STA305 / STA1004

Mark Ebden, 21–24 January 2019, Week 3

With grateful acknowledgment to Nathan Taback et al.

Aesop's Fable #210: The Boy Who Cried Wolf

This Week

- Example 3: Clinical Trials
  - Power of Statistical Tests
  - Sample-size calculations
- In-class work, on Questions 1, 2, and 3
  - The default setup is to work in pairs
  - See Section 8 of the Unit 3 Lecture notes

Course sign-posting

- Unit 1:
  - Introduced the three principles of this course:
  - Randomization, blocking, and replication
- Unit 2:
  - The **wheat-yield** example explored randomization
  - The **boys' shoes** example explored blocking
- Unit 3:
  - **Clinical trials** examples will explore replication
  - Two common problems: $n \rightarrow (1 - \beta)$, and $(1 - \beta) \rightarrow n$

What are clinical trials?

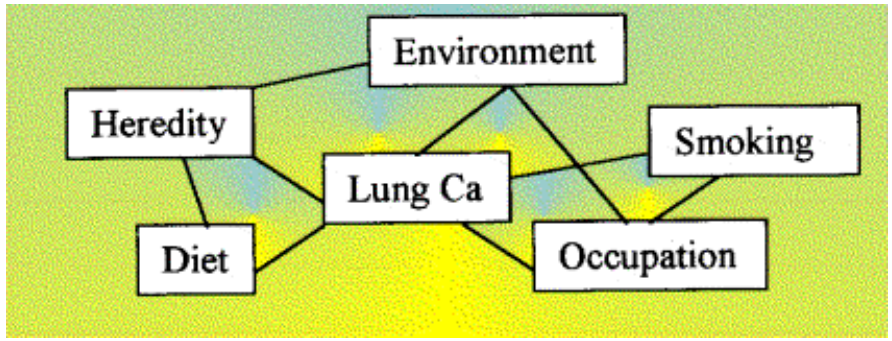
Clinical trials are prospective intervention studies with human subjects to investigate experimental drugs, new treatments, medical devices, or clinical procedures (G. Yin, *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*, 2012).

A couple terms to keep in mind:

- All experimental units receiving a given treatment = *arm*
- Assignment mechanism = *allocation* scheme

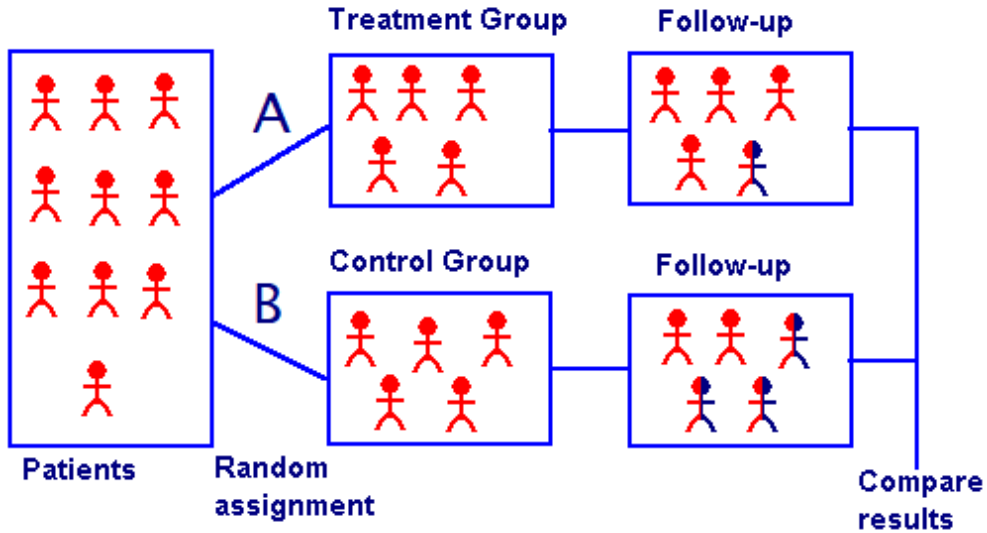
Some optional material is available in the appendix, from slide 75 onwards.

How can causation be assessed using a randomized design?



- Suppose that patients are randomized in a two-arm clinical trial in which one of the arms is the standard treatment and the other arm is an experimental treatment. The outcome is 5-month mortality.
- Suppose further that a statistically significant difference in the outcome between the two arms is observed, showing that the experimental treatment is more efficacious.
- The interpretation is that the experimental treatment *caused* patients to have a better outcome since the only difference between the two arms is the treatment. Randomization is supposed to ensure that the groups will be similar with respect to all the factors measured in the study and all the factors that are not measured. (Recall last week's NSW table, slide 12.)

Applying statistical hypotheses



Suppose that subjects are randomized to treatments A or B with equal probability. Let μ_A be the mean response in the group receiving drug A and μ_B be the mean response in the group receiving drug B. The null hypothesis is that there is no difference between A and B; the alternative claims there is a clinically meaningful difference between them.

$$H_0 : \mu_A = \mu_B \text{ versus } H_1 : \mu_A \neq \mu_B$$

We want to know if the standard treatment is better than the experimental treatment, or vice versa.

Statistical hypotheses

Recall that the type I error rate is defined as:

$$\begin{aligned}\alpha &= P(\text{type I error}) \\ &= P(\text{Reject } H_0 \mid H_0 \text{ is true})\end{aligned}$$

The type II error rate is defined as:

$$\begin{aligned}\beta &= P(\text{type II error}) \\ &= P(\text{Don't reject } H_0 \mid H_1 \text{ is true})\end{aligned}$$

Statistical hypotheses and power

Power (a.k.a. sensitivity) is defined as:

$$
\begin{aligned}
&= 1 - P(\text{Don't reject } H_0 \mid H_1 \text{ is true}) \\
&= P(\text{Reject } H_0 \mid H_1 \text{ is true})
\end{aligned}
$$

The probability that a fixed-level α test will reject H_0 when a particular alternative value of the parameter is true is called the *power* of the test to detect that alternative.

Why is Power Important in Phase III Clinical Trials?

- If a new treatment is to be used in patients then it should be compared to the standard treatment
- Evidence is required that the new treatment is effective and safe. The form of the evidence is the result of a hypothesis test
- Will the hypothesis test reject if a difference between the treatments really exists?
- **High power** will ensure that if a difference exists then the hypothesis test will have a high probability of rejecting H_0
- The most practical way to ensure that the test is powerful is to enrol enough patients in each arm of the trial (so we need to think about *sample size*)

How many patients should be enrolled in a Phase III clinical trial?

- In a phase III trial, **sample size** (n) is the most critical component of the study design. The sample size has implications for how many subjects will be exposed to a drug that has yet-to-be-confirmed efficacy
- Standard practice is to compute the smallest sample size required to detect a clinically important/significant treatment difference with sufficient power
- The investigator needs to specify a type I error rate, type II error rate, and the effect size (smallest difference that is expected between two groups)

How many patients should be enrolled in a Phase III clinical trial?

- If the sample size is too small then the trial might fail to discover a truly effective drug because the statistical test cannot reach the significance level (5%) due to a lack of power
- If the sample size is overestimated then resources are wasted and drug development is delayed since patient enrollment is often the main factor in time to complete a trial

Sample Size and Power in Phase III Clinical Trials

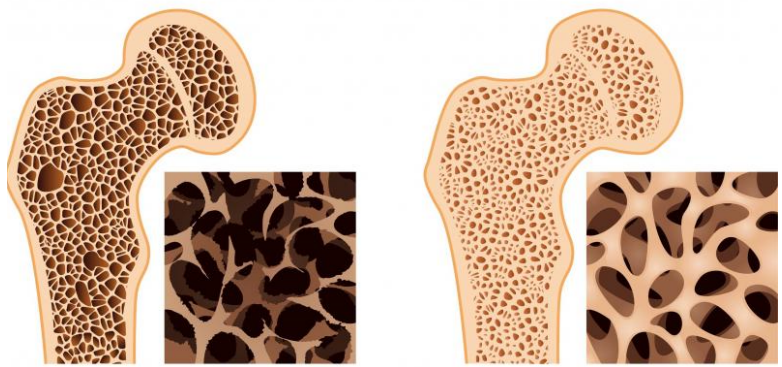
- The sample size is calculated under the alternative hypothesis based on the type I error rate α and power $1 - \beta$
- Specify a clinically meaningful difference that is to be detected at the conclusion of the trial
- Intuitively, if a small difference (effect size) is expected between the two treatments in comparison, a large sample size would be required, and vice versa. Why?

- Sample size also depends on the variance
- The larger the variance, the harder it is to detect the difference and thus a larger sample size is needed

Example

You're curious as to whether a 6-month exercise program can increase the total body bone-mineral content (TBBMC) of young women. A change in TBBMC of 1% would be considered important. Is 36 subjects a reasonably large sample size to detect such a change at $\alpha = 0.05$?

- Based on the results of a previous study, an individual's percent change in TBBMC over the 6-month period follows a $\mu = 0$, $\sigma = 2$ distribution that's roughly symmetric with rapidly dying tails
- Assume that the exercise program would affect only μ, if anything



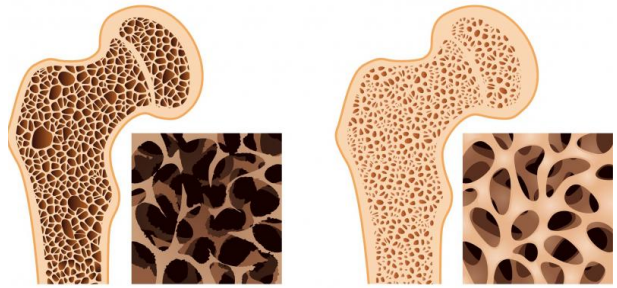
Hints: What kind of significance test would you use? Draw a picture. Calculate power to see whether it's high enough.

R commands for calculating power or sample size

Two common problems: $n \rightarrow (1-\beta)$, and $(1-\beta) \rightarrow n$.

Procedure	Conditions	Function
Calculating power:		
One-sample z-test: testing a mean, when sigma is known	Equal allocation	`pow.z.test`
One-sample t-test: testing a mean, when sigma is unknown	Equal allocation	`power.t.test(type = "one.sample")`, `onesampttestpow`
Two-sample t-test: testing a difference in means	Equal allocation	`power.t.test, twosampttestpow`
Any test in which the test statistic has known distribution	Equal or unequal allocation	`replicate(N,t.test(...)$p.value)`
Calculating sample size:		
Experiment with continuous outcomes	Equal allocation	`power.t.test, size2z.uneq.test(r=1)`
Experiment with continuous outcomes	Unequal allocation	`size2z.uneq.test`
Experiment with binary outcomes	Equal allocation	`power.prop.test`
Experiment with binary outcomes	Unequal allocation	`replicate(N,prop.test(...)$p.value)`

Power of the one-sample z-test



Consider the TBBMC example.

Let $X_1, ..., X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution with known σ. A test of the hypothesis

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

will reject at level α if and only if

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \geq z_{\alpha/2},$$

or

$$\left| \bar{X} - \mu_0 \right| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2},$$

where $z_{\alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the $\mathcal{N}(0, 1)$.

If you need to, review CLT (Week 1, slides 65f).

Power of the one-sample z-test

The power of the test at $\mu = \mu_1$ is

$$\begin{aligned} 1 - \beta &= 1 - P(\text{type II error}) \\ &= P(\text{Reject } H_0 \mid H_1 \text{ is true}) \\ &= P(\text{Reject } H_0 \mid \mu = \mu_1) \\ &= P\left(\left| \bar{X} - \mu_0 \right| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \;\middle|\; \mu = \mu_1 \right) \end{aligned}$$

Subtract the mean μ_1 and divide by $\sigma/\sqrt{n}$ to obtain:

$$1 - \beta = 1 - \Phi\left(z_{\alpha/2} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \right) + \Phi\left(-z_{\alpha/2} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \right),$$

where $\Phi(\cdot)$ is the $\mathcal{N}(0, 1)$ CDF.

Power of the one-sample z-test

The power function of the one-sample z-test is:

$$1 - \beta = 1 - \Phi\left(z_{\alpha/2} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) + \Phi\left(-z_{\alpha/2} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$$

What is the limit of the power function as:

- $n \to \infty$
- $\mu_1 \to \mu_0$
- $\sigma \to 0$

Which two are easiest to control?

Power of the one-sample z-test

The power function for a one-sample z-test can be calculated using R.

$$1-\beta = 1-\Phi\left(z_{\alpha/2}-\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right)+\Phi\left(-z_{\alpha/2}-\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right).$$

```
pow.z.test <- function(alpha,mu1,mu0,sigma,n){
  arg1 <- qnorm(1-alpha/2)-(mu1-mu0)/(sigma/sqrt(n))
  arg2 <- -1*qnorm(1-alpha/2)-(mu1-mu0)/(sigma/sqrt(n))
  1-pnorm(arg1)+pnorm(arg2)
}
```

NB: `arg1` is for the first Φ, `arg2` the second.

Power of the one-sample z-test

For example consider

$$H_0 : \mu = 0 \text{ versus } H_1 : \mu \neq 0$$

with $n = 30, \sigma = 0.2, \alpha = 0.05$. The power at $\mu_1 = 0.15$ can be calculated by calling the above function.

```
pow.z.test(.05,.15,0,.2,30)
```

```
[1] 0.9841413
```

What does this mean?

Poll question

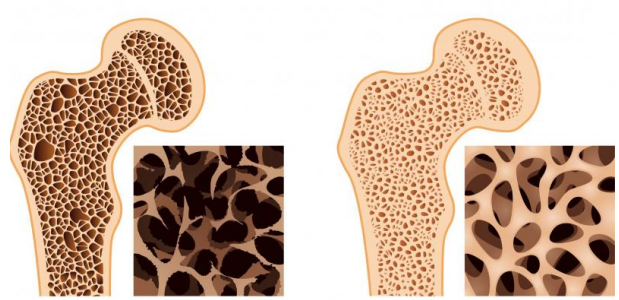
Question: $H_0 : \mu = 0$, $H_1 : \mu \neq 0$. At $\mu_1 = 0.15$, what does power $= 0.98$ mean?

Answer: In studies for which the true mean is...

- A: ... 0.00, the test will reject H_0 2% of the time
- B: ... 0.15, the test will reject H_0 98% of the time
- C: ... 0.15, the test will fail to reject H_0 98% of the time
- D: ... 0.00, the test will fail to reject H_0 2% of the time
- E: *I don't know*

To vote: visit pollev.com/loop

Power of the one-sample t-test



What if we didn't know σ?

Let $X_1, ..., X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution with unknown σ. A test of the hypothesis

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

will reject at level α if and only if

$$\left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| \geq t_{n-1,\, \alpha/2}$$

where $t_{n-1,\, \alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the t_{n-1}.

If you need to, review the t-distribution (Week 1, slides 70f).

Power of the one-sample t-test

The lefthand side can be rewritten as

$$\sqrt{n}\left[\frac{\bar{X}-\mu_0}{S}\right] = \frac{Z+\gamma}{\sqrt{V/(n-1)}}$$

where

$$Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$$

$$\gamma = \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}$$

$$V = \frac{(n-1)}{\sigma^2}S^2$$

Note that $Z \sim \mathcal{N}(0,1)$, and $V \sim \chi^2_{n-1}$, and that Z is independent of V.

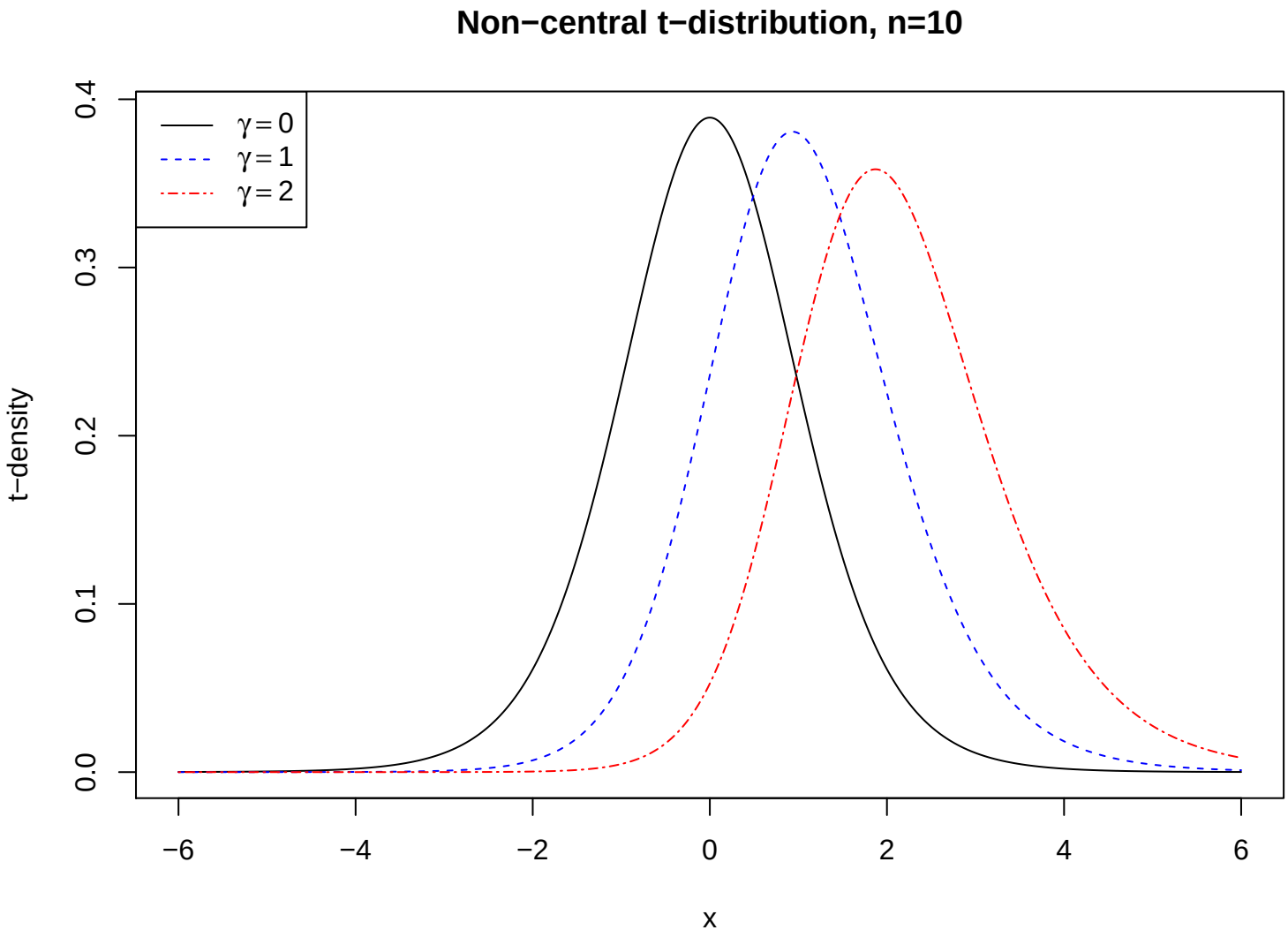
The parameter $\gamma = 0$ under H_0, and $\gamma \neq 0$ under H_1.

Power of the one-sample t-test

- If $\gamma = 0$ then $\sqrt{n}\left[\frac{\bar{X}-\mu_0}{S}\right] \sim t_{n-1}$. This is sometimes called the **central *t*-distribution**
- If $\gamma \neq 0$ (occurs under H_1) then $\sqrt{n}\left[\frac{\bar{X}-\mu_0}{S}\right] \sim t_{n-1,\gamma}$, where $t_{n-1,\gamma}$ is the **non-central *t*-distribution** with non-centrality parameter γ

Power of the one-sample t-test

A plot of the central t-distribution ($\gamma = 0$) and non-central t-distribution ($\gamma = 1, 2$) are shown in the graph below.



Power of the one-sample t-test

The power of the test at $\mu = \mu_1$ is

$$
\begin{aligned}
1 - \beta &= 1 - P(\text{type II error}) \\
&= P(\text{Reject } H_0 \mid H_1 \text{ is true}) \\
&= P(\text{Reject } H_0 \mid \mu = \mu_1) \\
&= P\left(\left| \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \right| \geq t_{n-1,\,\alpha/2} \mid \mu = \mu_1 \right) \\
&= P(t_{n-1,\gamma} \geq t_{n-1,\,\alpha/2}) + P(t_{n-1,\gamma} < -t_{n-1,\,\alpha/2})
\end{aligned}
$$

Power of the one-sample t-test

$$P(t_{n-1,\gamma} \geq t_{n-1,\alpha/2}) + P(t_{n-1,\gamma} < -t_{n-1,\alpha/2})$$

The following function calculates the power function for the one-sample t-test in R:

```
onesampttestpow <- function(alpha,n,mu0,mu1,sigma)
{delta <- mu1-mu0
t.crit <-qt(1-alpha/2,n-1) # gives t quantiles for RHS of either term
t.gamma <- sqrt(n)*(delta/sigma) # non-centrality parameter
t.power <- 1-pt(t.crit,n-1,ncp=t.gamma) # CDFs
            +pt(-t.crit,n-1,ncp=t.gamma)
return(t.power)
}
```

Power of the one-sample *t*-test

Consider the *t*-test for testing

$$H_0 : \mu = 0 \text{ versus } H_1 : \mu \neq 0$$

with $n = 10, \sigma = 0.2, \alpha = 0.05$. The power at $\mu = 0.15$ can be calculated by calling the function on the previous slide:

```
onesampttestpow(.05,10,0,.15,0.2)
```

```
[1] 0.5619339
```

Power of the one-sample *t*-test

Use the built-in function in R to calculate the power of *t*-test `power.t.test()`, in which `delta` $= \mu_1 - \mu_0$.

```
power.t.test(n = 10,delta = 0.15,sd = 0.2,
             sig.level = 0.05,type = "one.sample" )
```

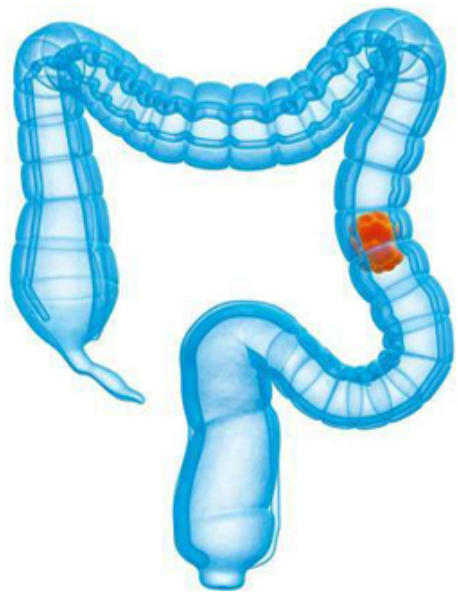
```
     One-sample t test power calculation

              n = 10
          delta = 0.15
             sd = 0.2
      sig.level = 0.05
          power = 0.5619339
    alternative = two.sided
```

Sign-posting: the R landscape

Procedure	Conditions	Function
Calculating power:		
One-sample z-test: testing a mean, when sigma is known	Equal allocation ✓	`pow.z.test`
One-sample t-test: testing a mean, when sigma is unknown	Equal allocation ✓	`power.t.test(type = "one.sample")`, `onesampttestpow`
Two-sample t-test: testing a difference in means	Equal allocation	`power.t.test, twosampttestpow`
Any test in which the test statistic has known distribution	Equal or unequal allocation	`replicate(N,t.test(...)$p.value)`
Calculating sample size:		
Experiment with continuous outcomes	Equal allocation	`power.t.test, size2z.uneq.test(r=1)`
Experiment with continuous outcomes	Unequal allocation	`size2z.uneq.test`
Experiment with binary outcomes	Equal allocation	`power.prop.test`
Experiment with binary outcomes	Unequal allocation	`replicate(N,prop.test(...)$p.value)`

Power of the two-sample t-test

Suppose a clinical trial to test a new treatment against the standard treatment for colon cancer is being designed. They wish to compare the tumour growth amount (in centimetres) under the new treatment versus that under the standard treatment.



Power of the two-sample t-test

- This is an example of a two-sample comparison with continuous outcomes. Let Y_{ik} be the observed outcome for the ith subject in the kth treatment group, for $i = 1, ..., n_k$, and $k = 1, 2$. The outcomes in the two groups are assumed to be independent and normally distributed with different means but an equal variance σ^2:

$$Y_{ik} \sim \mathcal{N}(\mu_k, \sigma^2).$$

- Let $\theta = \mu_1 - \mu_2$, the difference between the means of treatment 1 (the new therapy) and treatment 2 (the standard of care)
- To test whether the effects of the two treatments are the same, we formulate the null- and alternative hypotheses as

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta \neq 0.$$

Power of the two-sample t-test

Recall from Lecture 2 slide 48 that the two-sample t statistic is given by

$$T_n = \frac{\bar{Y}_1 - \bar{Y}_2}{s\sqrt{(1/n_1 + 1/n_2)}} \sim t_{n_1+n_2-2}.$$

More generally,

- $T_n \sim t_{n_1+n_2-2}$ under H_0
- $T_n \sim t_{n_1+n_2-2,\gamma}$ under H_1, with non-centrality parameter

$$\gamma = \frac{\mu_1 - \mu_2}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

Power of the two-sample t-test

H_0 is rejected if

$$|T_n| \geq t_{n_1+n_2-2,\,\alpha/2}$$

where $t_{df,\,\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the central t-distribution with df degrees of freedom. (Yin, pp. 164–165)

- Use `t.test()` to do the calculations

Power of the two-sample t-test

The power of the test is

$$1 - \beta = P(t_{n_1+n_2-2,\,\gamma} \geq t_{n_1+n_2-2,\,\alpha/2}) + P(t_{n_1+n_2-2,\,\gamma} < -t_{n_1+n_2-2,\,\alpha/2})$$

The sample size can be solved from this equation which does not have a closed form.

The sample size can be determined by specifying:

- type I and type II error rates
- the standard deviation
- the difference in treatment means that the clinical trial aims to detect

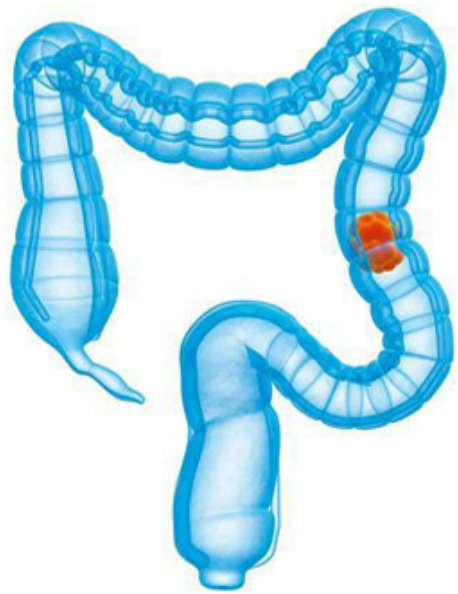
Power of the two-sample t-test

$$1-\beta = 1 - P(t_{n_1+n_2-2,\,\gamma} \geq t_{n_1+n_2-2,\,\alpha/2}) + P(t_{n_1+n_2-2,\,\gamma} < -t_{n_1+n_2-2,\,\alpha/2})$$

```
twosampttestpow <- function(alpha,n1,n2, mu1, mu2,sigma){
delta <- mu1-mu2
t.crit <-qt(1-alpha/2,n1+n2-2)
t.gamma <- delta/(sigma*sqrt(1/n1+1/n2))
t.power <- 1-pt(t.crit,n1+n2-2,ncp=t.gamma)+
             pt(-t.crit,n1+n2-2,ncp=t.gamma)
return(t.power)
}
```

Power of the two-sample t-test

Recall our clinical trial to test a new treatment against the standard treatment for colon cancer. The investigators feel that the smallest meaningful difference in tumour growth is 1 cm. The standard deviation of tumour growth is 3 cm. The investigators feel that they can enrol 50 subjects per arm. Will this clinical trial have adequate power to detect a difference between the treatments?



- What are the parameters of interest?
- What are the null and alternative hypotheses?
- How can the power of the study be calculated in R?

Power of the two-sample t-test

```
twosampttestpow(.05,50,50,1,2,3)
```

```
[1] 0.3785749
```

So, if the study were repeated 100 times and the difference is at least 1, then we can expect about 38 of the studies will reject H_0.

Power of the two-sample *t*-test

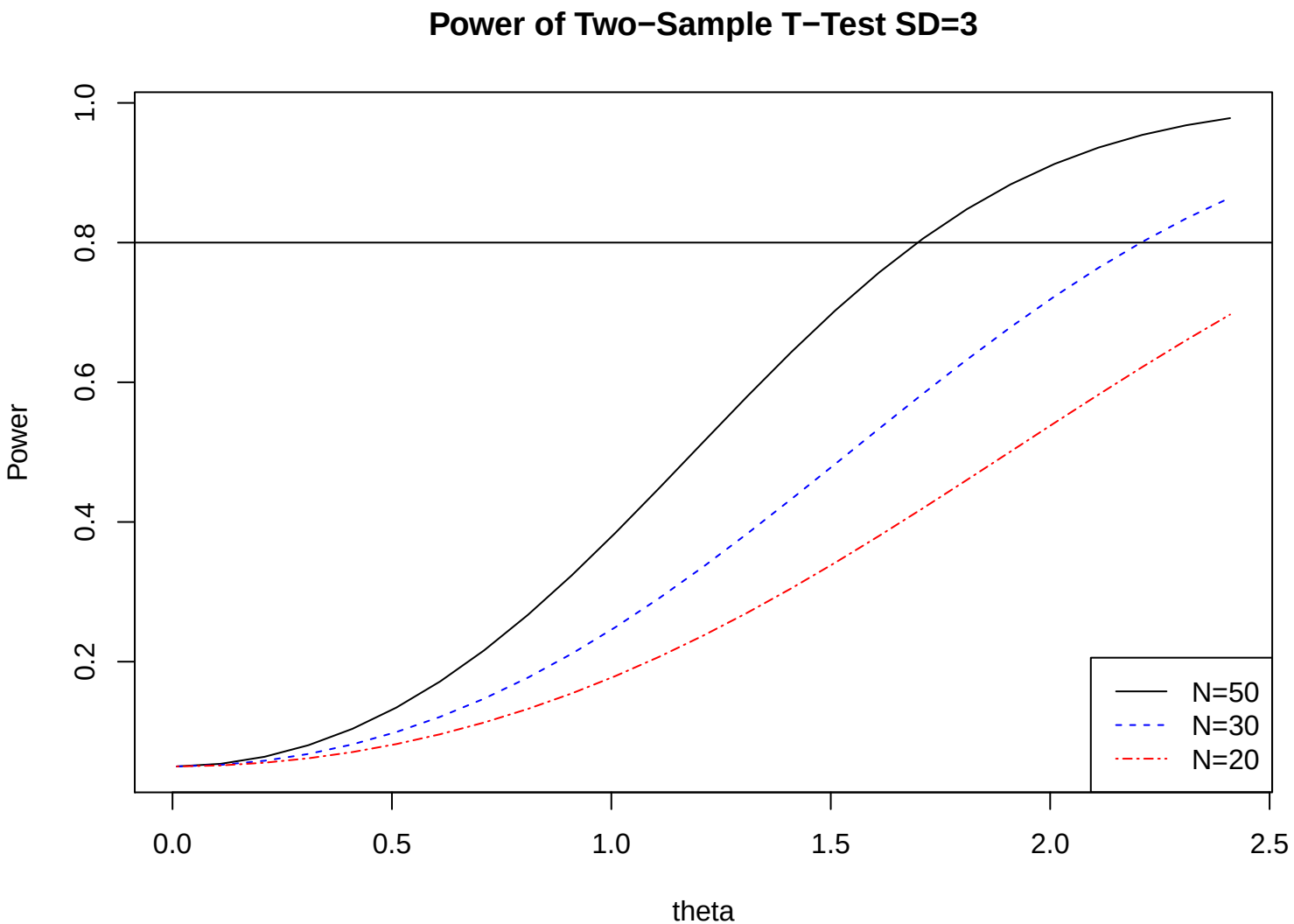
- `power.t.test()` can calculate the number of subjects required to achieve a certain power
- Suppose the investigators want to know how many subjects would have to be enrolled in each group to achieve 80% power under the same conditions

```
power.t.test(power = 0.8,delta = 1,sd = 3,sig.level = 0.05)
```

```
     Two-sample t test power calculation

              n = 142.2466
          delta = 1
             sd = 3
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

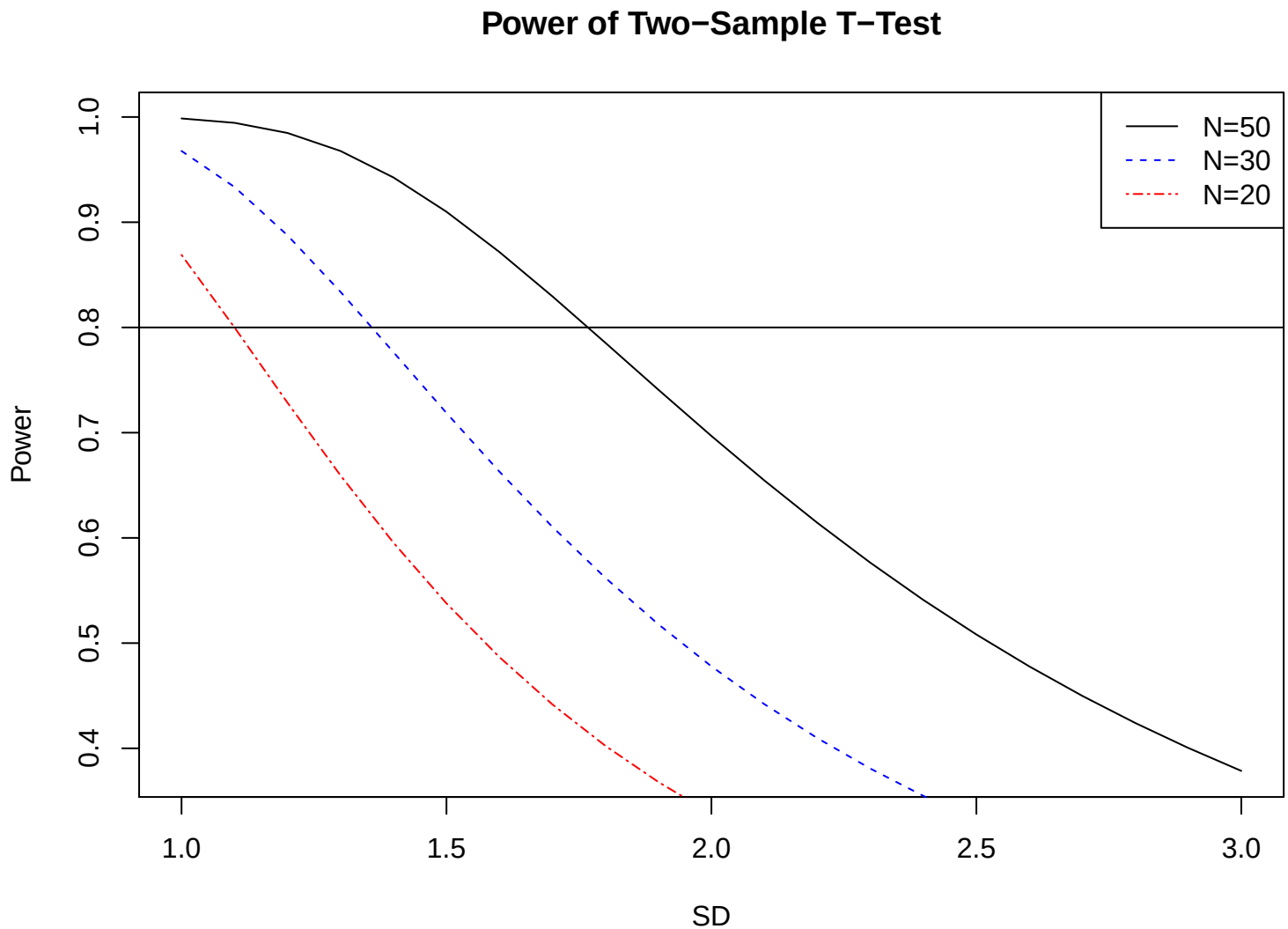
Power of the two-sample t-test

The following plot shows the power of the two-sample t-test as a function of the difference $\theta = \mu_1 - \mu_2$ to be detected and equal sample size per arm. As n decreases, power decreases.



Power of the two-sample t-test

This plot shows power as a function of σ and sample size per arm. As σ decreases, power increases.



A 'poweRful' function

- `power.t.test`: "Compute the power of the one- or two- sample *t*-test, or determine parameters to obtain a target power"
- `power.t.test (n, delta, sd, sig.level, power, type, alternative, ... )`
- Exactly one of the parameters n, delta, power, sd, and sig.level must be passed as NULL, and that parameter is determined from the others
- `type` is "two.sample", "one.sample", or "paired"
- `alternative` is "two.sided" or "one.sided"

Effect size

In some studies, instead of specifying separately the standard deviation and the difference in treatment means, the ratio

$$\text{ES} = \frac{\mu_1 - \mu_2}{\sigma}$$

can be specified. ES is called the effect size.

Cohen (1992) suggests that effect sizes of 0.2, 0.5, 0.8 correspond to small, medium, and large effects respectively

In general, an **effect size** (ES) is an index of the degree to which H_0 is false. The higher the ES, the more discrepancy there is with H_0. When ES $= 0$, there is no discrepancy.

Different definitions of ES (optional material)

The mathematical definition of ES depends on the hypothesis test you're using, and which author you read.

- If your test concerns the difference in means between two samples of real numbers, then ES is related to the concept of treatment effect. For example, Guosheng Yin's 2012 textbook describes ES as $m_A - m_B$. However, ES is defined as $(m_A - m_B)/\sigma$ by many other authors – including Jacob Cohen in a popular paper from 1992 (click). To help clarify which meaning is meant, you'll very occasionally see the term "scaled effect size" to emphasize what Jacob Cohen simply calls the effect size or the ES index.
- If your hypothesis test concerns a correlation coefficient, a difference in proportions, or one of various other possibilities, the expression changes accordingly. We don't cover most of these expressions but those interested can see page 157 of Cohen's paper for a list of ES formulae.

The 'Why' of using Effect Size

When designing an experiment, if possible you should estimate the probability that your plan will lead to achieving a statistically significant result. This will allow you to make adjustments to your experimental design before it's too late.

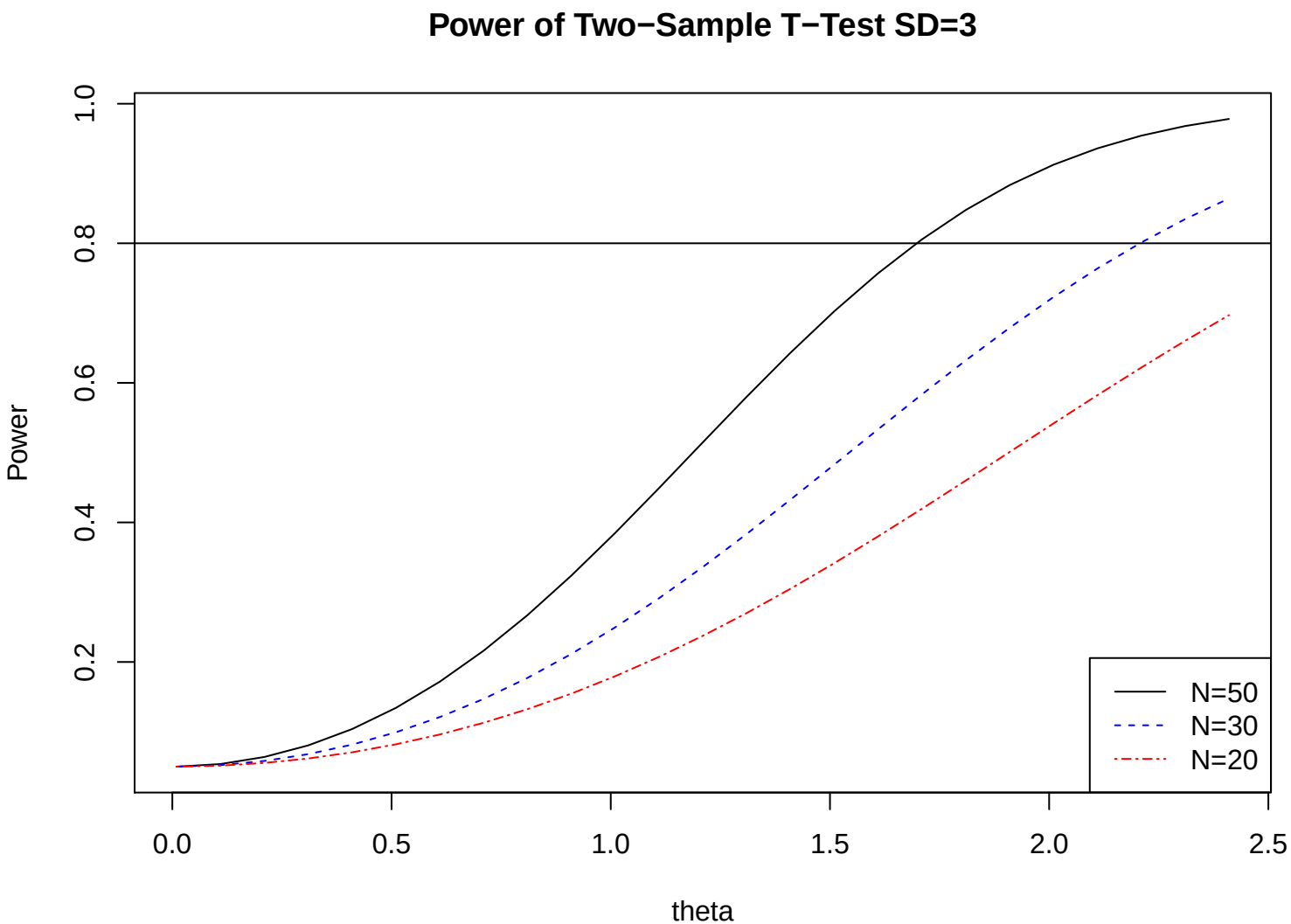
- For example, your estimate might reveal that you'd be unlikely to achieve a statistically significant result with 200 units
- In this case, you might cancel the experiment, redesign your experiment for 800 units before executing it, etc.
- "A stitch in time saves nine."

In making these calculations you must assume a particular ES, reflecting your belief about the expected departure from the null hypothesis. This is why we use the concept of effect size.

Power of the two-sample t-test

Power as a function of effect size can be investigated.

The plot shows that for $n_1 = n_2 = 10$ the two-sample t-test has at least 80% power for detecting effect sizes that are at least 1.3.



Sign-posting: the R landscape

Procedure	Conditions	Function
Calculating power:		
One-sample z-test: testing a mean, when sigma is known	Equal allocation ✓	`pow.z.test`
One-sample t-test: testing a mean, when sigma is unknown	Equal allocation ✓	`power.t.test(type = "one.sample")`, `onesampttestpow`
Two-sample t-test: testing a difference in means	Equal allocation ✓	`power.t.test, twosampttestpow`
Any test in which the test statistic has known distribution	Equal or unequal allocation	`replicate(N,t.test(...)$p.value)`
Calculating sample size:		
Experiment with continuous outcomes	Equal allocation	`power.t.test, size2z.uneq.test(r=1)`
Experiment with continuous outcomes	Unequal allocation	`size2z.uneq.test`
Experiment with binary outcomes	Equal allocation	`power.prop.test`
Experiment with binary outcomes	Unequal allocation	`replicate(N,prop.test(...)$p.value)`

Sample size - known variance and equal allocation

Allocation refers to: a clinical trial design strategy used to assign participants to an arm of a study.

If the variance is known then the test statistic is

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma\sqrt{(1/n_1 + 1/n_2)}} \sim \mathcal{N}(0,1).$$

This is the test statistic of the two-sample z-test.

The power at $\theta = \theta_1$ is given by

$$1-\beta = P\left(Z \geq z_{\alpha/2} - \frac{\theta_1}{\sigma\sqrt{1/n_1 + 1/n_2}}\right) + P\left(Z < -z_{\alpha/2} - \frac{\theta_1}{\sigma\sqrt{1/n_1 + 1/n_2}}\right).$$

Ignoring terms smaller than $\alpha/2$ and combining positive and negative θ,

$$\beta \approx \Phi\left(z_{\alpha/2} - \frac{|\theta_1|}{\sigma\sqrt{1/n_1 + 1/n_2}}\right).$$

Sample size - known variance and equal allocation

Recalling $z_{1-\beta} = \Phi^{-1}(\beta)$, the sample size is obtained by solving

$$z_\beta + z_{\alpha/2} = \left(\frac{|\theta_1|}{\sigma\sqrt{1/n_1 + 1/n_2}} \right).$$

If we assume that there will be an equal allocation of subjects to each group then $n_1 = n_2 = n/2$, the total sample size for the phase III trial is (using θ for θ_1, which you will recall is $\mu_1 - \mu_2$):

$$n = \frac{4\sigma^2 \left(z_\beta + z_{\alpha/2}\right)^2}{\theta^2}.$$

Sample size - known variance and unequal allocation

- In many trials it is desirable to put more patients into the experimental group to learn more about this treatment
- If the patient allocation between the two groups is $r = n_1/n_2$ then $n_1 = r \cdot n_2$ is the number of patients in the new-experimental treatment arm and n_2, the number in the standard control arm, is

$$n_2 = \frac{(1 + 1/r)\sigma^2 \left(z_\beta + z_{\alpha/2}\right)^2}{\theta^2}.$$

  An R function to compute the sample size in groups 1 and 2 for unequal allocation is

```
size2z.uneq.test <- function(theta,alpha,beta,sigma,r)
{ zalpha <- qnorm(1-alpha/2)
  zbeta <- qnorm(1-beta)
  n2 <- (1+1/r)*(sigma*(zalpha+zbeta)/theta)^2
  n1 <- r*n2
  return(c(n1,n2))}
```

Sample size - known variance and unequal allocation

What is the sample size required for 90% power to detect $\theta = 1$ with $\sigma = 2$ at the 5% level in a trial where two patients will be enrolled in the experimental arm for every patient enrolled in the control arm?

```
# sample size for theta =1, alpha=0.05,
# beta=0.1, sigma=2, r=2
# group 1 sample size (experimental group)
size2z.uneq.test(1,.05,.1,2,2)[1]
```

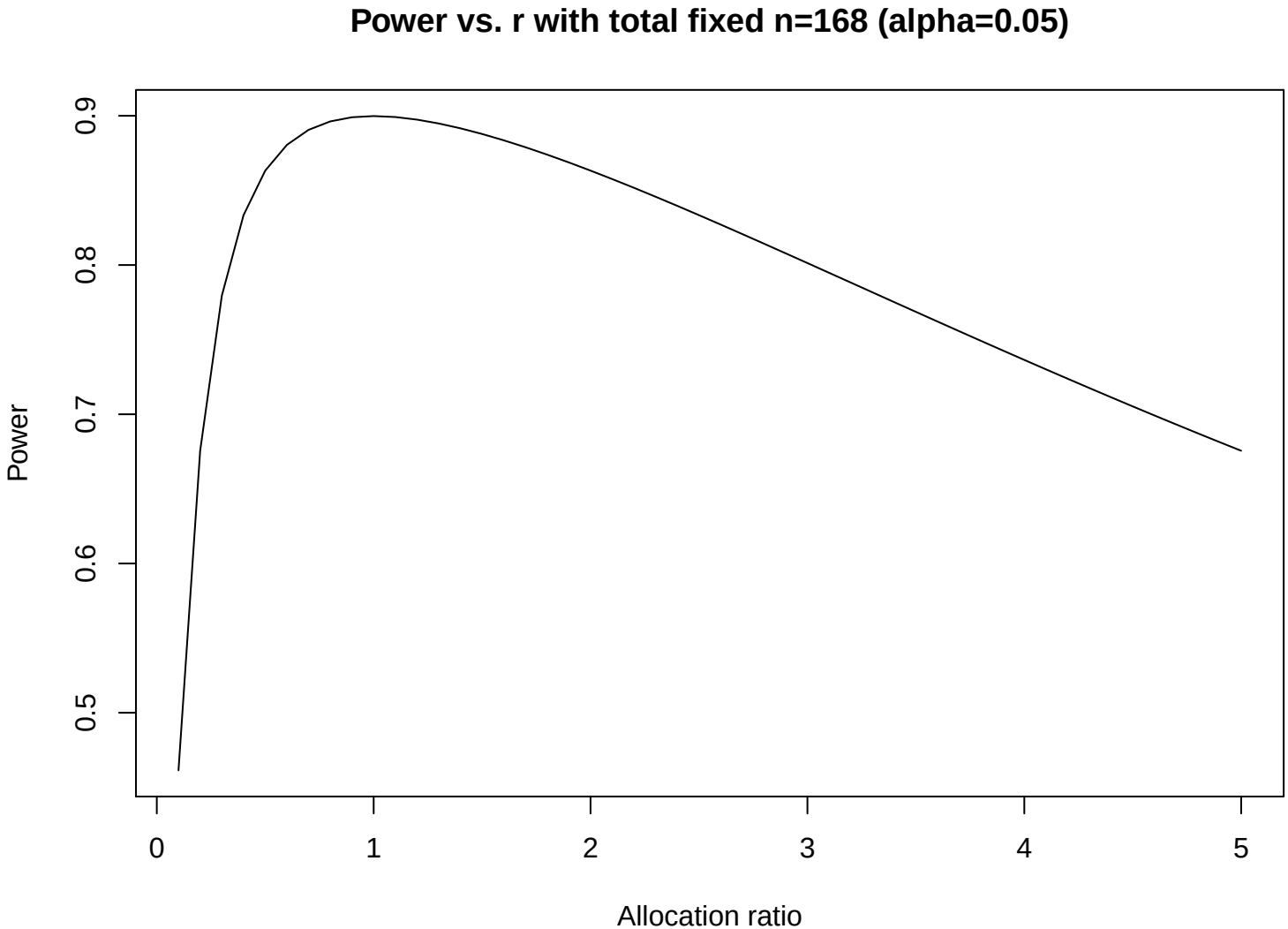
```
[1] 126.0891
```

```
# group 2 sample size (control group)
size2z.uneq.test(1,.05,.1,2,2)[2]
```

```
[1] 63.04454
```

Sample size - known variance and unequal allocation

Two-sample z-test's power as a function of the allocation ratio $r = n_1/n_2$:



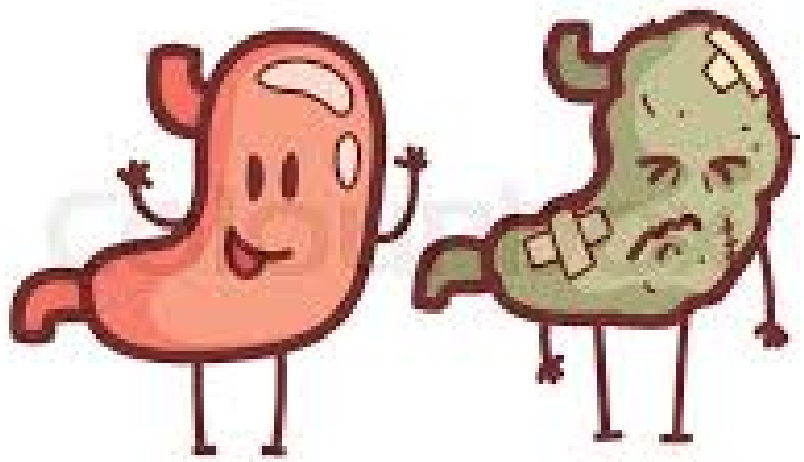
The plot shows that imbalance typically leads to loss of power.

Sign-posting: the R landscape

Procedure	Conditions	Function
Calculating power:		
One-sample z-test: testing a mean, when sigma is known	Equal allocation ✓	`pow.z.test`
One-sample t-test: testing a mean, when sigma is unknown	Equal allocation ✓	`power.t.test(type = "one.sample")`, `onesampttestpow`
Two-sample t-test: testing a difference in means	Equal allocation ✓	`power.t.test, twosampttestpow`
Any test in which the test statistic has known distribution	Equal or unequal allocation	`replicate(N,t.test(...)$p.value)`
Calculating sample size:		
Experiment with continuous outcomes	Equal allocation ✓	`power.t.test, size2z.uneq.test(r=1)`
Experiment with continuous outcomes	Unequal allocation ✓	`size2z.uneq.test`
Experiment with binary outcomes	Equal allocation	`power.prop.test`
Experiment with binary outcomes	Unequal allocation	`replicate(N,prop.test(...)$p.value)`

Comparing Proportions for Binary Outcomes

- In many clinical trials, the primary endpoint is dichotomous: for example, either a patient has responded to the treatment, or a patient has experienced toxicity



- Consider a two-arm randomized trial with binary outcomes. Let p_1 denote the response rate of the experimental drug, p_2 as that of the standard drug, and the difference is $\theta = p_1 - p_2$

Comparing Proportions for Binary Outcomes

Let Y_{ik} be the binary outcome for subject i in arm k, and $Y_{ik} \sim \text{Bern}(p_k)$ — that is,

$$Y_{ik} = \begin{cases} 1 & \text{with probability } p_k \\ 0 & \text{with probability } 1 - p_k, \end{cases}$$

for $i = 1, ..., n_k$ and $k = 1, 2$. The sum of independent and identically distributed Bernoulli random variables has a binomial distribution,

$$\sum_{i=1}^{n_k} Y_{ik} \sim \text{Bin}(n_k, p_k),\ k = 1, 2.$$

(Yin, pp 173–174)

Comparing Proportions for Binary Outcomes

The sample proportion for group k is

$$\hat{p}_k = \bar{Y}_k = (1/n_k) \sum_{i=1}^{n_k} Y_{ik},\ k = 1, 2,$$

and $\mathrm{E}\left(\bar{Y}_k\right) = p_k$ and $\mathrm{var}\left(\bar{Y}_k\right) = \frac{p_k(1-p_k)}{n_k}$.

The goal of the clinical trial is to determine whether there's a difference between the two groups using a binary endpoint. That is, we want to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.

The test statistic (assuming that H_0 is true) is:

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \sim \mathcal{N}(0, 1)$$

Comparing Proportions for Binary Outcomes

The test rejects at level α if and only if

$$|T| \geq z_{\alpha/2}.$$

Using the same argument as the case with continuous endpoints and ignoring terms smaller than $\alpha/2$ we can solve for β:

$$\beta \approx \Phi\left(z_{\alpha/2} - \frac{|\theta_1|}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}\right).$$

Comparing Proportions for Binary Outcomes

Use this formula to solve for sample size. If $n_1 = r \cdot n_2$ then

$$n_2 = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{\theta^2} \left(p_1(1-p_1)/r + p_2(1-p_2)\right).$$

Comparing Proportions for Binary Outcomes

- The built-in R function `power.prop.test()` can be used to calculate sample size or power
- For example suppose that the standard treatment for a disease has a response rate of 20%, and an experimental treatment is anticipated to have a response rate of 30%
- The researchers want both arms to have an equal number of subjects. How many patients should be enrolled if the study will conduct a two-sided test at the 5% level with 80% power?

```
power.prop.test(p1 = 0.2,p2 = 0.3,power = 0.8)
```

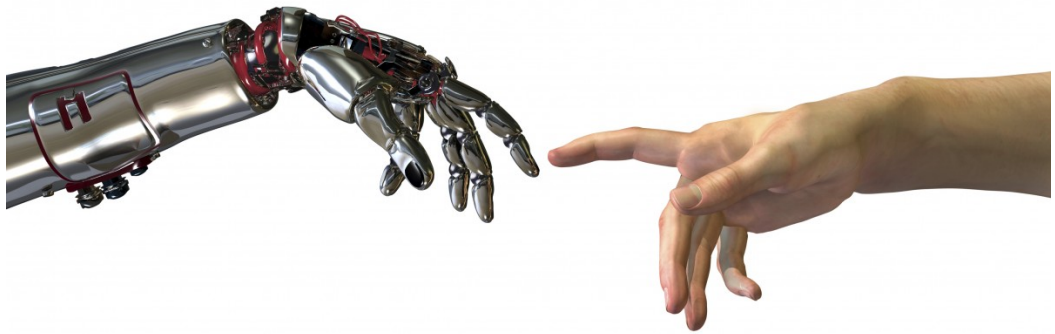
```
     Two-sample comparison of proportions power calculation

              n = 293.1513
             p1 = 0.2
             p2 = 0.3
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Sign-posting: the R landscape

Procedure	Conditions	Function
Calculating power:		
One-sample z-test: testing a mean, when sigma is known	Equal allocation ✓	`pow.z.test`
One-sample t-test: testing a mean, when sigma is unknown	Equal allocation ✓	`power.t.test(type = "one.sample")`, `onesampttestpow`
Two-sample t-test: testing a difference in means	Equal allocation ✓	`power.t.test, twosampttestpow`
Any test in which the test statistic has known distribution	Equal or unequal allocation	`replicate(N,t.test(...)$p.value)`
Calculating sample size:		
Experiment with continuous outcomes	Equal allocation ✓	`power.t.test, size2z.uneq.test(r=1)`
Experiment with continuous outcomes	Unequal allocation ✓	`size2z.uneq.test`
Experiment with binary outcomes	Equal allocation ✓	`power.prop.test`
Experiment with binary outcomes	Unequal allocation	`replicate(N,prop.test(...)$p.value)`

Calculating Power by Simulation

- If the test statistic and distribution of the test statistic are known then the power of the test can be calculated via simulation



- Consider a two-sample t-test with 30 subjects per arm and the standard deviation of the clinical outcome is known to be 1
- What is the power of the test $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$, at $\mu_1 - \mu_2 = 0.5$, at the 5% significance level?
- Answer: the proportion of times that the test correctly rejects the null hypothesis in repeated sampling (i.e. when $\mu_1 - \mu_2 = 0.5$)

Calculating Power by Simulation

We can simulate a single study using the `rnorm()` command. Let's assume that $n_1 = n_2 = 30, \mu_1 = 3.5, \mu_2 = 3, \sigma = 1, \alpha = 0.05$.

```
set.seed(2301)
t.test(rnorm(30,mean=3.5,sd=1),rnorm(30,mean=3,sd=1),var.equal = T)
```

```
    Two Sample t-test

data:  rnorm(30, mean = 3.5, sd = 1) and rnorm(30, mean = 3, sd = 1)
t = 2.1462, df = 58, p-value = 0.03605
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03458122 0.99248595
sample estimates:
mean of x mean of y
 3.339362  2.825828
```

Should you reject H_0?

Calculating Power by Simulation

- Suppose that 10 studies are simulated
- What proportion of these 10 studies will reject the null hypothesis at the 5% level?
- To investigate how many times the two-sample t-test will reject at the 5% level the `replicate()` command will be used to generate 10 studies and calculate the p-value in each study
- It will still be assumed that $n_1 = n_2 = 30, \mu_1 = 3.5, \mu_2 = 3, \sigma = 1, \alpha = 0.05$

```
set.seed(2301)
pvals <- replicate(10,t.test(rnorm(30,mean=3.5,sd=1),
                             rnorm(30,mean=3,sd=1),
                             var.equal = T)$p.value)
pvals # print out 10 p-values
```

```
 [1] 0.03604893 0.15477655 0.01777959 0.40851999 0.34580930 0.11131007
 [7] 0.14788381 0.00317709 0.09452230 0.39173723
```

```
#power is the number of times the test rejects at the 5% level
sum(pvals<=0.05)/10
```

```
[1] 0.3
```

Calculating Power by Simulation

But, since we only simulated 10 studies the estimate of power will have a large standard error. So let's try simulating 10,000 studies so that we can obtain a more precise estimate of power.

```
set.seed(2301)
pvals <- replicate(10000,t.test(rnorm(30,mean=3.5,sd=1),
                                rnorm(30,mean=3,sd=1),
                                var.equal = T)$p.value)
sum(pvals<=0.05)/10000
```

```
[1] 0.4881
```

Calculating Power by Simulation

This is much closer to the theoretical power obtained from `power.t.test()`.

```
power.t.test(n = 30,delta = 0.5,sd = 1,sig.level = 0.05)
```

```
     Two-sample t test power calculation

              n = 30
          delta = 0.5
             sd = 1
      sig.level = 0.05
          power = 0.477841
    alternative = two.sided

NOTE: n is number in *each* group
```

Calculating Power by Simulation

- The built-in R functions `power.t.test()` and `power.prop.test()` don't have an option for calculating power where there is unequal allocation of subjects between arms
- One option is to simulate power for the scenarios that are of interest. Another option is to write your own function using the formula on slide 51 or 59

Calculating Power by Simulation

Suppose the standard treatment for a disease has a response rate of 20%, and an experimental treatment is anticipated to have a response rate of 30%

- Recall that the researchers had wanted both arms to have an equal number of subjects
- Our power calculation on slide 60 revealed that the study would require 294×2 patients for 80% power

What would happen to the power if the researchers put 441 patients in the experimental arm and 147 patients in the control arm? (Either study involves 588 patients.)

Calculating Power by Simulation

- The number of subjects in the experimental arm that have a positive response to treatment will be an observation from a Bin$(441, 0.3)$
- The number of subjects that have a positive response to the standard treatment will be an observation from a Bin$(147, 0.2)$
- We can obtain simulated responses from these distributions using the `rbinom()` command in R

```
set.seed(2301)
rbinom(1,441,0.30)
```

```
[1] 125
```

```
rbinom(1,147,0.20)
```

```
[1] 30
```

Calculating Power by Simulation

The *p*-value for this simulated study can be obtained using `prop.test()`.

```
set.seed(2301)
prop.test(x=c(rbinom(1,441,0.3),rbinom(1,147,0.2)),
          n=c(441,147),correct = F)
```

```
    2-sample test for equality of proportions without continuity
    correction

data:  c(rbinom(1, 441, 0.3), rbinom(1, 147, 0.2)) out of c(441, 147)
X-squared = 3.5774, df = 1, p-value = 0.05857
alternative hypothesis: two.sided
95 percent confidence interval:
 0.001815455 0.156914704
sample estimates:
   prop 1    prop 2
0.2834467 0.2040816
```

Calculating Power by Simulation

- A power simulation repeats this process a large number of times
- In the example below we simulate 10,000 hypothetical studies to calculate power

```
set.seed(2301)
pvals <- replicate(10000,
prop.test(x=c(rbinom(n = 1,size = 441,prob = 0.30),
              rbinom(n=1,size=147,prob=0.20)),
              n=c(441,147),correct=F)$p.value)
sum(pvals<=0.05)/10000
```

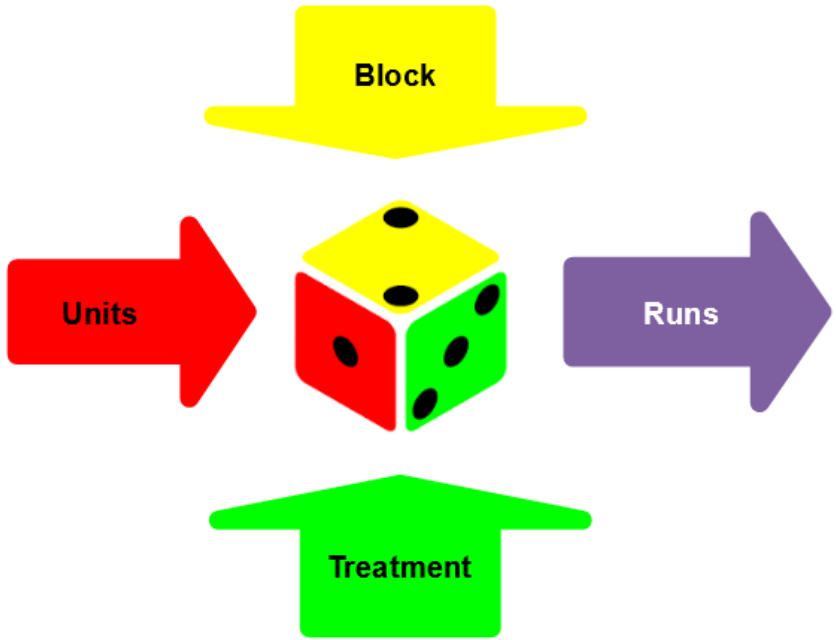
```
[1] 0.6749
```

If the researchers decide to have a 3:1 allocation ratio of patients in the treatment to control arm then the power will be _____.

The R landscape

Procedure	Conditions	Function
Calculating power:		
One-sample z-test: testing a mean, when sigma is known	Equal allocation	✓ `pow.z.test`
One-sample t-test: testing a mean, when sigma is unknown	Equal allocation	✓ `power.t.test(type = "one.sample")`, `onesampttestpow`
Two-sample t-test: testing a difference in means	Equal allocation	✓ `power.t.test, twosampttestpow`
Any test in which the test statistic has known distribution	Equal or unequal allocation	✓ `replicate(N,t.test(...)$p.value)`
Calculating sample size:		
Experiment with continuous outcomes	Equal allocation	✓ `power.t.test, size2z.uneq.test(r=1)`
Experiment with continuous outcomes	Unequal allocation	✓ `size2z.uneq.test`
Experiment with binary outcomes	Equal allocation	✓ `power.prop.test`
Experiment with binary outcomes	Unequal allocation	✓ `replicate(N,prop.test(...)$p.value)`

What next?



Next steps

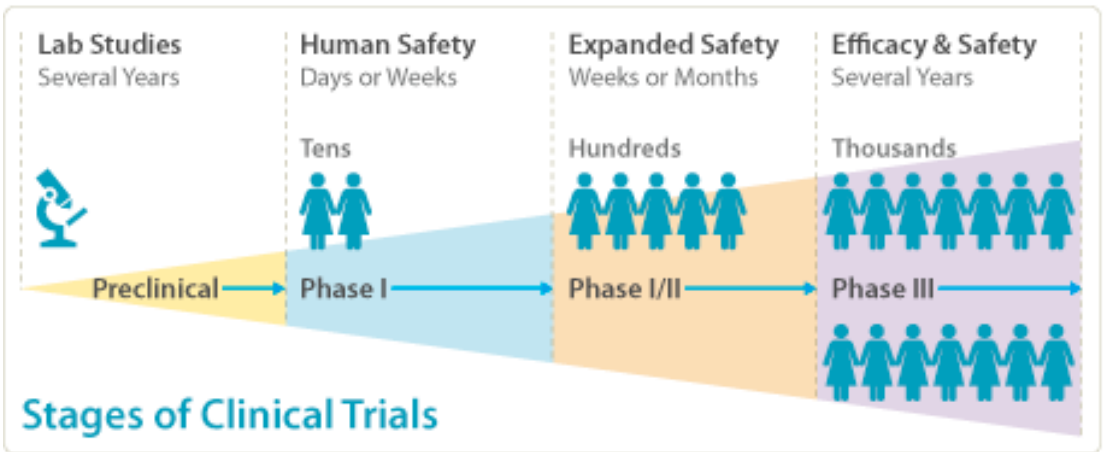
In-pairs work: Questions 1, 2, and 3 (Unit 3 notes)

Appendix: How do clinical trials work?

Clinical trials are a useful & inspiring topic in the design of scientific studies.

Phases of clinical trials: e.g. Developing a new drug

- **Preclinical studies:** In-vitro (e.g. slides, test tubes) and in-vivo (living organism such as rodents) studies on a wide range of doses of experimental agents. This stage of study provides preliminary toxicity and efficacy data including pharmacokinetics (PK) and pharmacodynamics (PD) information
- **Phase I:** Usually first study in humans to investigate the toxicity and side effects of the new agent. Identify maximum tolerated dose (MTD)
- **Phase II:** Assess if drug has sufficient efficacy. The drug is usually administered around the MTD. If drug does not show efficacy or is too toxic then further testing is discontinued



Optional material

Phases of clinical trials

- **Phase III:** If drug passes phase II testing then it is compared to the current standard of care or placebo. These are long-term, large-scale randomized studies that may involve hundreds or thousands of patients.
- If the drug is proven to be effective (e.g. two positive phase III trials required for FDA approval) the company will file an application with regulatory agencies to sell the drug. If approved then the drug will be available to the general population in the country where it was approved.
- **Phase IV:** After approval a study might follow a large number of patients over a longer period of time to monitor side effects and drug interactions. For example, findings from these studies might add a warning label to the drug.

One of the FDA's counterparts here is Health Canada's Health Products and Food Branch (HPFB).

Optional material

Phases of clinical trials

- The four phases are usually conducted sequentially and separately
- Each trial requires an independent study design and a study protocol
- Every aspect of trial design, monitoring, and data analysis calls upon statistical methods
- In randomized clinical trials, an example of a treatment group (**arm**) is all of the 0's in this treatment vector:

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Optional material

Phases of clinical trials

- Experimental design plays a very important role in setting up clinical trials
- Two-arm clinical trials use all of the theory of randomization that we learned about last week. Randomization is used particularly to design phase III clinical trials since causation can usually be assessed using a randomized design

Optional material