Hypothesis testing

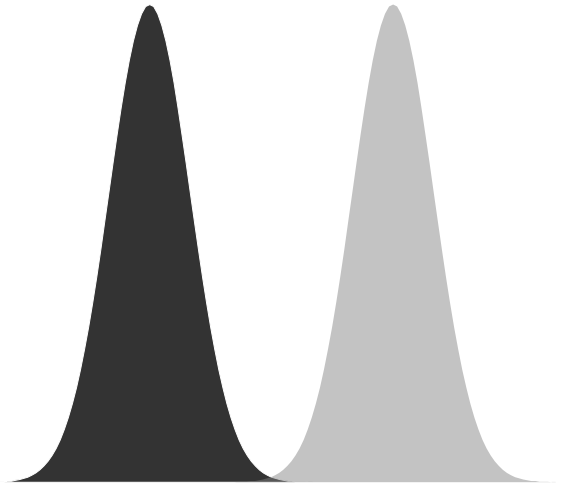
- Effect size
- Power of test
- Sample size
- Philosophy of hypothesis testing
- Coding convention of R

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

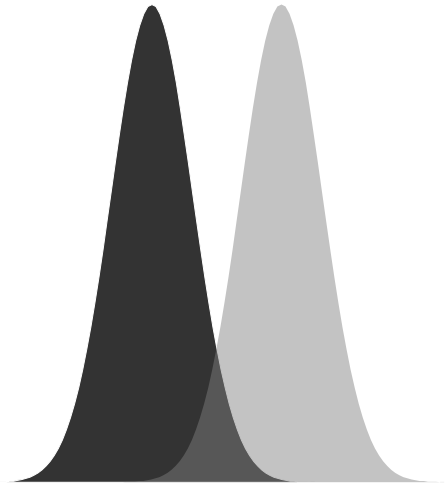
- The standardized difference
- Doesn't depend on the size of the sample
- Difference between the mean of your sample and the mean of the population if the null were true, divided by the standard deviation of the population

Effect size

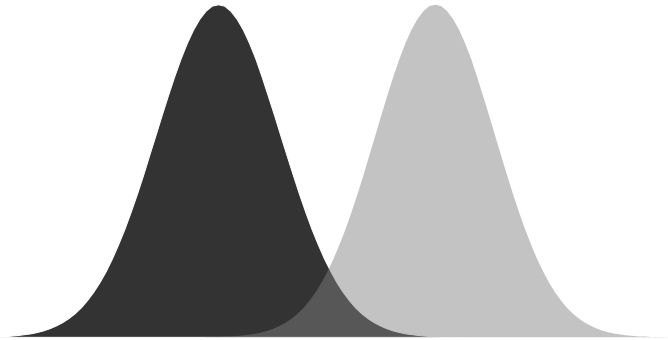
$$d = \frac{\bar{x} - \mu}{s_{pooled}}$$



Big effect size



Small effect size



Small effect size

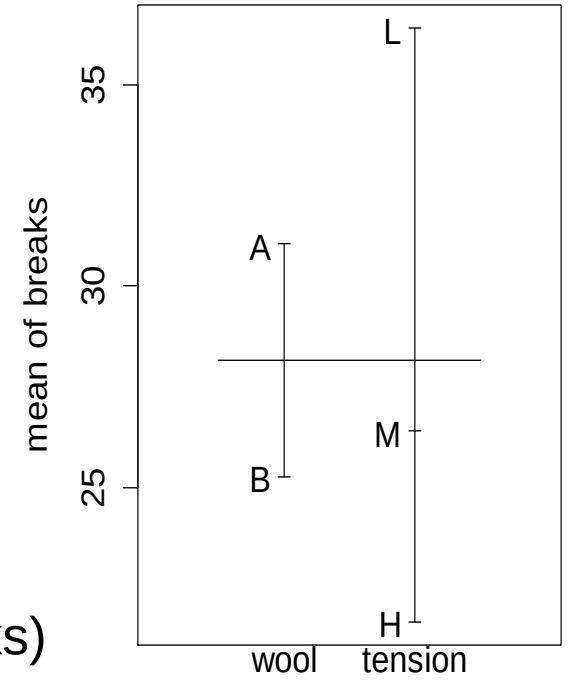
R script

```
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)
```



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.

Remember - Type I and Type II Errors

		NULL HYPOTHESIS	
		TRUE	FALSE
DECISION	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

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.

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

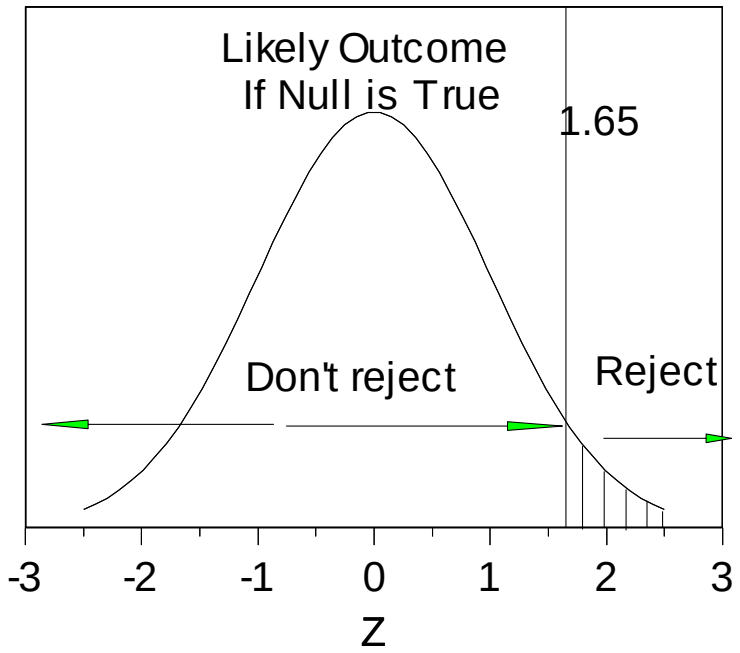
$$Bound = 138 + 1.65\sigma_M = 141.3$$

$$\alpha = p(reject\ H_0 \mid \mu = 138)$$

$$\alpha = p(reject\ H_0 \mid H_0) = .05$$

$$\beta = p(accept\ H_0 \mid \mu = 142)$$

$$\beta = p(accept\ H_0 \mid H_1)$$

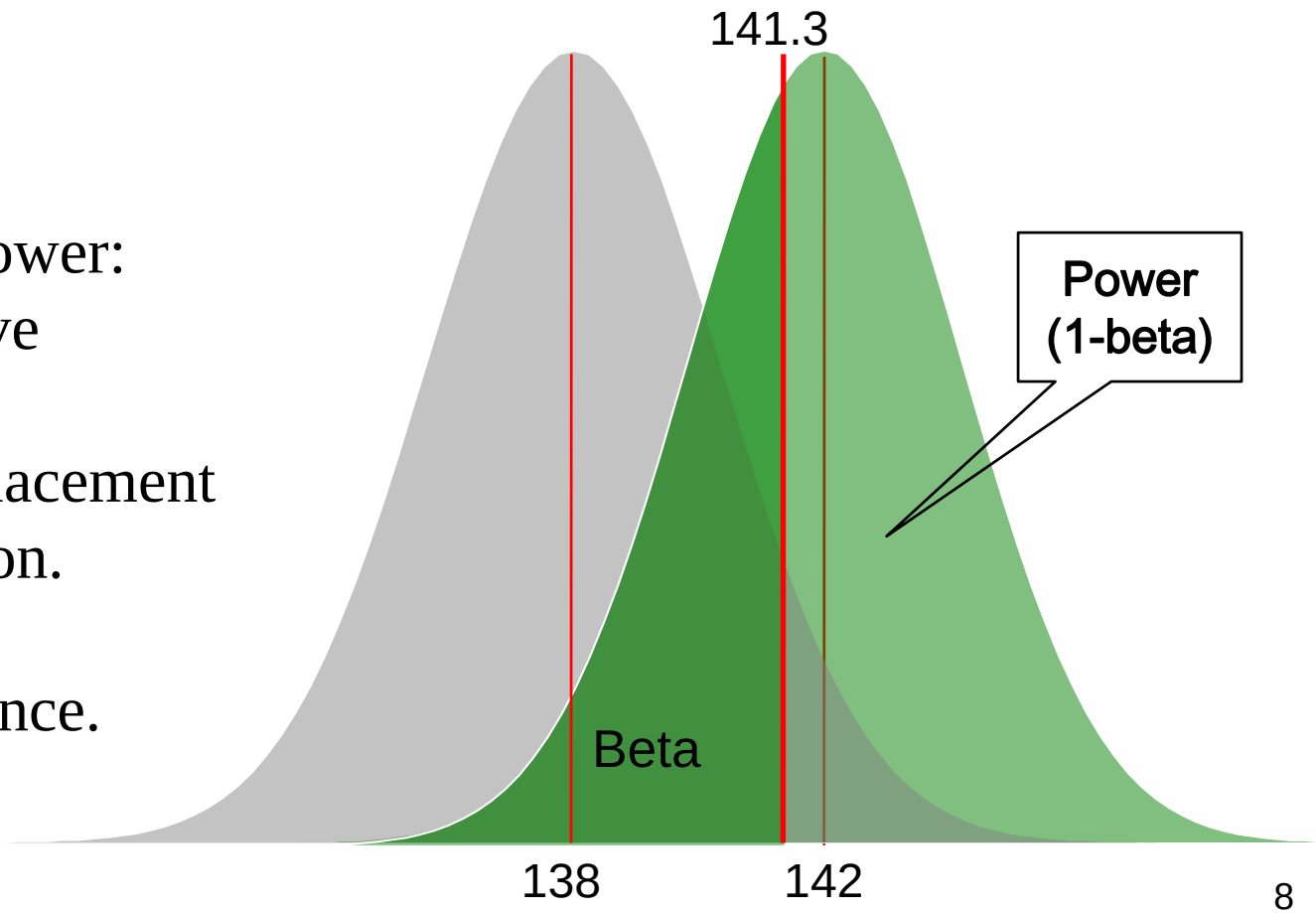


Power

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.

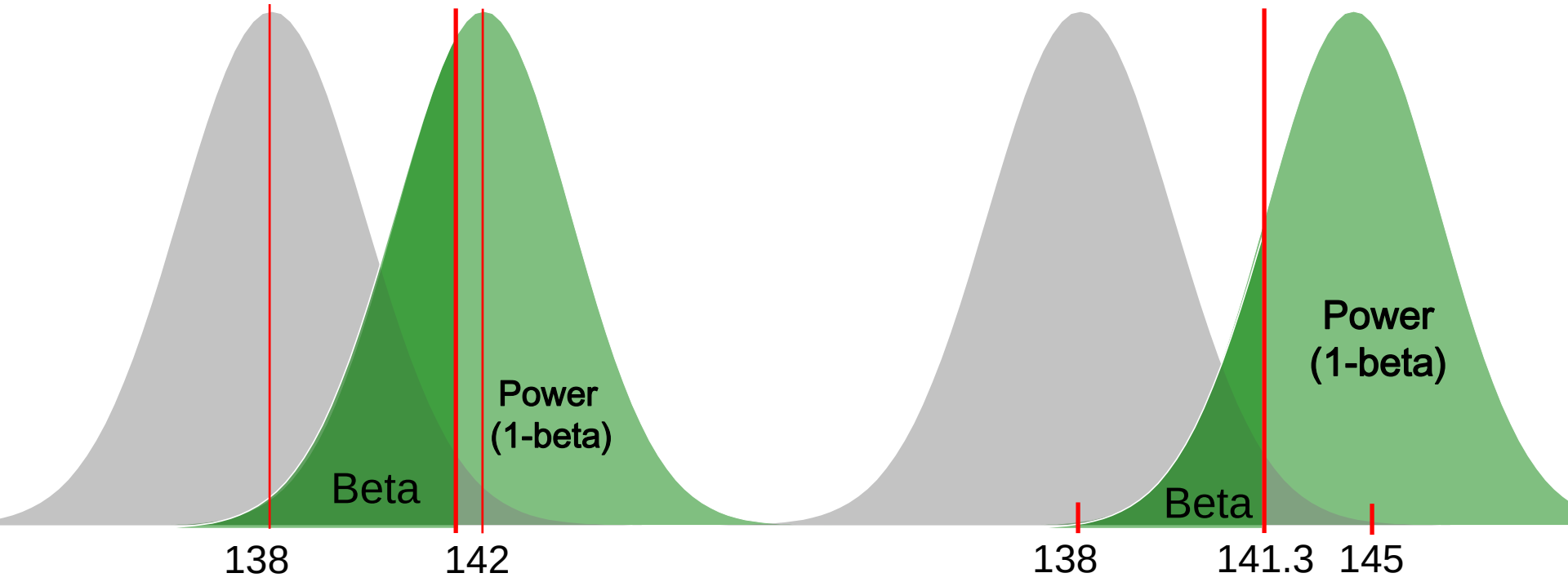
Four things affect power:

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



Power

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

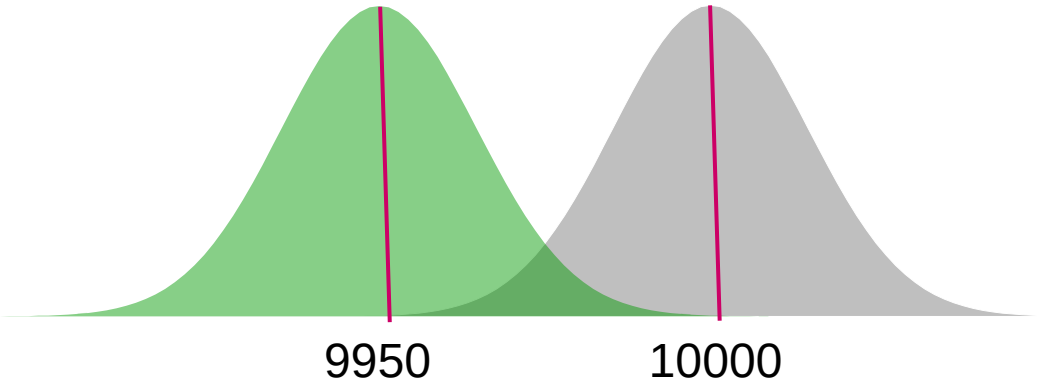


Example

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

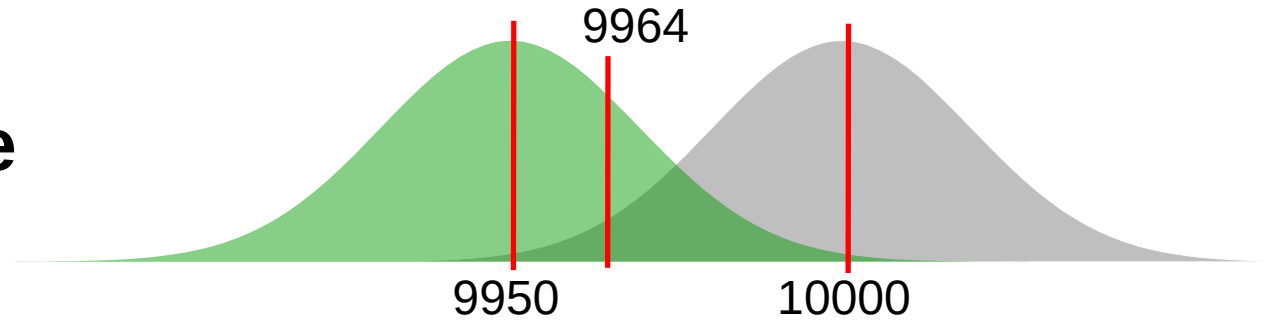
Assume actual mean light bulb lifetime is 9,950 hours and the population standard deviation is 120 hours.

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



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

Example



Solution

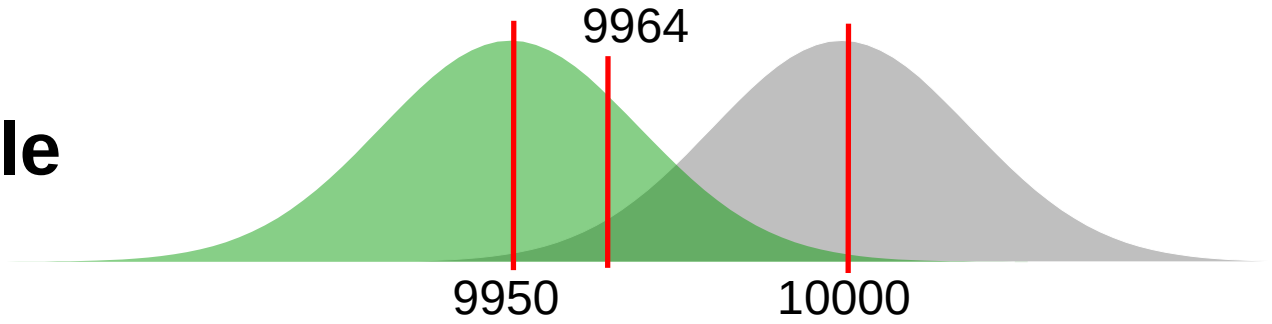
The null hypothesis is that $\mu \geq 10000$. We begin with computing the test statistic.

```
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.

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

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%.

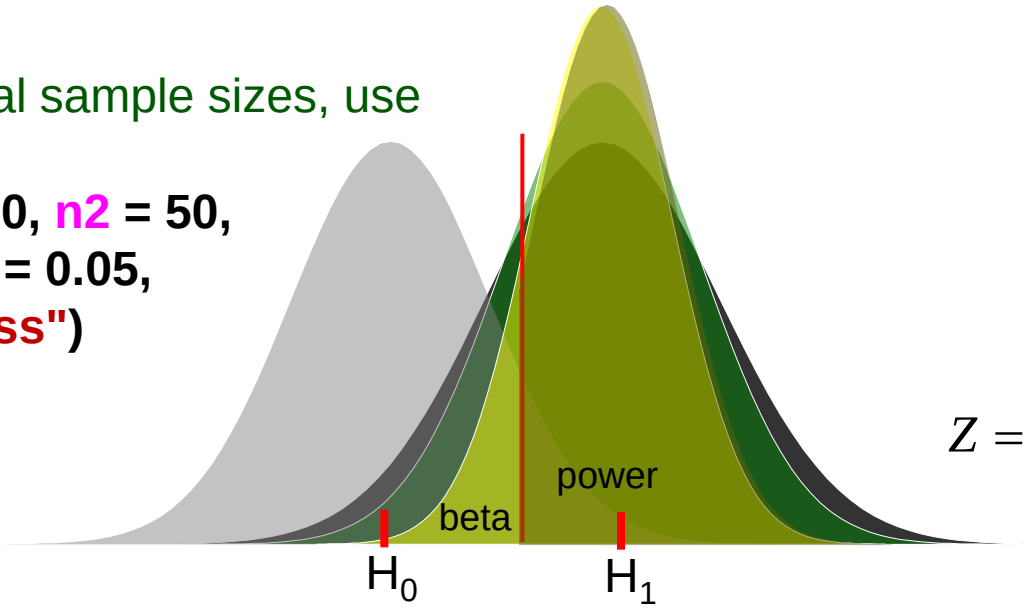
http://www.r-tutor.com/elementary-statistics/hypothesis-testing/lower-tail-test-population-mean-unknown-variance

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 the effect size, 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")
```



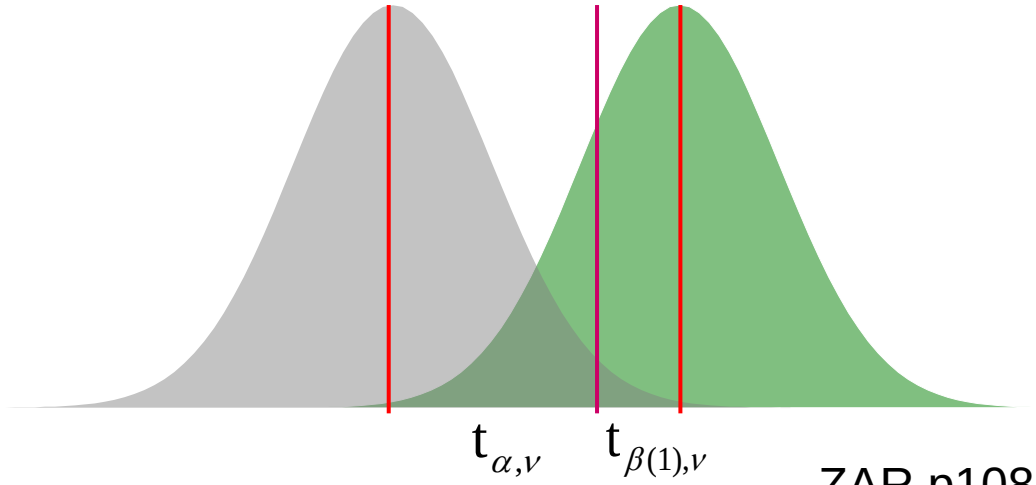
$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}$$

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

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



ZAR p108

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

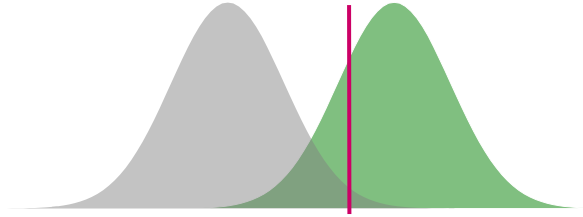
What is the probability of detecting a true difference (i.e. a difference between μ and μ_0) of at least 1.0 g for the experiment of Example 7.2?

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

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

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

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

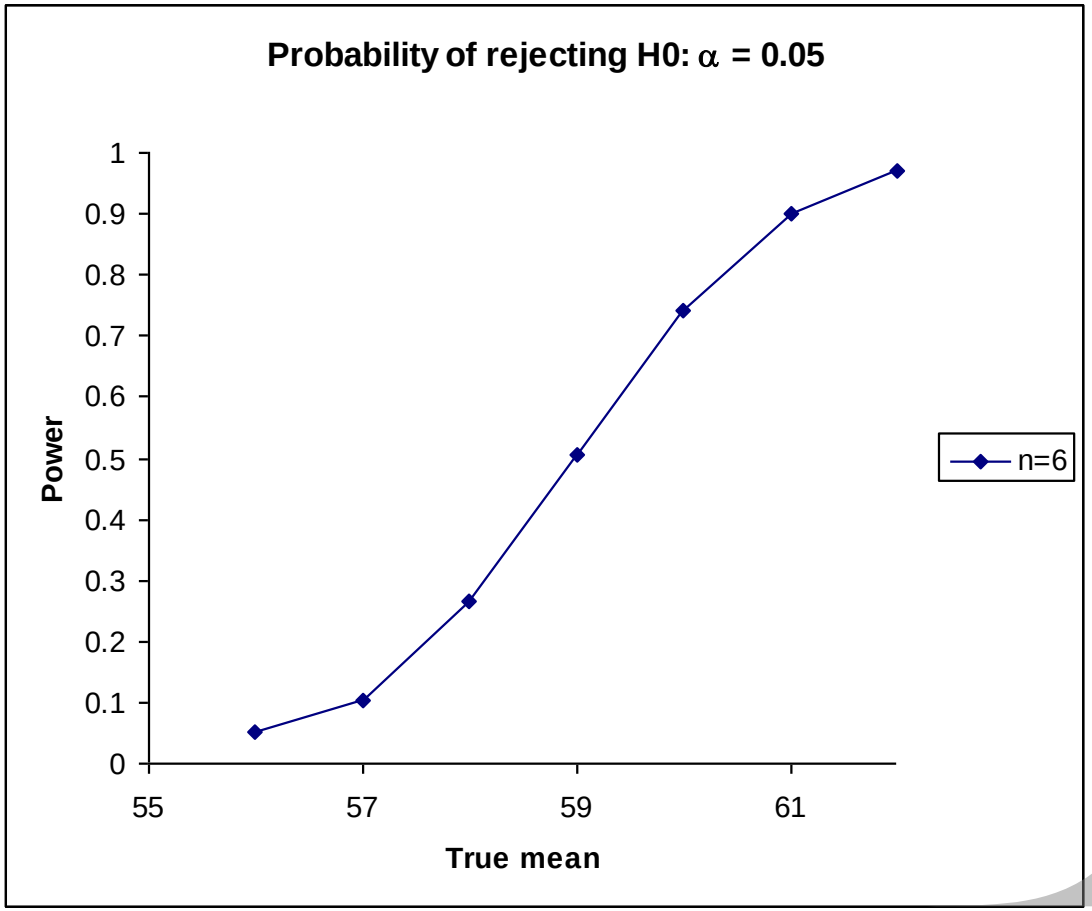
Traditionally accepted power

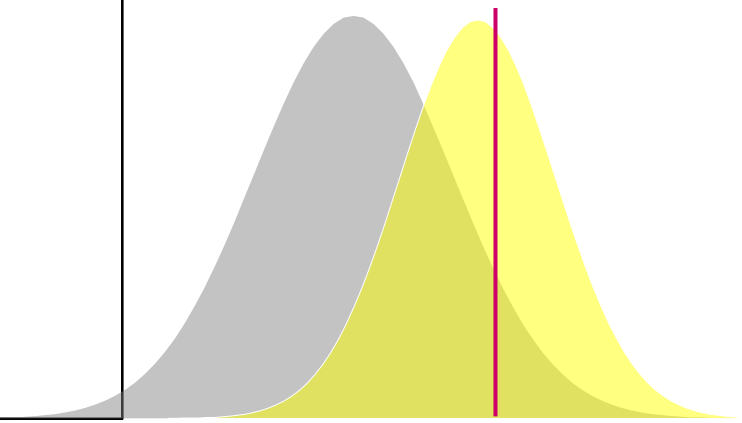
- 80% is the commonly used standard for power of test.
- It is below 95%. It is a conservative standard.
- Scientists want to increase power to detect the difference, in order to stop some production due to its significant negative impact on environment. People in industry want to decrease power, so as to keep their benefit by ignoring the impact.

Power and the effect size

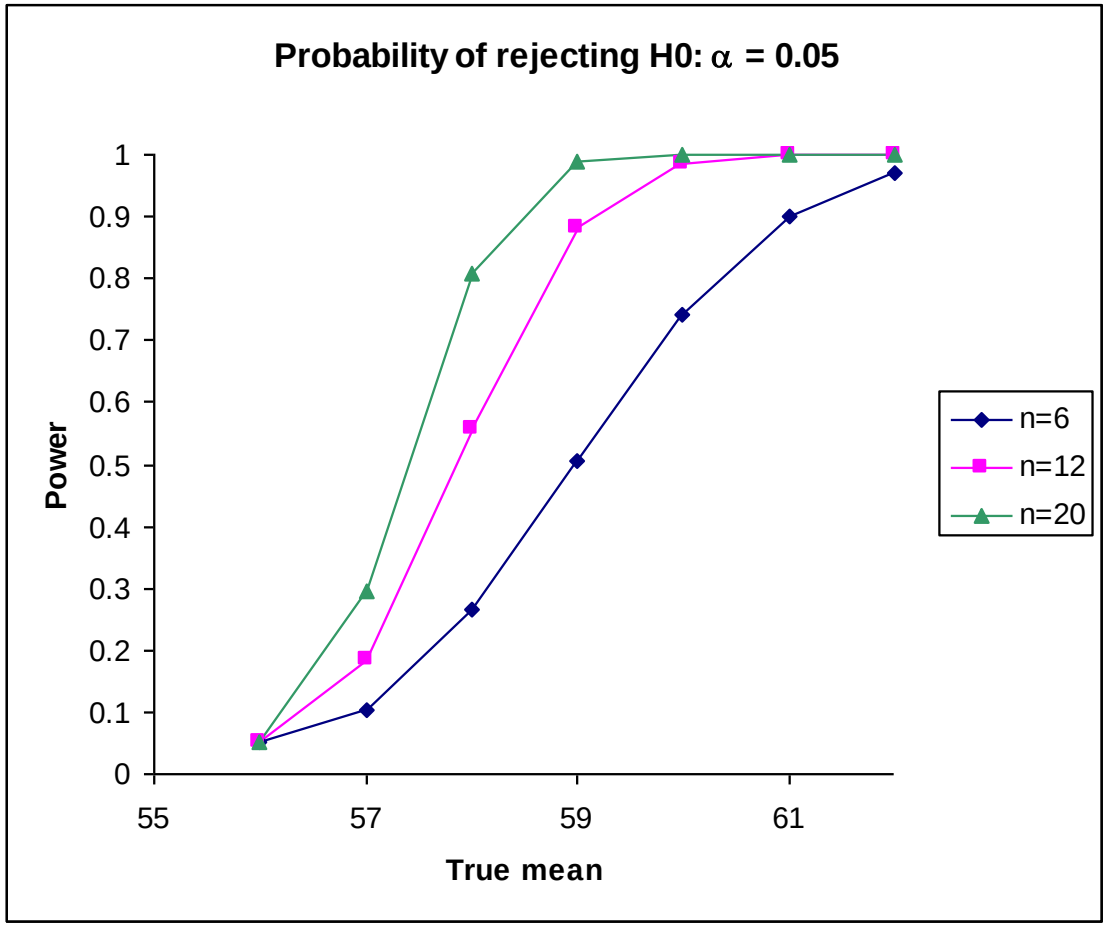
$\sigma = 3$

$H_0: \mu \leq 56$



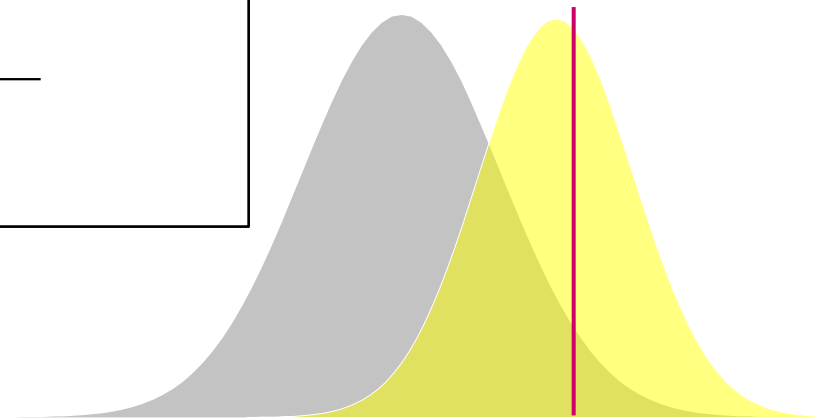


Effect of sample size on power

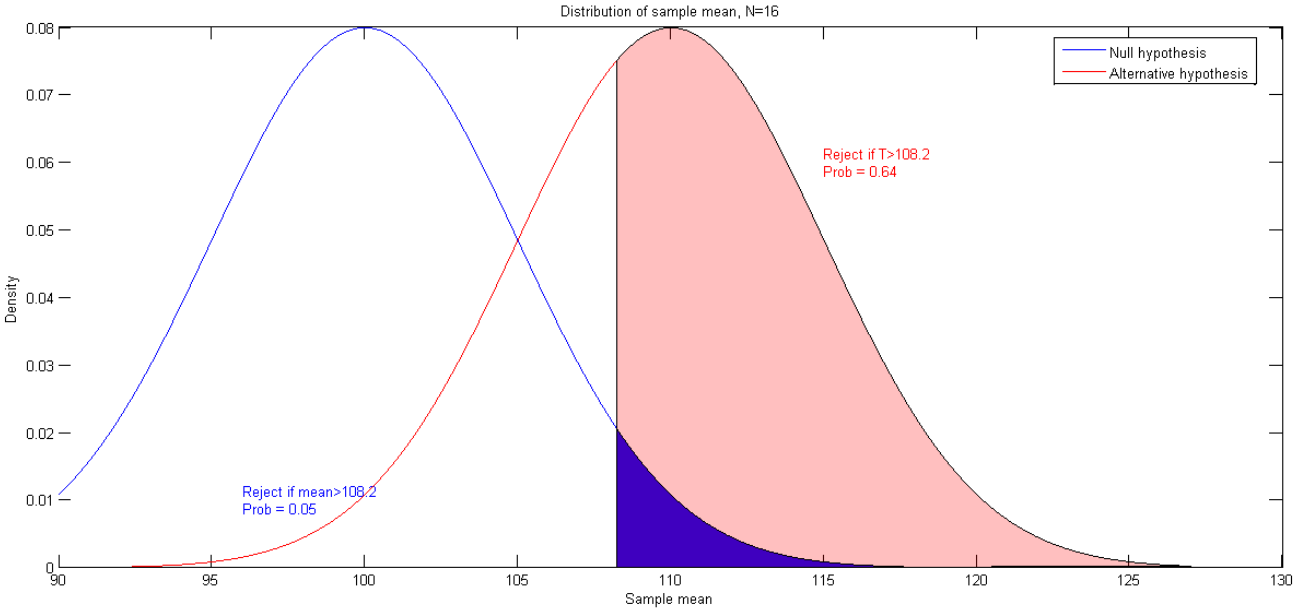


$\sigma = 3$

$H_0: \mu \leq 56$



Sample size *n* and beta



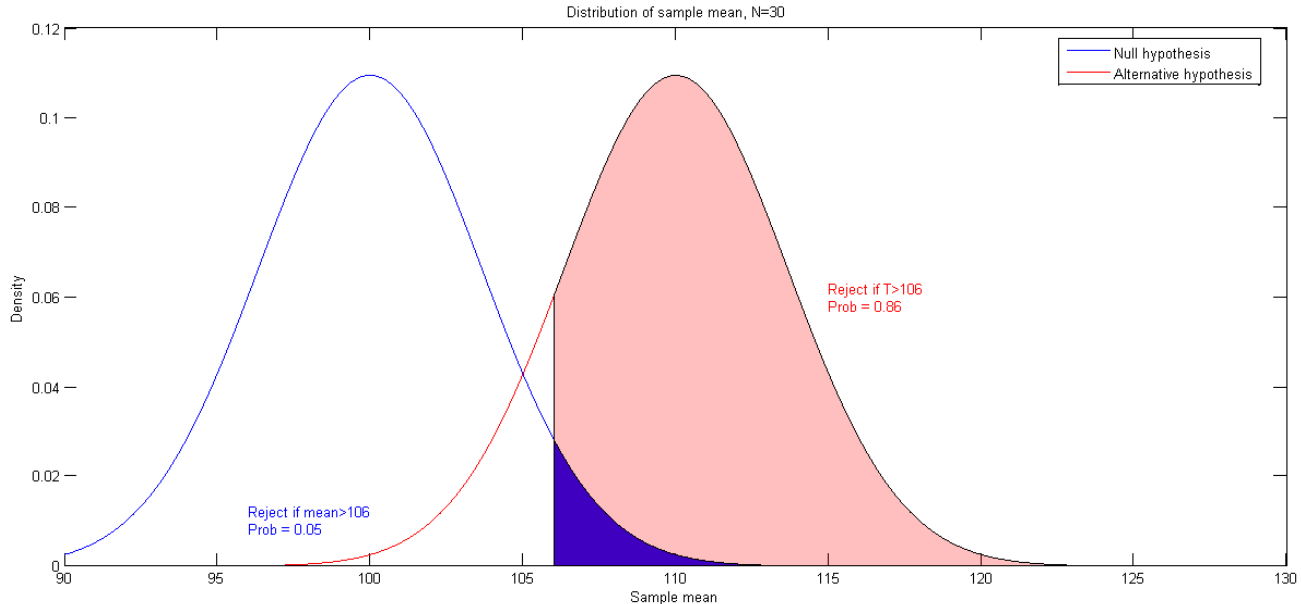
H_0: mean=100

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).



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))

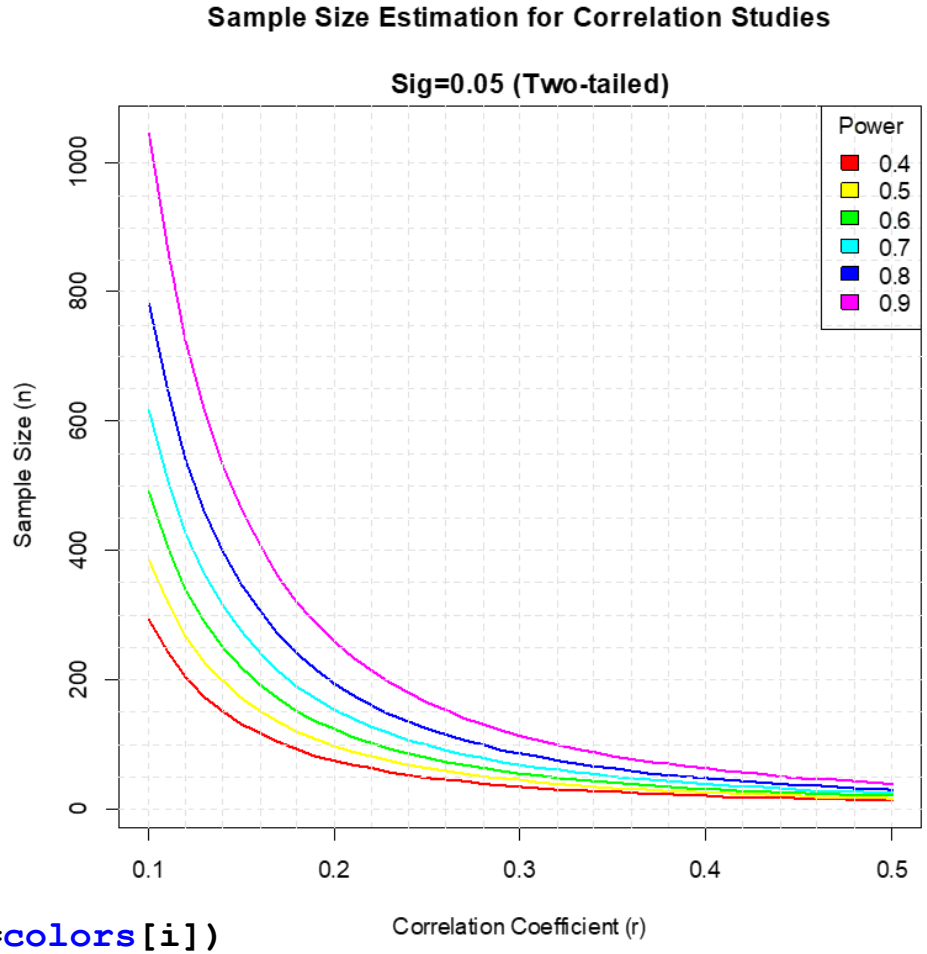
pnorm(106, mean=110, sd=20/sqrt(30), lower.tail = F)

Power

```
# 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




```
# 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)
```

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:

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

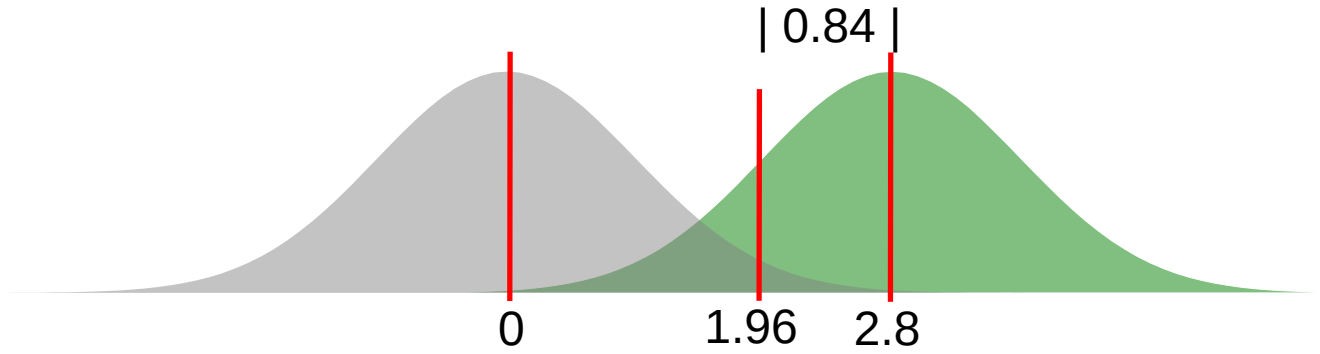
$$t_{\beta(1),\nu} = \frac{\delta}{\sqrt{\frac{s^2}{n}}} - t_{\alpha,\nu}$$

Example - Rosiglitazone vs. Placebo

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

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

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



Source: Carr, et al (2004)

ZAR p106

Sample size – one sample case

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

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 – α = 0.95, and α = 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 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.

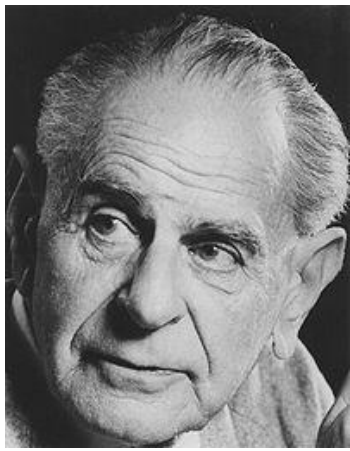
Philosophy of hypothesis testing

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.



http://en.wikipedia.org/wiki/Karl_Popper

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.

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

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

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

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 $\alpha = 0.05$.

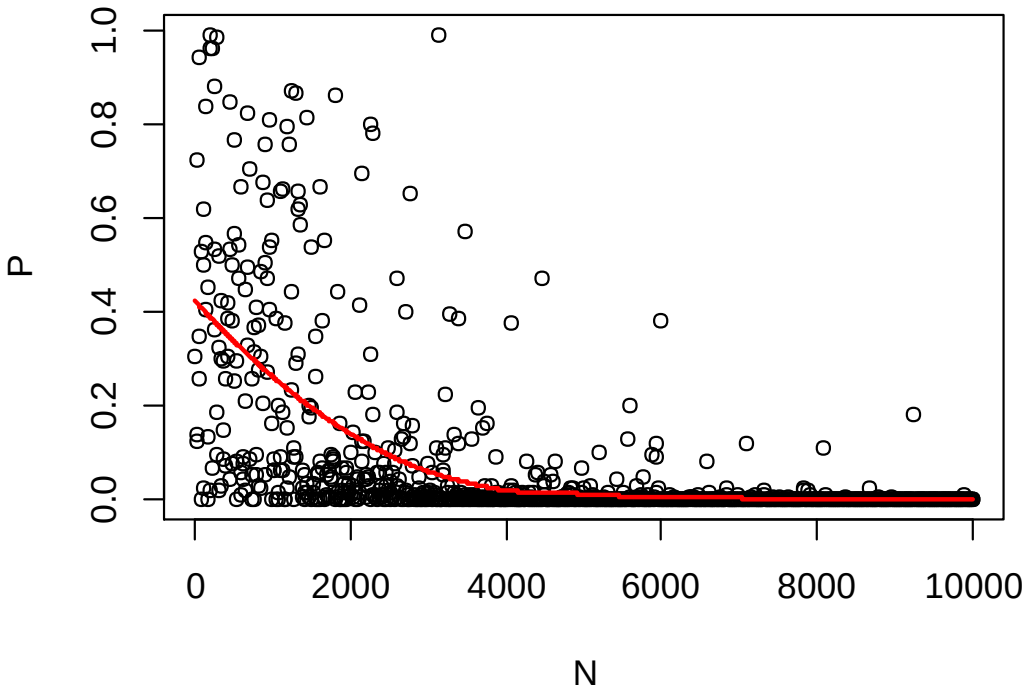
The *p* values should be reported.

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

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.

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

Hypothesis?

H_0: μ_{down} is <= 8 ppb

H_1: μ_{down} is > 8 ppb

Which test to use?

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?

Coding convention of R

Notation and Naming

File Names

File names should end in .R and, of course, be meaningful.

GOOD: **predict_ad_revenue.R**

BAD: foo.R

Identifiers

Don't use underscores (_) or hyphens (-) in identifiers. Identifiers should be named according to the following conventions. Variable names should have all lower case letters and words separated with dots (.); function names have initial capital letters and no dots (CapWords); constants are named like functions but with an initial k.

variable.name

GOOD: **avg.clicks**

BAD: avg_Clicks , avgClicks

FunctionName

GOOD: **CalculateAvgClicks**

BAD: calculate_avg_clicks , calculateAvgClicks

Make function names verbs.

kConstantName (e.g. kCarryingCapacity)

Coding convention of R

Indentation

When indenting your code, use two spaces. Never use tabs or mix tabs and spaces.

```
for(count in seq(0,10000,by=1)) {
  code line
  code line
}
```

Spacing

Place spaces around all binary operators (=, +, -, <-, etc.).

Do not place a space before a comma, but always place one after a comma.

GOOD:

```
tabPrior <- table(df[df$daysFromOpt < 0, "campaignid"])
total <- sum(x[, 1])
total <- sum(x[1, ])
```

BAD:

```
tabPrior <- table(df[df$daysFromOpt<0, "campaignid"])  # Needs spaces around '<'
tabPrior <- table(df[df$daysFromOpt < 0,"campaignid"])  # Needs a space after the comma
tabPrior<- table(df[df$daysFromOpt < 0, "campaignid"])  # Needs a space before <-
tabPrior<-table(df[df$daysFromOpt < 0, "campaignid"])  # Needs spaces around <-
total <- sum(x[,1])  # Needs a space after the comma
total <- sum(x[ ,1])  # Needs a space after the comma, not before
```

Coding convention of R

Extra spacing (i.e., more than one space in a row) is okay if it improves alignment of equals signs or arrows (<-).

```
plot(x     = xCoord,
     y     = dataMat[, makeColName(metric, ptiles[1], "roiOpt")],
     ylim  = ylim,
     xlab  = "dates",
     ylab  = metric,
     main = (paste(metric, " for 3 samples ", sep="")))
```

Coding convention of R

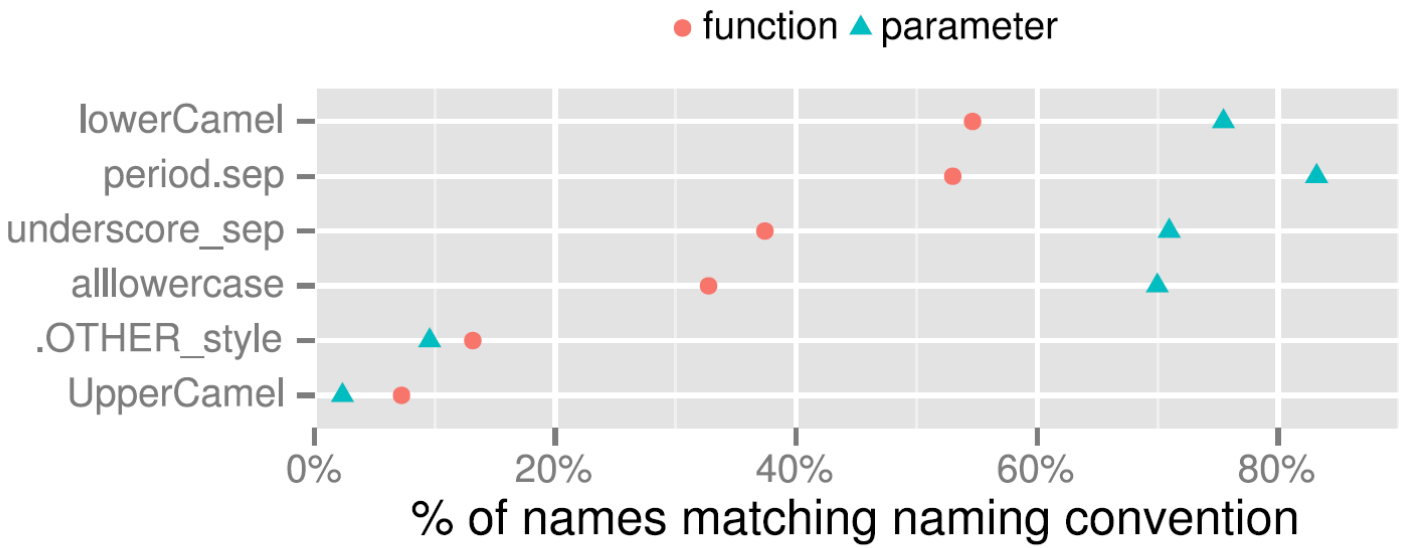


Figure 1: The percentage of function and parameter names from CRAN that matches the five naming conventions.

Bååth, Rasmus. 2012. **The State of Naming Conventions in R.** The R Journal Vol. 4/2, 74-75

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

```
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
```