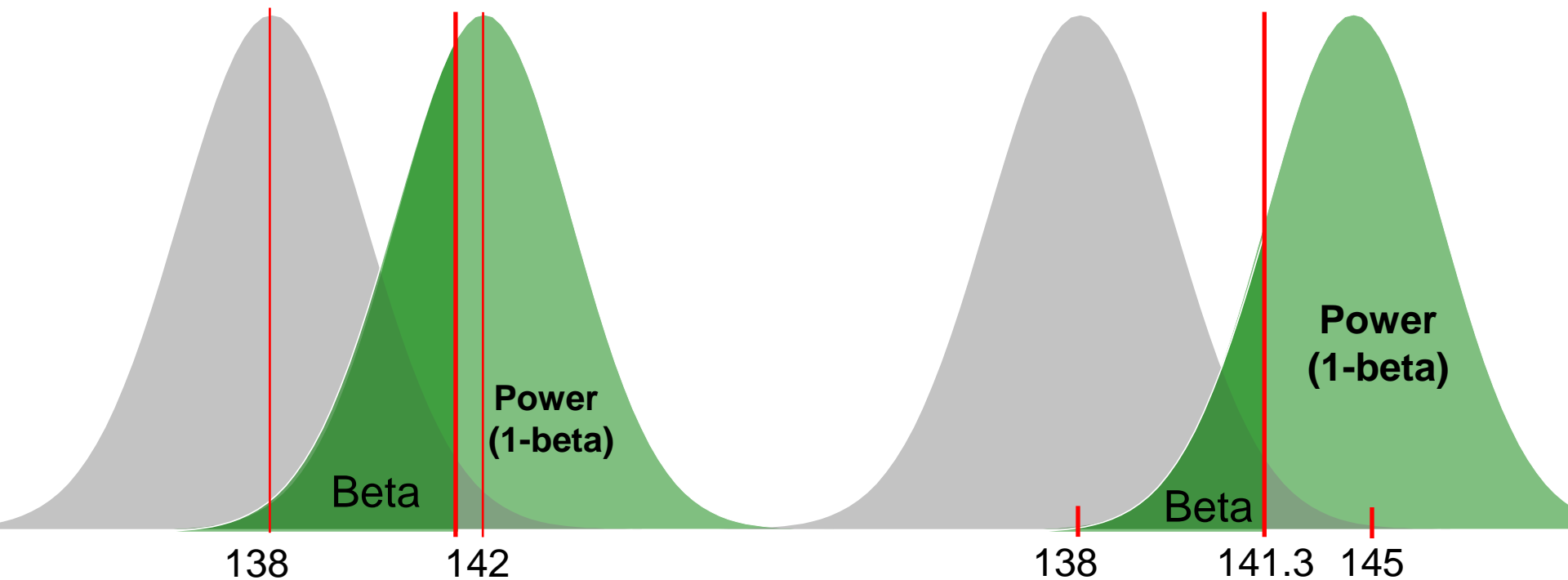
Example

Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours.

Assume actual mean light bulb lifetime is 9,950 hours and the population standard deviation is 120 hours.

At .05 significance level, what is the probability of having type II error for a sample size of 30 light bulb?

9950   10000

http://www.r-tutor.com/elementary-statistics/hypothesis-testing/lower-tail-test-population-mean-unknown-variance
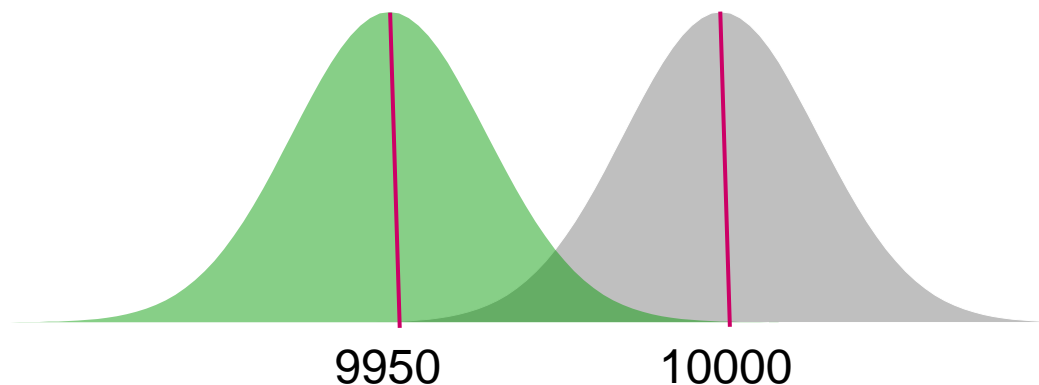
Example
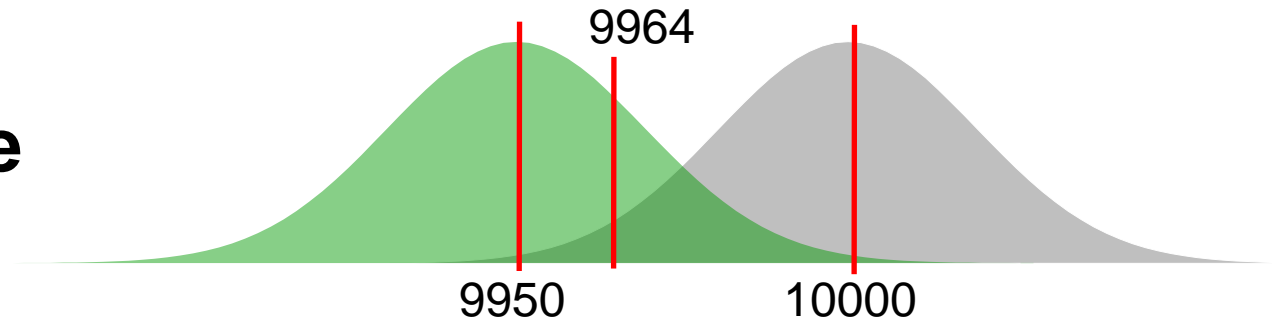
Solution

The null hypothesis is that μ ≥ 10000. We begin with computing the test statistic.

n = 30   # sample size
sigma = 120   # population standard deviation
sem = sigma/sqrt(n); sem # standard error
alpha = .05   # significance level
mu0 = 10000   # hypothetical lower bound
q = qnorm(alpha, mean=mu0, sd=sem); q
[1] 9964

Therefore, so long as the sample mean is less than 9964 in the hypothesis test, the null hypothesis will be rejected.

Example
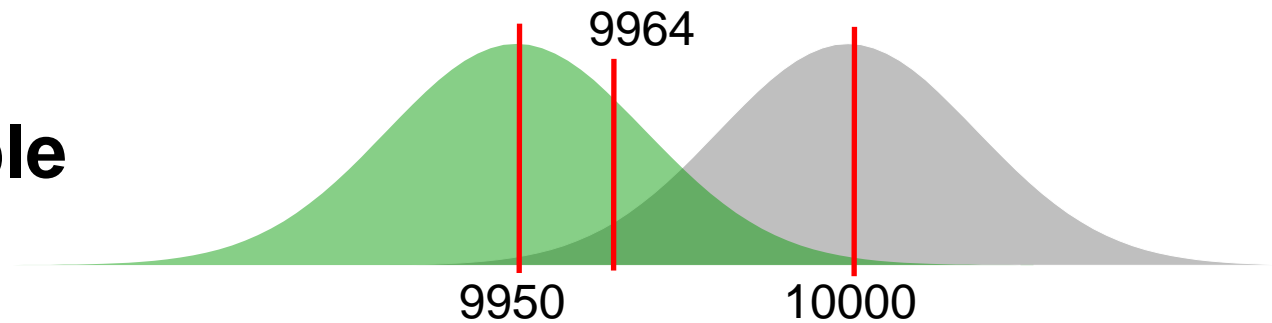
Since we assume that the actual population mean is 9950, we can compute the probability of the sample mean above 9964, and thus found the probability of type II error.

mu = 9950   # assumed actual mean
pnorm(q, mean=mu, sd=sem, lower.tail=FALSE)

[1] 0.26196

Answer

If the light bulbs sample size is 30, the actual mean light bulb lifetime is 9,950 hours and the population standard deviation is 120 hours, then the probability of type II error for testing the null hypothesis $\mu \geq 10000$ at .05 significance level is 26.2%, and the power of the hypothesis test is 73.8%.

http://www.r-tutor.com/elementary-statistics/hypothesis-testing/lower-tail-test-population-mean-unknown-variance

Power

Sample size, alpha level, effect size, and population variability affect the power.

power.t.test (n = 20, delta = 1.5, sd = 2, sig.level = 0.05,
type = "one.sample", alternative = "two.sided", strict = TRUE)
\# where n is the sample size, delta is true difference in means, and type indicates
\# a two-sample t-test, one-sample t-test or paired t-test.
> power = 0.8888478

\# If you have unequal sample sizes, use
library (pwr)
pwr.t2n.test (n1 = 30, n2 = 50,
d = -.5, sig.level = 0.05,
alternative = "less")

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}$$

power

beta

H_0 H_1

50
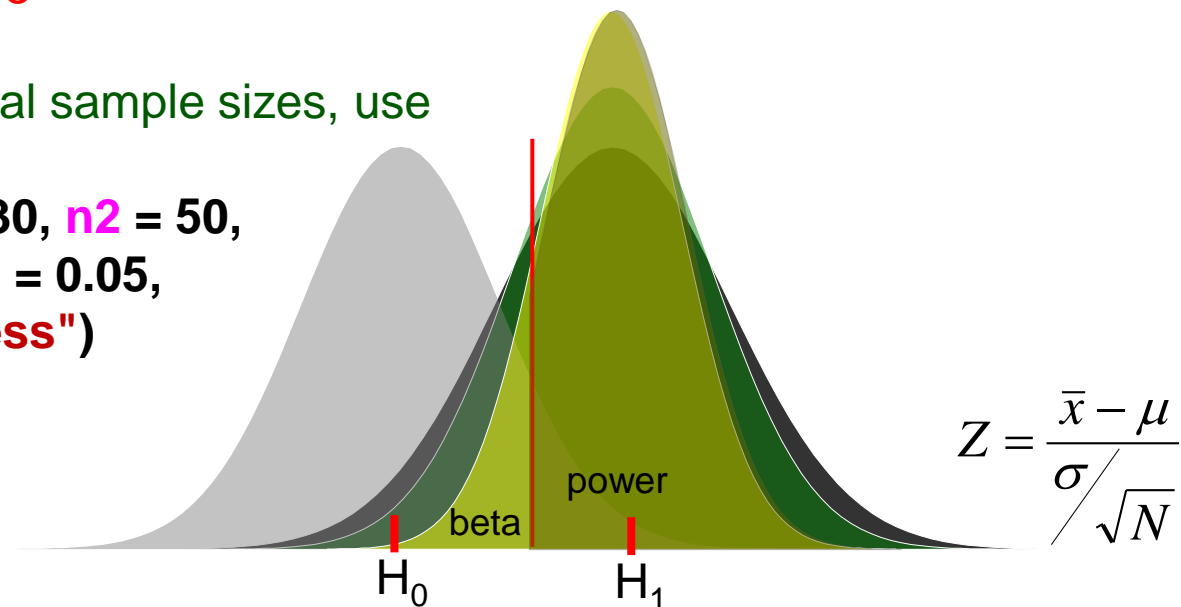
Key factors affecting power

1. **Sample size** is the easiest factor to manipulate. The larger your sample, the greater your power. But, like the precision of a confidence interval, power only goes up as fast as $\sqrt{n}$.

2. **The difference in means** you are looking for also affects the power. You should know what kind of difference you are looking for before you plan a study.

3. **The variation of your measurements** also affects the power. If individuals vary a great deal within group, it will take a larger sample size to see the differences between groups.

4. **The level you require for the P value** affects power; if you make the level 0.01 instead of 0.05 it will be harder to reject and power will go down.

Power of one sample t test

$$t_{\beta(1),v} = \frac{\delta}{\sqrt{\dfrac{s^2}{n}}} - t_{\alpha,v}$$

$t_{\alpha,v}$ $t_{\beta(1),v}$
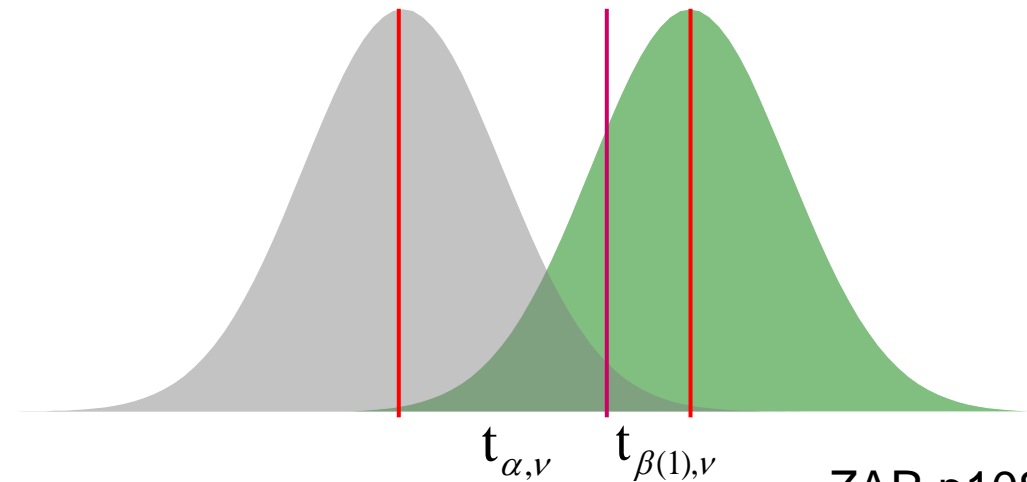
ZAR p108

EXAMPLE7.9 Estimation of power of a one-sample *t* test.

What is the probability of detecting a true difference (i.e. a difference between μ and μ_0) of at least 1.0 g for the experiment of Example 7.2?

For $n = 12$, $v = 11$, $t_{0.05(2),11} = 2.201$, and $s^2 = 1.5682$ g^2 , and we use the above equation to find

$$t_{\beta(1),11} = \frac{1.0}{\sqrt{\dfrac{1.5682}{12}}} - 2.201 = 0.57$$

By considering 0.57 to be a normal deviate, we conclude $\beta = 0.28$ and that the power of the test $1 - \beta = 0.72$.

52

Setting error levels
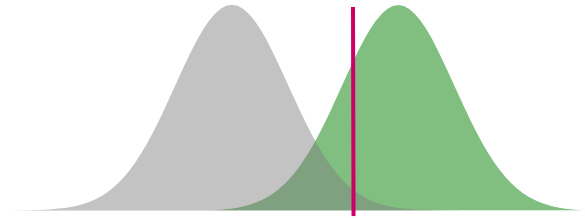
- α is controlled by setting critical *P*-value for rejecting null hypothesis

- β decreased by
  - increasing α
  - Increasing sample size (n)
  - Decreasing sample variance, var(x)
  - increasing effect size, Δ

- Tradeoff between α and β

- Need to balance costs associated with type I and type II errors

Power and the effect size

$\sigma = 3$

$H_0: \mu \leq 56$

Probability of rejecting H0: $\alpha = 0.05$

Power vs True mean, with legend n=6. X-axis: True mean (55, 57, 59, 61). Y-axis: Power (0 to 1).
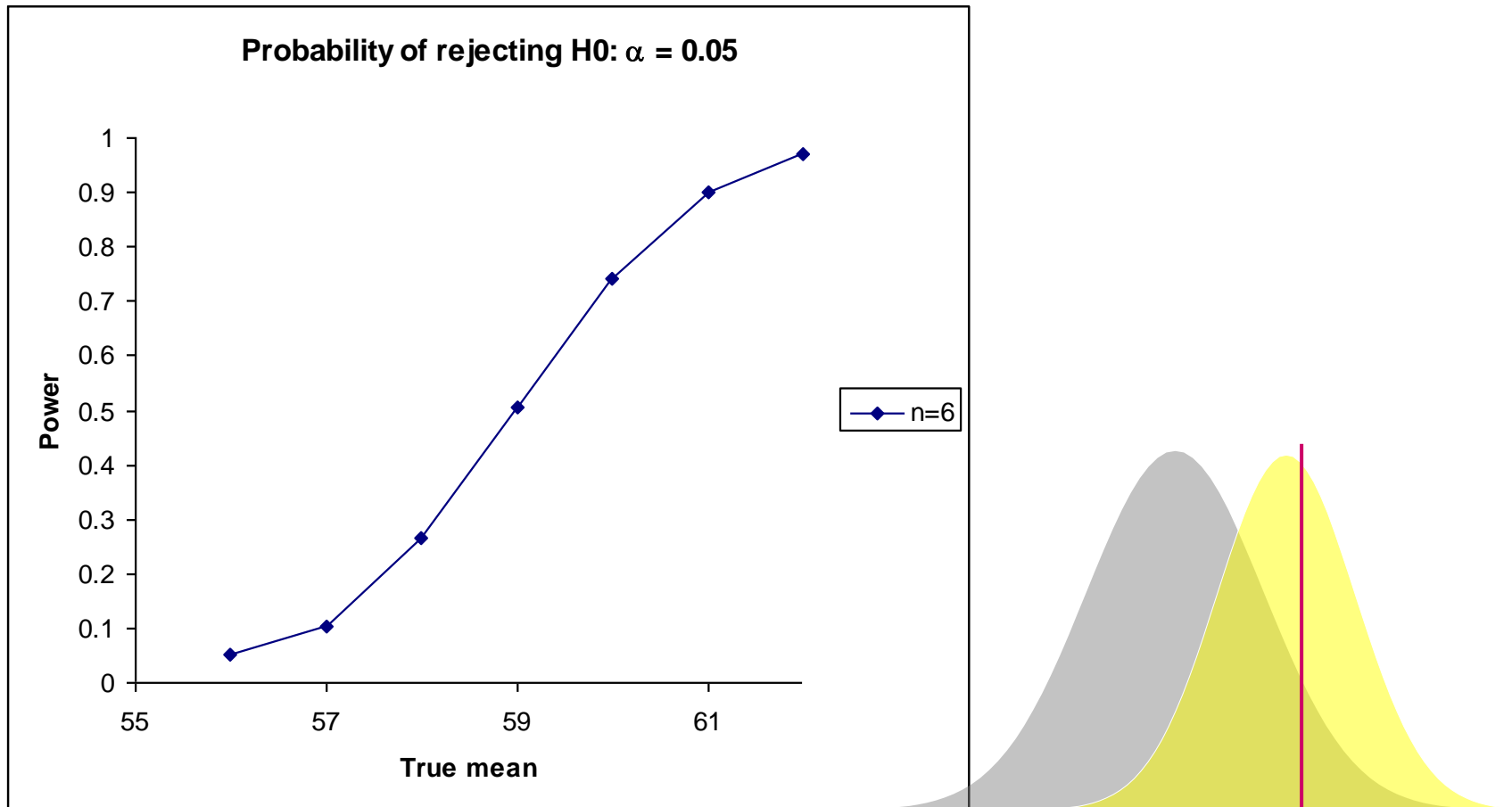
Effect of sample size on power

Probability of rejecting H0: α = 0.05
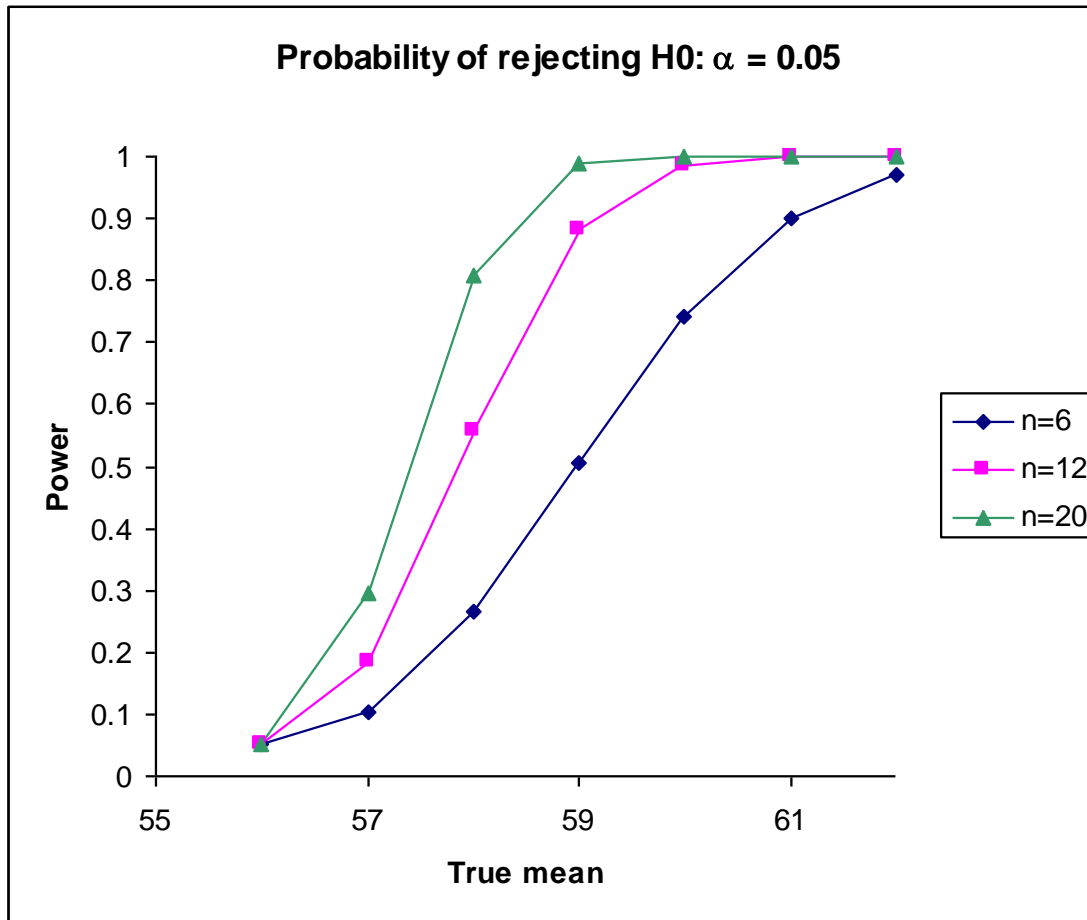
$\sigma = 3$

$H_0: \mu \leq 56$

Sample size *n* and beta

Distribution of sample mean, N=16
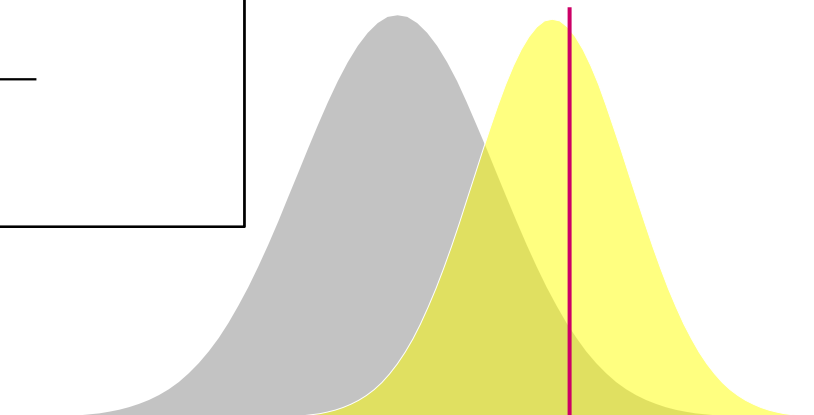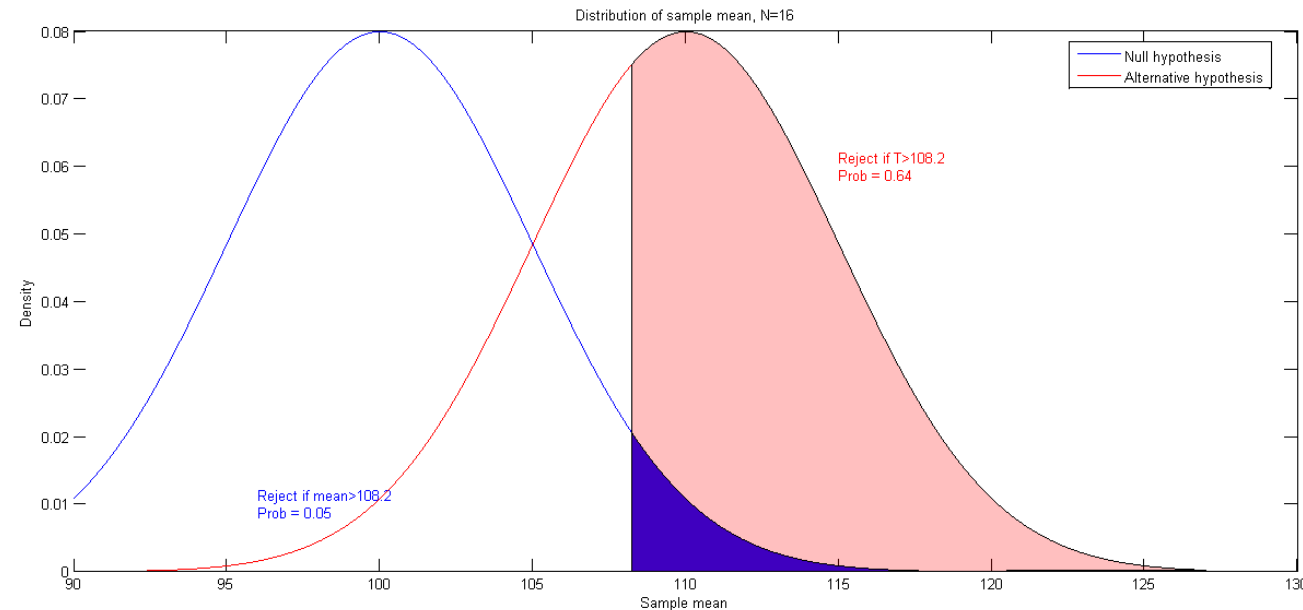
H_0: mean=100

H_1: mean=110

s=20; N=16;

$mean_{critical}$=108 at 0.05 sig.

At 108 there is a 64% chance that it belongs to the alternative population (mean=110).

Distribution of sample mean, N=30
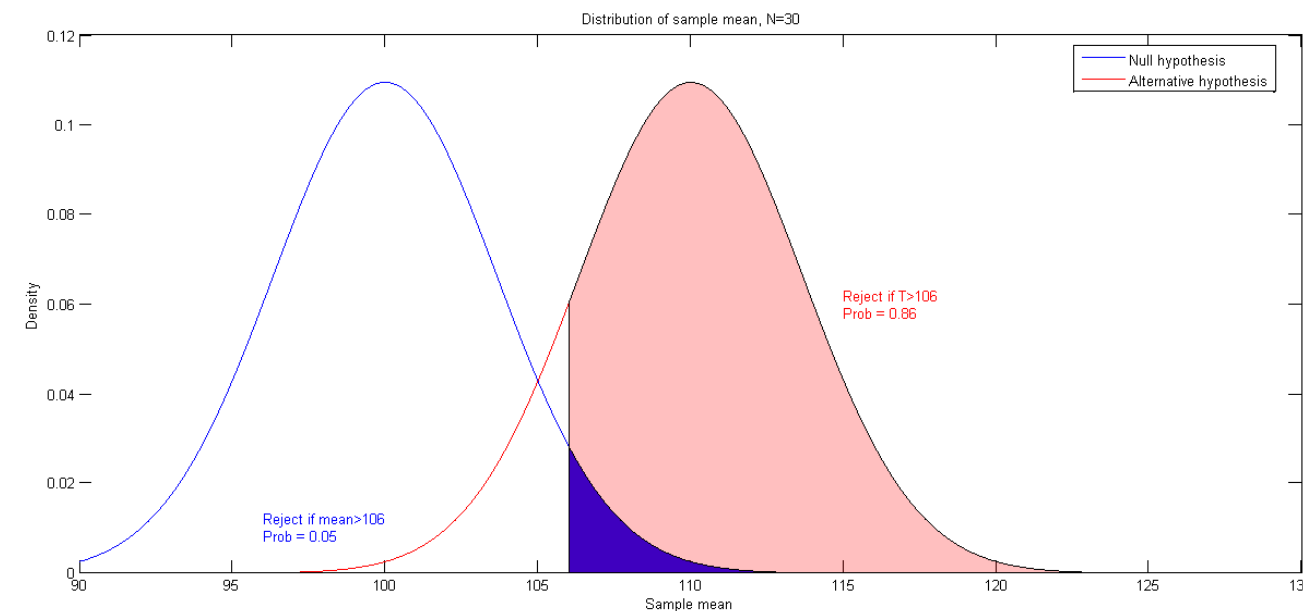
N=30, $mean_{critical}$=106

at 0.05 sig.

At 106 there is a 86% chance that it belongs to the alternative population (mean=110).

qnorm(0.95, mean=100, sd=20/sqrt(30))

pnorm(106, mean=110, sd=20/sqrt(30),

        lower.tail = F)

56

Power

```r
# Quick-R http://www.statmethods.net/stats/power.html
# Plot sample size curves for detecting correlations of
# various sizes.
library(pwr)
# range of correlations
r <- seq(.1,.5,.01)
nr <- length(r)
# power values
p <- seq(.4,.9,.1)
np <- length(p)
# obtain sample sizes
samsize <- array(numeric(nr*np), dim=c(nr,np))
for (i in 1:np){
  for (j in 1:nr){
     result <- pwr.r.test(n = NULL, r = r[j],
     sig.level = .05, power = p[i],
     alternative = "two.sided")
     samsize[j,i] <- ceiling(result$n)
}}
samsize
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	292	384	489	616	782	1046
[2,]	241	318	404	509	646	864
[3,]	203	267	340	427	542	725
[4,]	173	227	289	364	462	617
[5,]	149	196	249	313	398	532
[6,]	130	171	217	273	346	463
[7,]	114	150	191	240	304	406
[8,]	101	133	169	212	269	359
[9,]	91	119	151	189	240	320
[10,]	81	106	135	169	215	287
[11,]	74	96	122	153	194	258
[12,]	67	87	110	138	175	234
[13,]	61	79	101	126	160	213
[14,]	56	73	92	115	146	194
[15,]	51	67	84	106	134	178
[16,]	47	62	78	97	123	164
[17,]	44	57	72	90	113	151
[18,]	41	53	67	83	105	140
[19,]	38	49	62	77	97	130
[20,]	35	46	58	72	91	120
[21,]	33	43	54	67	85	112
[22,]	31	40	50	63	79	105
[23,]	29	38	47	59	74	98
[24,]	28	35	44	55	69	92
[25,]	26	33	42	52	65	86
[26,]	25	31	39	49	61	81
[27,]	23	30	37	46	58	77
[28,]	22	28	35	44	55	72
[29,]	21	27	33	41	52	68
[30,]	20	25	32	39	49	65
[31,]	19	24	30	37	46	61
[32,]	18	23	29	35	44	58
[33,]	17	22	27	34	42	55
[34,]	17	21	26	32	40	52
[35,]	16	20	25	30	38	50
[36,]	15	19	24	29	36	47
[37,]	15	18	23	28	34	45
[38,]	14	18	22	26	33	43
[39,]	13	17	21	25	31	41
[40,]	13	16	20	24	30	39
[41,]	12	16	19	23	29	38

Plot the curves

Sample Size Estimation for Correlation Studies

Sig=0.05 (Two-tailed)

```
# set up graph
xrange <- range(r)
yrange <- round(range(samsize))
colors <- rainbow(length(p))
plot(xrange, yrange, type="n",
  xlab = "Correlation Coefficient (r)",
  ylab = "Sample Size (n)" )
# add power curves
for (i in 1:np){
  lines(r, samsize[,i],type="l",lwd=2,col=colors[i])
}
# add annotation (grid lines, title, legend)
abline(v=0, h=seq(0,yrange[2],50), lty=2, col="grey89")
abline(h=0, v=seq(xrange[1],xrange[2],.02), lty=2,col="grey89")
title("Sample Size Estimation for Correlation Studies\n
      Sig=0.05 (Two-tailed)")
legend("topright", title="Power", as.character(p), fill=colors)
```

58
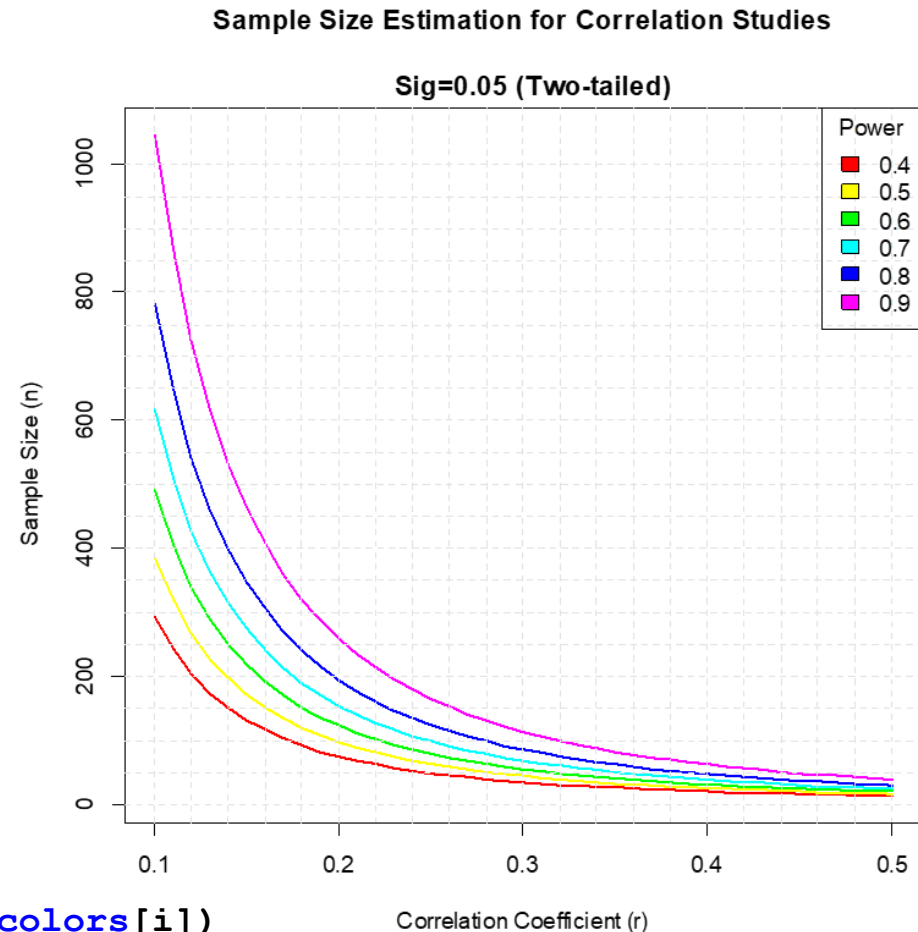
R functions for power calculations
library(pwr)

pwr.2p.test   two proportions (equal n)

pwr.2p2n.test   two proportions (unequal n)

pwr.anova.test   balanced one way ANOVA

pwr.chisq.test   chi-square test

pwr.f2.test   general linear model

pwr.p.test   proportion (one sample)

pwr.r.test   correlation

pwr.t.test   t-tests (one sample, 2 sample, paired)

pwr.t2n.test   t-test (two samples with unequal n)

Power of test

In this example the hypothesis test is: H_0: mu = 6, H_a: mu not = 6.
The standard deviation is 2, and the sample size is 20.

We will use a 95% confidence level and wish to find the power to detect a true mean that differs from 6 by an amount of 1.5.

R script:
```
a <- 6; s <- 2; n <- 20
diff <- qt(0.975, df = n-1)*s/sqrt(n)
left <- a-diff ; right <- a+diff
> left [1] 5.063971
> right [1] 6.936029
```

Next we find the Z-scores for the left and right values assuming that the true mean is 6+1.5=7.5:

```
assumed <- a + 1.5
tleft  <- (assumed - right)/(s/sqrt(n)) #1.261
p <- pt(-tleft, df = n-1); p #0.1112583
```

The probability that we make a type II error if the true mean is 7.5 is approximately 11.1%.
So the power of the test is **1-p #0.888**

In this example, the power of the test is approximately 88.8%.
If the true mean differs from 6 by 1.5 then the probability that we will reject the null hypothesis is approximately 88.8%.

Sample size

Sample size calculations for fixed power

Goal - Choose sample sizes to have a favorable chance of detecting a *clinically meaning difference*

Step 1 - Define an important difference in means:

– **Case 1:** σ approximated from prior experience or pilot study - difference can be stated in units of the data

– **Case 2:** σ unknown - difference must be stated in units of standard deviations of the data

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

Step 2 - Choose the desired power to detect the clinically meaningful difference (1-b, typically at least .80). For 2-sided test:

$$n_1 = n_2 = \frac{\sigma^2 \left(z_{\alpha/2} + z_\beta \right)^2}{d^2}$$

$$t_{\beta(1),v} = \frac{\delta}{\sqrt{\frac{s^2}{n}}} - t_{\alpha,v}$$

Example - Rosiglitazone vs. Placebo

- **Treat:** Rosiglitazone vs. Placebo
- **Response:** Change in Limb fat mass
- **Clinically Meaningful Difference:** 0.5 (std dev's)
- **Desired Power:** $1-\beta = 0.80$
- **Significance Level:** $\alpha = 0.05$

$$z_{\alpha/2} = 1.96 \quad z_{\beta} = z_{.20} = .84$$

$$n_1 = n_2 = \frac{1 \times (1.96 + 0.84)^2}{(0.5)^2} = 31$$

| 0.84 |

0   1.96   2.8
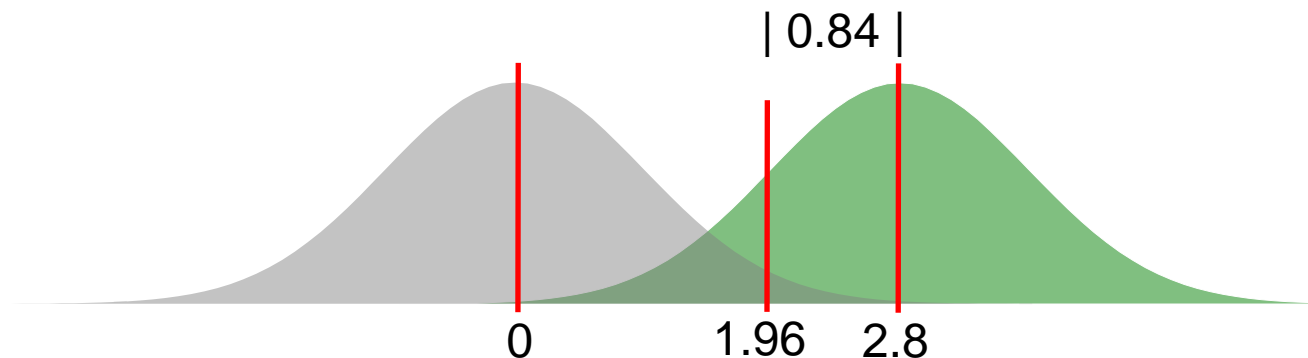
Sample size – one sample t test

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \qquad n = \frac{\sigma^2 z^2_{\alpha(2),(n-1)}}{(\bar{x} - \mu)^2}$$

EXAMPLE 7.6 Determination of sample size needed to achieve stated precision in estimating a population mean, using the data of Example 7.3

To estimate μ with a 95% confidence interval no wider than 0.5 kg, then $d = 0.25$ kg, $1 - \alpha = 0.95$, and $\alpha = 0.05$. From Example 7.3 we have an estimate of the population variance: $s^2 = 0.4008$ kg^2 .

Let us guess that a sample of 40 is necessary; then,

$$t_{0.05(2),39} = 2.023$$

So we estimate:

$$n = \frac{(0.4008)(2.0023)^2}{(0.25)^2} = 26.2$$

Next, we might estimate n = 27, for which $t_{0.05(2),26} = 2.056$ and we calculate:

$$n = \frac{(0.4008)(2.056)^2}{(0.25)^2} = 27.1$$

Therefore, we conclude that a sample size greater than 27 is required to achieve the specified confidence interval.

Philosophy

Philosophy of hypothesis testing

Ronald Aylmer Fisher, Jerzy Neyman and Egon Pearson had developed their approaches for hypothesis testing by the 1930s.

Cox (1977) termed Fisher's procedure: significance testing
    Neyman and Pearson's procedure: hypothesis testing.

The philosophical justification for the continued use of hypothesis testing is based on Popper's proposals for **falsification** tests of hypotheses.

Popper asserted that a hypothesis, proposition, or theory is scientific only if it is falsifiable.

For example, "all men are mortal" is unfalsifiable;
"All men are immortal," by contrast, is falsifiable.

Philosophy of hypothesis testing

Sir Karl Raimund Popper (1902-1994) was an Austro-British philosopher and a professor at the London School of Economics. He is regarded as one of the greatest philosophers of science of the 20th century.

Popper used the term **critical rationalism** (vs. comprehensive rationalism) to describe his philosophy, against classical empiricism, and the classical observation-induction method.

Popper criticised psychologism, naturalism, inductionism, and logical positivism, and put forth his theory of potential falsifiability as the criterion separating science from non-science.

Popper, Karl. R. (1934) *The Logic of Scientific Discovery*. Hutchinson, London.
Popper, Karl. R. (1945) *The Open Society and Its Enemies* . Routledge, London.
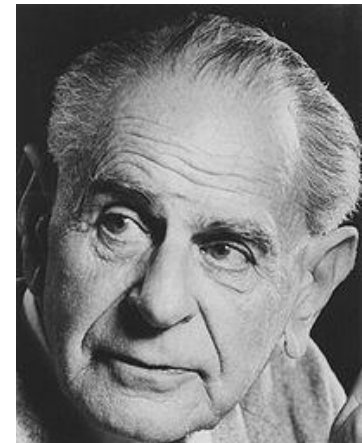
Critiques to hypothesis testing

The philosophical justification for testing the null hypothesis is still a controversial issue.

Over the past 60 years an increasing number of articles have questioned the utility of hypothesis testing (Anderson et al. 2000).

Schmidt (1996) have felt that the misuse of hypothesis testing was sufficiently widespread to justify its being banned from use within the journals of the American Psychological Association (APA).

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. Journal of Wildlife Management, 6: 912–923.
Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1: 115–129.

Silly null hypotheses

The H_0 is simply the complement of the research hypothesis about which we are trying to make a decision (Chow 1988, 1991; Mulaik *et al.* 1997).

Chow, S.L. (1988) Significance test or effect size? *Psychological Bulletin* **103**: 105–110.
Chow, S.L. (1991) Some reservations about power analysis. *American Psychologist* **46**: 1088.
Mulaik, S.A., Raju, N.S. & Harshman, R.A. (1997) There is a time and a place for significance testing. In:
        What if there were no significance tests? (Harlow, L.L., Mulaik, S.A. & Steiger, J.H. eds.), pp. 65–115.
        Lawrence Erlbaum, New Jersey.

Silly null hypotheses

Typical null hypothesis is almost always false.

Excessive use of p values.

The size and direction of observed differences should be reported, not the naked *p values* (Anderson et al., 2001).

Need to consider:
• Effect size
• *P* value
• Sample size

Anderson, D. R., Link, W. A., Johnson, D. H., & Burnham, K. P. (2001). Suggestions for presenting the results of data analysis. *Journal of Wildlife Management, 65: 373*-378.

Reporting effect sizes

Thompson (2000) reported that over the past few years, more than a dozen journals in education-related fields have instituted policies that require authors to provide effect sizes in addition to *p values*

- Contemporary Educational Psychology
- Educational and Psychological Measurement
- Journal of Agricultural Education
- Journal of Applied Psychology
- Journal of Consulting & Clinical Psychology
- Journal of Early Intervention
- Journal of Experimental Education
- Journal of Learning Disabilities, Language Learning
- Measurement and Evaluation in Counseling and Development
- The Professional Educator and Research in the Schools

Requiring authors to always provide effect size information may distract or mislead readers, when such information adds little to the correct interpretation of the data.

For example, a major use of hypothesis testing is in testing model fit, such as using a likelihood ratio to compare a restricted model to its more general parent. What does effect size mean in this context?

Thompson, B. (2000). Various editorial policies regarding statistical significance tests and effect sizes. Retrieved November 20, 2001, from http://www.coe.tamu.edu/~bthompson/journals.htm

Fisher's original plan

Fisher (1926) adopted the α of 0.05 to screen for potentially useful innovations.

Fisher understood science as a continuous process. He believed hypothesis testing only made sense in the context of a continuing series of experiments that were aimed at checking the effects of specific treatments.

He used statistical tests to come to one of three conclusions.

- When p <0.05, he declared that an effect has been demonstrated;

- When 0.05 < p < 0.2, he concluded that if there is an effect, it is too small to be detected with an experiment this size; he discussed how to design the next experiment to estimate the effect better;

- When p > 0.2, no effect.

Fisher, R. A. (1926). The arrangement of field experiments. Journal of the Ministry of Agriculture of Great Britain, 33: 503-513.

Arbitrary α Levels

One long-standing criticism has been the arbitrary use of 0.05 as the criterion for rejecting or not rejecting H_0. Fisher originally suggested 0.05 but later argued against using a single significance level for every statistical decision-making process.

The fact that many persons misuse hypothesis testing by simply making reject / fail to reject decisions on single studies is probably due to the Neyman-Pearson legacy of such dichotomous decisions.

Researchers should not be bound by the chains of α = 0.05.

The *p* values should be reported.

Notes

- A large sample size can help get a small p-value

- Failing to reject H_0 means:
  - There is not enough evidence to reject H_0
  - Does NOT mean H_0 is true

A journal banning hypothesis testing

The Basic and Applied Social Psychology (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014).

Now BASP is banning the NHSTP (Trafimow & Marks, 2015).

"But prior to publication, authors will have to remove all vestiges of the NHSTP (p-values, t-values, F-values, statements about "significant" differences or lack thereof, and so on)." (Trafimow & Marks, 2015).

"Confidence intervals suffer from an inverse inference problem that is not very different from that suffered by the NHSTP." (Trafimow & Marks, 2015).

"Bayesian procedures are neither required nor banned from BASP." (Trafimow & Marks, 2015).

Trafimow, D. 2014. Editorial . Basic and Applied Social Psychology, 36(1), 1-2.
Trafimow, D. and Marks, M. 2015. Editorial . Basic and Applied Social Psychology, 37(1), 1-2.

What the journal BASP suggests

"BASP will require strong descriptive statistics, including effect sizes.

We also encourage the presentation of frequency or distributional data when this is feasible.

Finally, we encourage the use of larger sample sizes than is typical in much psychology research, because as the sample size increases, descriptive statistics become increasingly stable and sampling error is less of a problem.

we will stop requiring particular sample sizes, because it is possible to imagine circumstances where more typical sample sizes might be justifiable."

Trafimow, D. and Marks, M. 2015. Editorial . Basic and Applied Social Psychology, 37(1), 1-2.

What the journal BASP expects

"We believe that the $p < .05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research.

We hope and anticipate that banning the NHSTP will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking.

The NHSTP has dominated psychology for decades; we hope that by instituting the first NHSTP ban, we demonstrate that psychology does not need the crutch of the NHSTP, and that other journals follow suit."

Trafimow, D. and Marks, M. 2015. Editorial . Basic and Applied Social Psychology, 37(1), 1-2.

The argument about the p value

$$t = \frac{\overline{X} - \mu}{\sigma / \sqrt{N}}$$

Charles J. Krebs in Beijing on 2014-Nov-26

Nature 567, 305-307 (2019)

COMMENT

EVOLUTION Cooperation and conflict from ants and chimps to us **p.308**

HISTORY To fight denial, study Galileo and Arendt **p.309**

CHEMISTRY Three more unsung women — of astatine discovery **p.311**

PUBLISHING As well as ORCID ID and English, list authors in their own script **p.311**

Scientists rise up against statistical significance - Nature

www.nature.com

Retire statistical significance

Valentin Amrhein, **Sander Greenland**, **Blake McShane** and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

When was the last time you heard a seminar speaker claim there was 'no difference' between two groups because the difference was 'statistically non-significant'?

If your experience matches ours, there's a good chance that this happened at the last talk you attended. We hope that at least someone in the audience was perplexed if, as frequently happens, a plot or table showed that there actually was a difference.

How do statistics so often lead scientists to deny differences that those not educated in statistics can plainly see? For several generations, researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome)[1]. Nor do statistically significant results 'prove' some other hypothesis. Such misconceptions have famously warped the

literature with overstated claims and, less famously, led to claims of conflicts between studies where none exists.

We have some proposals to keep scientists from falling prey to these misconceptions.

PERVASIVE PROBLEM
Let's be clear about what must stop: we should never conclude there is 'no difference' or 'no association' just because a *P* value is larger than a threshold such as 0.05 ▶

BEWARE FALSE CONCLUSIONS

Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.

'Significant' study (low *P* value)

'Non-significant' study (high *P* value)

The observed effect (or point estimate) is the same in both studies, so they are not in conflict, even if one is 'significant' and the other is not.

Decreased effect ◀ No effect ▶ Increased effect   ©nature
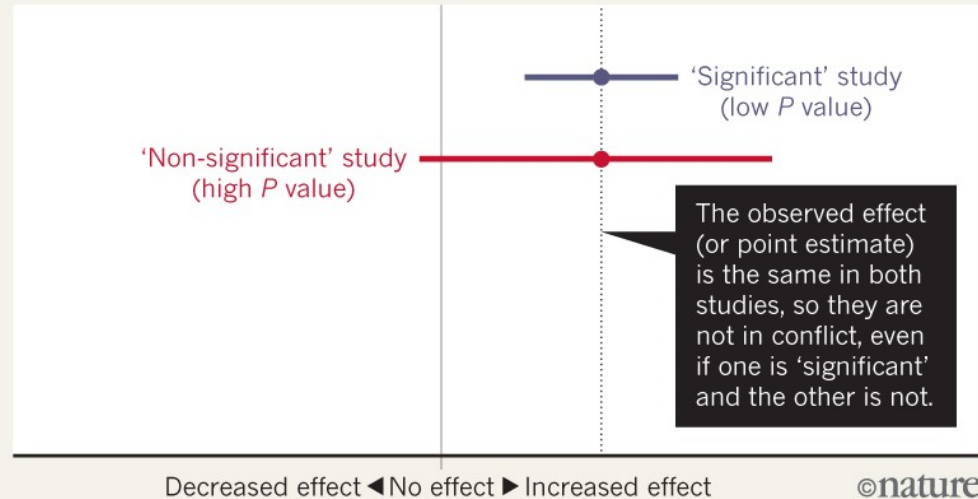
WRONG INTERPRETATIONS

An analysis of 791 articles across 5 journals* found that around half mistakenly assume non-significance means no effect.

Appropriately interpreted **49%**

Wrongly interpreted **51%**

ARTICLES **791**
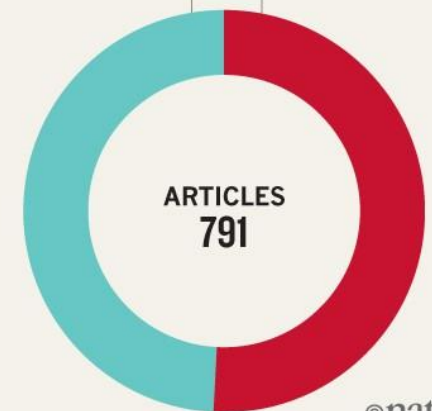
*Data taken from: P. Schatz *et al.* *Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler *et al.* *Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra *et al.* *Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi *et al.* *Eur. Sociol. Rev.* **33**, 1–15 (2017).

©nature

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

Q: Why do so many colleges and grad schools teach $p = 0.05$?
A: Because that's still what the scientific community and journal editors use.
Q: Why do so many people still use $p = 0.05$?
A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

The ASA Board was also stimulated by highly visible discussions over the last few years. For example, ScienceNews (Siegfried 2010) wrote: "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation." A November 2013, article in Phys.org Science News Wire (2013) cited "numerous deep flaws" in null hypothesis significance testing. A ScienceNews article (Siegfried 2014) on February 7, 2014, said "statistical techniques for testing hypotheses …have more flaws than Facebook's privacy policies." A week later, statistician and "Simply Statistics" blogger Jeff Leek responded. "The problem is not that people use P-values poorly," Leek wrote, "it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis" (Leek 2014). That same week, statistician and science writer Regina Nuzzo published an article in *Nature* entitled "Scientific Method: Statistical Errors" (Nuzzo 2014). That article is now one of the most highly viewed *Nature* articles, as reported by altmetric.com (http://www.altmetric.com/details/2115792#score).

Of course, it was not simply a matter of responding to some articles in print. The statistical community has been deeply concerned about issues of *reproducibility* and *replicability* of scientific conclusions. Without getting into definitions and distinctions of these terms, we observe that much confusion and even doubt about the validity of science is arising. Such doubt can lead to radical choices, such as the one taken by the editors of *Basic and Applied Social Psychology*, who decided to ban *p*-values (null hypothesis significance testing) (Trafimow and Marks 2015). Misunderstanding or misuse of statistical inference is only one cause of the "reproducibility crisis" (Peng 2015), but to our community, it is an important one.

When the ASA Board decided to take up the challenge of developing a policy statement on *p*-values and statistical significance, it did so recognizing this was not a lightly taken step. The ASA has not previously taken positions on specific matters of statistical practice. The closest association has come to this is a statement on the use of value-added models (VAM) for educational assessment (Morganstein and Wasserstein

2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on *p*-values and statistical significance would shed light on an aspect of our field that is too often misunderstood and misused in the broader research community, and, in the process, provides the community a service. The intended audience would be researchers, practitioners, and science writers who are not primarily statisticians. Thus, this statement would be quite different from anything previously attempted.

The Board tasked Wasserstein with assembling a group of experts representing a wide variety of points of view. On behalf of the Board, he reached out to more than two dozen such people, all of whom said they would be happy to be involved. Several expressed doubt about whether agreement could be reached, but those who did said, in effect, that if there was going to be a discussion, they wanted to be involved.

Over the course of many months, group members discussed what format the statement should take, tried to more concretely visualize the audience for the statement, and began to find points of agreement. That turned out to be relatively easy to do, but it was just as easy to find points of intense disagreement.

The time came for the group to sit down together to hash out these points, and so in October 2015, 20 members of the group met at the ASA Office in Alexandria, Virginia. The 2-day meeting was facilitated by Regina Nuzzo, and by the end of the meeting, a good set of points around which the statement could be built was developed.

The next 3 months saw multiple drafts of the statement, reviewed by group members, by Board members (in a lengthy discussion at the November 2015 ASA Board meeting), and by members of the target audience. Finally, on January 29, 2016, the Executive Committee of the ASA approved the statement.

The statement development process was lengthier and more controversial than anticipated. For example, there was considerable discussion about how best to address the issue of multiple *potential* comparisons (Gelman and Loken 2014). We debated at some length the issues behind the words "a *p*-value near 0.05 taken by itself offers only weak evidence against the null

American Statistical Association (ASA) took actions

- In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

Q: Why do so many colleges and grad schools teach *p* = 0.05?

A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people still use *p* = 0.05?

A: Because that's what they were taught in college or grad school.

- October 2015, 20 members of ASA met at the ASA Office in Alexandria, Virginia. By the end of 2-day meeting, a good set of points around which the statement could be built was developed.

- Wasserstein RL, Lazar NA. 2016. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician 70:129-133.

The ASA Statement on p-Values (2016)

1. *P*-values can indicate how incompatible the data are with a specified statistical model.

2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.

4. Proper inference requires full reporting and transparency.

5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Wasserstein RL, Lazar NA. 2016. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician 70:129-133.

October 2017, the American Statistical Association (ASA) held the Symposium on Statistical Inference, a two-day gathering that laid the foundations for this special issue of *The American Statistician*.

"Don't" Is Not Enough

Editorial--Statistical Inference in the 21st Century: A World Beyond p < 0.05
The American Statistician, Volume 73, Issue sup1 (2019) Volume 73, 2019

- Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the *p*-value passed some arbitrary threshold such as $p < 0.05$).

- Don't believe that an association or effect exists just because it was statistically significant.

- Don't believe that an association or effect is absent just because it was not statistically significant.

- Don't believe that your *p*-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.

- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar. 2019. Moving to a World Beyond "p < 0.05". The American Statistician 73:1-19.

Amrhein, V., D. Trafimow, and S. Greenland. 2019. Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. The American Statistician **73**:262-270.

Anderson, A. A. 2019. Assessing Statistical Results: Magnitude, Precision, and Model Uncertainty. The American Statistician **73**:118-121.

Benjamin, D. J., and J. O. Berger. 2019. Three Recommendations for Improving the Use of p-Values. The American Statistician **73**:186-191.

Betensky, R. A. 2019. The p-Value Requires Context, Not a Threshold. The American Statistician **73**:115-117.

Billheimer, D. 2019. Predictive Inference and Scientific Reproducibility. The American Statistician **73**:291-295.

Blume, J. D., R. A. Greevy, V. F. Welty, J. R. Smith, and W. D. Dupont. 2019. An Introduction to Second-Generation p-Values. The American Statistician **73**:157-167.

Brownstein, N. C., T. A. Louis, A. O'Hagan, and J. Pendergast. 2019. The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making. The American Statistician **73**:56-68.

Calin-Jageman, R. J., and G. Cumming. 2019. The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known. The American Statistician **73**:271-280.

Campbell, H., and P. Gustafson. 2019. The World of Research Has Gone Berserk: Modeling the Consequences of Requiring "Greater Statistical Stringency" for Scientific Publication. The American Statistician **73**:358-373.

Colquhoun, D. 2019. The False Positive Risk: A Proposal Concerning What to Do About p-Values. The American Statistician **73**:192-201.

Fraser, D. A. S. 2019. The p-value Function and Statistical Inference. The American Statistician **73**:135-147.

Fricker, R. D., K. Burke, X. Han, and W. H. Woodall. 2019. Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their p-Value Ban. The American Statistician **73**:374-384.

Gannon, M. A., C. A. de Bragança Pereira, and A. Polpo. 2019. Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels. The American Statistician **73**:213-222.

Goodman, S. N. 2019. Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics. The American Statistician **73**:26-30.

Goodman, W. M., S. E. Spruill, and E. Komaroff. 2019. A Proposed Hybrid Effect Size Plus p-Value Criterion: Empirical Evidence Supporting its Use. The American Statistician **73**:168-185.

Greenland, S. 2019. Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values. The American Statistician **73**:106-114.

Hubbard, D. W., and A. L. Carriquiry. 2019. Quality Control for Scientific Research: Addressing Reproducibility, Responsiveness, and Relevance. The American Statistician **73**:46-55.

Hubbard, R. 2019. Will the ASA's Efforts to Improve Statistical Practice be Successful? Some Evidence to the Contrary. The American Statistician **73**:31-35.

Hubbard, R., B. D. Haig, and R. A. Parsa. 2019. The Limited Role of Formal Statistical Inference in Scientific Inference. The American Statistician **73**:91-98.

Hurlbert, S. H., R. A. Levine, and J. Utts. 2019. Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires. The American Statistician **73**:352-357.

Ioannidis, J. P. A. 2019. What Have We (Not) Learnt from Millions of Scientific Papers with P Values? The American Statistician **73**:20-25.

Johnson, V. E. 2019. Evidence From Marginally Significant t Statistics. The American Statistician **73**:129-134.

Kennedy-Shaffer, L. 2019. Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing. The American Statistician **73**:82-90.

Kmetz, J. L. 2019. Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of p-Values. The American Statistician **73**:36-45.

Krueger, J. I., and P. R. Heck. 2019. Putting the P-Value in its Place. The American Statistician **73**:122-128.

Lavine, M. 2019. Frequentist, Bayes, or Other? The American Statistician **73**:312-318.

Locascio, J. J. 2019. The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration. The American Statistician **73**:346-351.

Manski, C. F. 2019. Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing. The American Statistician **73**:296-304.

Manski, C. F., and A. Tetenov. 2019. Trial Size for Near-Optimal Choice Between Surveillance and Aggressive Treatment: Reconsidering MSLT-II. The American Statistician **73**:305-311.

Matthews, R. A. J. 2019. Moving Towards the Post $p < 0.05$ Era via the Analysis of Credibility. The American Statistician **73**:202-212.

Maurer, K., L. Hudiburgh, L. Werwinski, and J. Bailer. 2019. Content Audit for p-value Principles in Introductory Statistics. The American Statistician **73**:385-391.

McShane, B. B., D. Gal, A. Gelman, C. Robert, and J. L. Tackett. 2019a. Abandon Statistical Significance. The American Statistician **73**:235-245.

McShane, B. B., J. L. Tackett, U. Böckenholt, and A. Gelman. 2019b. Large-Scale Replication Projects in Contemporary Psychological Research. The American Statistician **73**:99-105.

O'Hagan, A. 2019. Expert Knowledge Elicitation: Subjective but Scientific. The American Statistician **73**:69-81.

Pogrow, S. 2019. How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings. The American Statistician **73**:223-234.

Rose, S., and T. G. McGuire. 2019. Limitations of P-Values and R-squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment. The American Statistician **73**:152-156.

Rougier, J. 2019. p-Values, Bayes Factors, and Sufficiency. The American Statistician **73**:148-151.

Ruberg, S. J., F. E. Harrell, M. Gamalo-Siebers, L. LaVange, J. Jack Lee, K. Price, and C. Peck. 2019. Inference and Decision Making for 21st-Century Drug Development and Approval. The American Statistician **73**:319-327.

Steel, E. A., M. Liermann, and P. Guttorp. 2019. Beyond Calculations: A Course in Statistical Thinking. The American Statistician **73**:392-401.

Tong, C. 2019. Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science. The American Statistician **73**:246-261.

Trafimow, D. 2019. Five Nonobvious Changes in Editorial Practice for Editors and Reviewers to Consider When Evaluating Submissions in a Post $p < 0.05$ Universe. The American Statistician **73**:340-345.

van Dongen, N. N. N., J. B. van Doorn, Q. F. Gronau, D. van Ravenzwaaij, R. Hoekstra, M. N. Haucke, D. Lakens, C. Hennig, R. D. Morey, S. Homer, A. Gelman, J. Sprenger, and E.-J. Wagenmakers. 2019. Multiple Perspectives on Inference for Two Simple Statistical Scenarios. The American Statistician **73**:328-339.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar. 2019. Moving to a World Beyond "$p < 0.05$". The American Statistician **73**:1-19.

Ziliak, S. T. 2019. How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little "p" Is Not Enough. The American Statistician **73**:281-290.

In 2019 the President of the American Statistical Association (ASA), Karen Kafadar, established a task force to address concerns that a 2019 editorial in *The American Statistician* (an ASA journal) might be mistakenly interpreted as official ASA policy.

- P-values are valid statistical measures that provide convenient conventions for communicating the uncertainty inherent in quantitative results.

- Indeed, P-values and significance tests are among the most studied and best understood statistical procedures in the statistics literature.

- They are important tools that have advanced science through their proper application.

"The most reckless and treacherous of all theorists is he who professes to let facts and figures speak for themselves, who keeps in the background the part he has played, perhaps unconsciously, in selecting and grouping them." (Alfred Marshall, 1885)

Benjamini, Y., R. D. De Veaux, B. Efron, S. Evans, M. Glickman, B. I. Graubard, X. He, X.-L. Meng, N. Reid, S. M. Stigler, S. B. Vardeman, C. K. Wikle, T. Wright, L. J. Young, and K. Kafadar. 2021. The ASA president's task force statement on statistical significance and replicability. The Annals of Applied Statistics **15**:1084-1085, 1082.

P value and sample size

sample sizes for distinguishing 2% difference of uniform distributions
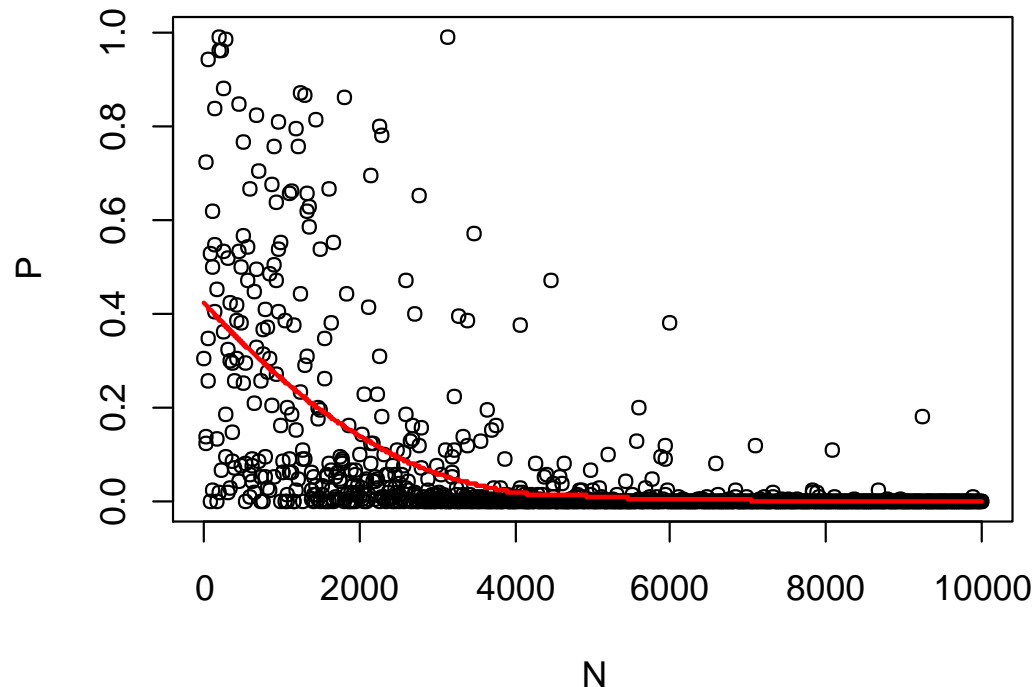

```
N = 10 * c(1:1000)
P = numeric(1000)

for(i in 1:1000){
  t = t.test(sample(0:100, N[i], rep=T),
             sample(2:102, N[i], rep=T))
  P[i] = t$p.value
}
plot(N, P)
lo = loess(P~N)
x1 = seq(min(N), max(N),
         (max(N) - min(N))/1000)
lines(x1, predict(lo,x1), col = 'red', lwd=2)
```

You need to know

- How to turn a question into hypotheses

- Every test has assumptions

  - A statistician can check all the assumptions
  - If the data does not meet the assumptions there are non-parametric versions of the tests

Common mistakes in hypothesis testing

- Lack of independence

- Violation of normality
  - Highly skewed data

- Assume equal variances and the variances are not equal
  (Did not do variance test)

ALWAYS graph your data first to assess symmetry and variance

Exercise

Which test to use?

Example 1: A scientist takes 10 measurements of downstream contamination levels in a river at only one point each time. She is interested in whether the fish have a higher level of PCB's than the known average of the upstream level, which is 8 ppb.

Hypothesis?

H_0: μ_{down} is <= 8 ppb
H_1: μ_{down} is > 8 ppb

Which test to use?

Measurement #	Downstream
1	10
2	12
3	6
4	9
5	15
6	8
7	4
8	9
9	11
10	7

Can only use one sample t-test. Right-handed t-test

Is the sample large enough to have a Power of 0.8 at an effect size of 1 ppb for alpha = .05?

Did we take enough measurements to correctly reject the null 80% of the time when the true mean is 9 for apha=.05?

Which test to use?

Example 2:

Comparing contamination levels upstream and downstream of toxic waste facility using measurements taken on a single day

Measurement #	Downstream	Upstream
1	10	6
2	12	10
3	6	8
4	9	7
5	15	9
6	8	7
7	4	3
8	9	9
9	11	9
10	7	10
	9.1	7.8

Mean of Downstream cont. level = 9.1

Mean of Upstream cont. level= 7.8

Which test to use?

     Two-sample?

Significance level?

Which test to use?

Example 2:

Comparing contamination levels upstream and downstream of toxic waste facility using the 10 measurements

Measurements #	Downstream	Upstream	Difference
1	10	6	4
2	12	10	2
3	6	8	-2
4	9	7	2
5	15	9	6
6	8	7	1
7	4	3	1
8	9	9	0
9	11	9	2
10	7	10	-3

Which test to use?

Paired!!

Significance level?

Assignment

General objectives: calculate the power of a t test

You develop your data set (e.g. weight of 30 rats), provide a brief introduction to the data set, formally state the hypotheses that you are going to test (e.g. Ho = 20g, and Ha=22g).

Check the normality of your data; calculate the standard deviation. Set the alpha level to be 0.05, calculate the power.

Indicate in your results and discussion section what you found, i.e. did you reject your null, and the conclusions that you have drawn from the analysis.

R script

```r
sample = rnorm(30) # You'd better have your own data
m = mean(sample)
s = sd(sample)
n = length(sample)

diff <- qt(0.975, df = n-1)*s/sqrt(n)
left <- m-diff ; right <- m+diff

assumed <- m + 2 # difference between H0 and Ha is 2
tleft  <- (assumed - right)/(s/sqrt(n))
p <- pt(-tleft, df = n-1)
power = 1-p
```