

2 Phase I and II clinical trials

2.1 Phases of Clinical Trials

The process of drug development can be broadly classified as pre-clinical and clinical. Pre-clinical refers to experimentation that occurs before it is given to human subjects; whereas, clinical refers to experimentation with humans. This course will consider only clinical research. It will be assumed that the drug has already been developed by the chemist or biologist, tested in the laboratory for biologic activity (in vitro), that preliminary tests on animals have been conducted (in vivo) and that the new drug or therapy is found to be sufficiently promising to be introduced into humans.

Within the realm of clinical research, clinical trials are classified into four phases.

- **Phase I:** To explore possible toxic effects of drugs and determine a tolerated dose for further experimentation. Also during Phase I experimentation the pharmacology of the drug may be explored.
- **Phase II:** Screening and feasibility by initial assessment for therapeutic effects; further assessment of toxicities.
- **Phase III:** Comparison of new intervention (drug or therapy) to the current standard of treatment; both with respect to efficacy and toxicity.
- **Phase IV:** (post marketing) Observational study of morbidity/adverse effects.

These definitions of the four phases are not hard and fast. Many clinical trials blur the lines between the phases. Loosely speaking, the logic behind the four phases is as follows:

A new promising drug is about to be assessed in humans. The effect that this drug might have on humans is unknown. We might have some experience on similar acting drugs developed in the past and we may also have some data on the effect this drug has on animals but we are not sure what the effect is on humans. To study the initial effects, a Phase I study is conducted. Using increasing doses of the drug on a small number of subjects, the possible side effects of the drug are documented. It is during this phase that the tolerated dose is determined for future

experimentation. The general dogma is that the therapeutic effect of the drug will increase with dose, but also the toxic effects will increase as well. Therefore one of the goals of a Phase I study is to determine what the maximum dose should be that can be reasonably tolerated by most individuals with the disease. The determination of this dose is important as this will be used in future studies when determining the effectiveness of the drug. If we are too conservative then we may not be giving enough drug to get the full therapeutic effect. On the other hand if we give too high a dose then people will have adverse effects and not be able to tolerate the drug.

Once it is determined that a new drug can be tolerated and a dose has been established, the focus turns to whether the drug is good. Before launching into a costly large-scale comparison of the new drug to the current standard treatment, a smaller feasibility study is conducted to assess whether there is sufficient efficacy (activity of the drug on disease) to warrant further investigation. This occurs during phase II where drugs which show little promise are screened out.

If the new drug still looks promising after phase II investigation it moves to Phase III testing where a comparison is made to a current standard treatment. These studies are generally large enough so that important treatment differences can be detected with sufficiently large probability. These studies are conducted carefully using sound statistical principles of experimental design established for clinical trials to make objective and unbiased comparisons. It is on the basis of such Phase III clinical trials that new drugs are approved by regulatory agencies (such as FDA) for the general population of individuals with the disease for which this drug is targeted.

Once a drug is on the market and a large number of patients are taking it, there is always the possibility of rare but serious side effects that can only be detected when a large number are given treatment for sufficiently long periods of time. It is important that a monitoring system be in place that allows such problems, if they occur, to be identified. This is the role of Phase IV studies.

A brief discussion of phase I studies and designs and Pharmacology studies will be given based on the slides from Professor Marie Davidian, an expert in pharmacokinetics. Slides on phase I and pharmacology will be posted on the course web page.

2.2 Phase II clinical trials

After a new drug is tested in phase I for safety and tolerability, a dose finding study is sometimes conducted in phase II to identify a lowest dose level with good efficacy (close to the maximum efficacy achievable at tolerable dose level). In other situations, a phase II clinical trial uses a fixed dose chosen on the basis of a phase I clinical trial. The total dose is either fixed or may vary depending on the weight of the patient. There may also be provisions for modification of the dose if toxicity occurs. The study population are patients with a specified disease for which the treatment is targeted.

The primary objective is to determine whether the new treatment should be used in a large-scale comparative study. Phase II trials are used to assess

- feasibility of treatment
- side effects and toxicity
- logistics of administration and cost

The major issue that is addressed in a phase II clinical trial is whether there is enough evidence of efficacy to make it worth further study in a larger and more costly clinical trial. In a sense this is an initial screening tool for efficacy. During phase II experimentation the treatment efficacy is often evaluated on **surrogate markers**; i.e on an outcome that can be measured quickly and is believed to be related to the clinical outcome.

Example: Suppose a new drug is developed for patients with lung cancer. Ultimately, we would like to know whether this drug will extend the life of lung cancer patients as compared to currently available treatments. Establishing the effect of a new drug on survival would require a long study with relatively large number of patients and thus may not be suitable as a screening mechanism. Instead, during phase II, the effect of the new drug may be assessed based on tumor shrinkage in the first few weeks of treatment. If the new drug shrinks tumors sufficiently for a sufficiently large proportion of patients, then this may be used as evidence for further testing.

In this example, tumor shrinkage is a surrogate marker for overall survival time. The belief is that if the drug has no effect on tumor shrinkage it is unlikely to have an effect on the patient's

overall survival and hence should be eliminated from further consideration. Unfortunately, there are many instances where a drug has a short term effect on a surrogate endpoint but ultimately may not have the long term effect on the clinical endpoint of ultimate interest. Furthermore, sometimes a drug may have beneficial effect through a biological mechanism that is not detected by the surrogate endpoint. Nonetheless, there must be some attempt at limiting the number of drugs that will be considered for further testing or else the system would be overwhelmed.

Other examples of surrogate markers are

- Lowering blood pressure or cholesterol for patients with heart disease
- Increasing CD4 counts or decreasing viral load for patients with HIV disease

Most often, phase II clinical trials do not employ formal comparative designs. That is, they do not use parallel treatment groups. Often, phase II designs employ more than one stage; i.e. one group of patients are given treatment; if no (or little) evidence of efficacy is observed, then the trial is stopped and the drug is declared a failure; otherwise, more patients are entered in the next stage after which a final decision is made whether to move the drug forward or not.

2.2.1 Statistical Issues and Methods

One goal of a phase II trial is to estimate an endpoint related to treatment efficacy with sufficient precision to aid the investigators in determining whether the proposed treatment should be studied further.

Some examples of endpoints are:

- proportion of patients responding to treatment (response has to be unambiguously defined)
- proportion with side effects
- average decrease in blood pressure over a two week period

A statistical perspective is generally taken for estimation and assessment of precision. That is, the problem is often posed through a statistical model with population parameters to be estimated and confidence intervals for these parameters to assess precision.

Example: Suppose we consider patients with esophageal cancer treated with chemotherapy prior to surgical resection. A complete response is defined as an absence of macroscopic or microscopic tumor at the time of surgery. We suspect that this may occur with 35% (guess) probability using a drug under investigation in a phase II study. The 35% is just a guess, possibly based on similar acting drugs used in the past, and the goal is to estimate the actual response rate with sufficient precision, in this case we want the 95% confidence interval to be within 15% of the truth.

As statisticians, we view the world as follows: We start by positing a statistical model; that is, let π denote the population complete response rate. We conduct an experiment: n patients with esophageal cancer are treated with the chemotherapy prior to surgical resection and we collect data: the number of patients who have a complete response.

The result of this experiment yields a random variable X , the number of patients in a sample of size n that have a complete response. A popular model for this scenario is to assume that

$$X \sim \text{binomial}(n, \pi);$$

that is, the random variable X is distributed with a binomial distribution with sample size n and success probability π . The goal of the study is to estimate π and obtain a confidence interval.

I believe it is worth stepping back a little and discussing how the actual experiment and the statistical model used to represent this experiment relate to each other and whether the implicit assumptions underlying this relationship are reasonable.

Statistical Model

What is the population? All people now and in the future with esophageal cancer who would be eligible for the treatment.

What is π ? (the population parameter)

If all the people in the hypothetical population above were given the new chemotherapy, then π would be the proportion who would have a complete response. This is a hypothetical construct. Neither can we identify the population above or could we actually give them all the chemotherapy. Nonetheless, let us continue with this mind experiment.

We assume the random variable X follows a binomial distribution. Is this reasonable? Let us review what it means for a random variable to follow a binomial distribution.

X being distributed as a binomial $b(n, \pi)$ means that X corresponds to the number of successes (complete responses) in n independent trials where the probability of success for each trial is equal to π . This would be satisfied, for example, if we were able to identify every member of the population and then, using a random number generator, chose n individuals at random from our population to test and determine the number of complete responses.

Clearly, this is not the case. First of all, the population is a hypothetical construct. Moreover, in most clinical studies the sample that is chosen is an **opportunistic** sample. There is generally no attempt to randomly sample from a specific population as may be done in a survey sample. Nonetheless, a statistical perspective may be a useful construct for assessing variability. I sometimes resolve this in my own mind by thinking of the hypothetical population that I can make inference on as all individuals who might have been chosen to participate in the study with whatever process that was actually used to obtain the patients that were actually studied. However, this limitation must always be kept in mind when one extrapolates the results of a clinical experiment to a more general population.

Philosophical issues aside, let us continue by assuming that the posited model is a reasonable approximation to some question of relevance. Thus, we will assume that our data is a realization of the random variable X , assumed to be distributed as $b(n, \pi)$, where π is the population parameter of interest.

Reviewing properties about a binomial distribution we note the following:

- $E(X) = n\pi$, where $E(\cdot)$ denotes expectation of the random variable.
- $Var(X) = n\pi(1 - \pi)$, where $Var(\cdot)$ denotes the variance of the random variable.
- $P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$, where $P(\cdot)$ denotes probability of an event, and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- Denote the sample proportion by $p = X/n$, then

$$- E(p) = \pi$$

$$- \text{Var}(p) = \pi(1 - \pi)/n$$

- When n is sufficiently large, the distribution of the sample proportion $p = X/n$ is well approximated by a normal distribution with mean π and variance $\pi(1 - \pi)/n$:

$$p \sim N(\pi, \pi(1 - \pi)/n).$$

This approximation is useful for inference regarding the population parameter π . Because of the approximate normality, the estimator p will be within 1.96 standard deviations of π approximately 95% of the time. (Approximation gets better with increasing sample size). Therefore the population parameter π will be within the interval

$$p \pm 1.96\{\pi(1 - \pi)/n\}^{1/2}$$

with approximately 95% probability. Since the value π is unknown to us, we approximate using p to obtain the approximate 95% confidence interval for π , namely

$$p \pm 1.96\{p(1 - p)/n\}^{1/2}.$$

Going back to our example, where our best guess for the response rate is about 35%, if we want the precision of our estimator to be such that the 95% confidence interval is within 15% of the true π , then we need

$$1.96\left\{\frac{(.35)(.65)}{n}\right\}^{1/2} = .15,$$

or

$$n = \frac{(1.96)^2(.35)(.65)}{(.15)^2} = 39 \text{ patients.}$$

Since the response rate of 35% is just a guess which is made before data are collected, the exercise above should be repeated for different feasible values of π before finally deciding on how large the sample size should be.

Exact Confidence Intervals

If either $n\pi$ or $n(1 - \pi)$ is small, then the normal approximation given above may not be adequate for computing accurate confidence intervals. In such cases we can construct **exact** (usually conservative) confidence intervals.

We start by reviewing the definition of a confidence interval and then show how to construct an exact confidence interval for the parameter π of a binomial distribution.

Definition: The definition of a $(1 - \alpha)$ -th confidence region (interval) for the parameter π is as follows:

For each realization of the data $X = k$, a region of the parameter space, denoted by $\mathcal{C}(k)$ (usually an interval) is defined in such a way that the random region $\mathcal{C}(X)$ contains the true value of the parameter with probability greater than or equal to $(1 - \alpha)$ regardless of the value of the parameter. That is,

$$P_{\pi}\{\mathcal{C}(X) \supset \pi\} \geq 1 - \alpha, \text{ for all } 0 \leq \pi \leq 1,$$

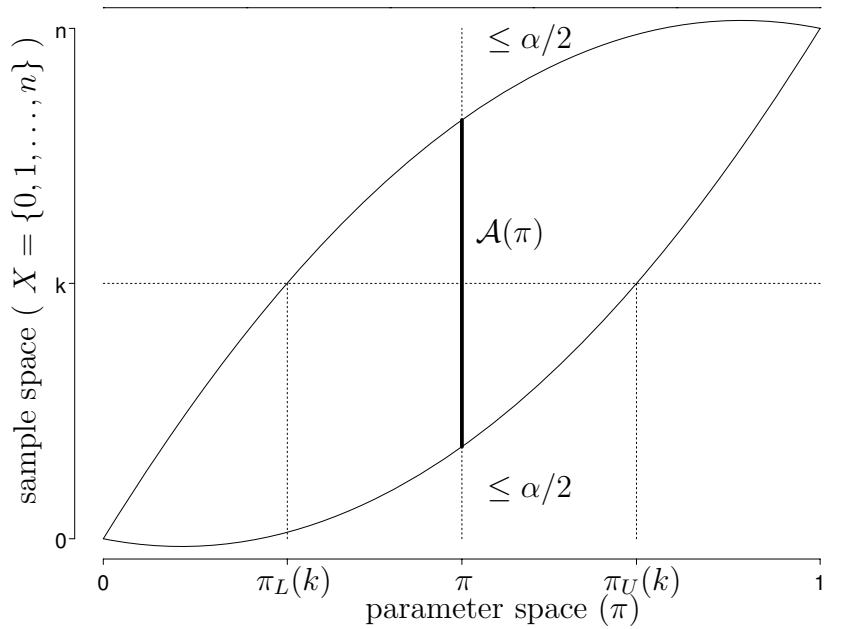
where $P_{\pi}(\cdot)$ denotes probability calculated under the assumption that $X \sim b(n, \pi)$ and \supset denotes “contains”. The confidence interval is the random interval $\mathcal{C}(X)$. After we collect data and obtain the realization $X = k$, then the corresponding confidence interval is defined as $\mathcal{C}(k)$.

This definition is equivalent to defining an acceptance region (of the sample space) for each value π , denoted as $\mathcal{A}(\pi)$, that has probability greater than equal to $1 - \alpha$, i.e.

$$P_{\pi}\{X \in \mathcal{A}(\pi)\} \geq 1 - \alpha, \text{ for all } 0 \leq \pi \leq 1,$$

in which case $\mathcal{C}(k) = \{\pi : k \in \mathcal{A}(\pi)\}$.

We find it useful to consider a graphical representation of the relationship between confidence intervals and acceptance regions.

Figure 2.1: *Exact confidence intervals*

Another way of viewing a $(1 - \alpha)$ -th confidence interval is to find, for each realization $X = k$, all the values π^* for which the value k would not reject the hypothesis $H_0 : \pi = \pi^*$. Therefore, a $(1 - \alpha)$ -th confidence interval is sometimes more appropriately called a $(1 - \alpha)$ -th credible region (interval).

If $X \sim b(n, \pi)$, then when $X = k$, the $(1 - \alpha)$ -th confidence interval is given by

$$\mathcal{C}(k) = [\pi_L(k), \pi_U(k)],$$

where $\pi_L(k)$ denotes the lower confidence limit and $\pi_U(k)$ the upper confidence limit, which are defined as

$$P_{\pi_L(k)}(X \geq k) = \sum_{j=k}^n \binom{n}{j} \pi_L(k)^j \{1 - \pi_L(k)\}^{n-j} = \alpha/2,$$

and

$$P_{\pi_U(k)}(X \leq k) = \sum_{j=0}^k \binom{n}{j} \pi_U(k)^j \{1 - \pi_U(k)\}^{n-j} = \alpha/2.$$

The values $\pi_L(k)$ and $\pi_U(k)$ need to be evaluated numerically as we will demonstrate shortly.

Remark: Since X has a discrete distribution, the way we define the $(1-\alpha)$ -th confidence interval above will yield

$$P_{\pi}\{\mathcal{C}(X) \supset \pi\} > 1 - \alpha$$

(strict inequality) for most values of $0 \leq \pi \leq 1$. Strict equality cannot be achieved because of the discreteness of the binomial random variable.

Example: In a Phase II clinical trial, 3 of 19 patients respond to α -interferon treatment for multiple sclerosis. In order to find the exact confidence 95% interval for π for $X = k$, $k = 3$, and $n = 19$, we need to find $\pi_L(3)$ and $\pi_U(3)$ satisfying

$$P_{\pi_L(3)}(X \geq 3) = .025; \quad P_{\pi_U(3)}(X \leq 3) = .025.$$

Many textbooks have tables for $P(X \leq c)$, where $X \sim b(n, \pi)$ for some n 's and π 's. Alternatively, $P(X \leq c)$ can be obtained using statistical software such as SAS or R. Either way, we see that $\pi_U(3) \approx .40$. To find $\pi_L(3)$ we note that

$$P_{\pi_L(3)}(X \geq 3) = 1 - P_{\pi_L(3)}(X \leq 2).$$

Consequently, we must search for $\pi_L(3)$ such that

$$P_{\pi_L(3)}(X \leq 2) = .975.$$

This yields $\pi_L(3) \approx .03$. Hence the “exact” 95% confidence interval for π is

$$[.03, .40].$$

In contrast, the normal approximation yields a confidence interval of

$$\frac{3}{19} \pm 1.96 \left(\frac{\frac{3}{19} \times \frac{16}{19}}{19} \right)^{1/2} = [-.006, .322].$$

2.2.2 Gehan's Two-Stage Design

Discarding ineffective treatments early

If it is unlikely that a treatment will achieve some minimal level of response or efficacy, we may want to stop the trial as early as possible. For example, suppose that a 20% response rate is the lowest response rate that is considered acceptable for a new treatment. If we get no responses in n patients, with n sufficiently large, then we may feel confident that the treatment is ineffective. Statistically, this may be posed as follows: How large must n be so that if there are 0 responses among n patients we are relatively confident that the response rate is not 20% or better? If $X \sim b(n, \pi)$, and if $\pi \geq .2$, then

$$P_\pi(X = 0) = (1 - \pi)^n \leq (1 - .2)^n = .8^n.$$

Choose n so that $.8^n = .05$ or $n \ln(8) = \ln(.05)$. This leads to $n \approx 14$ (rounding up). Thus, with 14 patients, it is unlikely ($\leq .05$) that no one would respond if the true response rate was greater than 20%. Thus 0 patients responding among 14 might be used as evidence to stop the phase II trial and declare the treatment a failure.

This is the logic behind Gehan's two-stage design. Gehan suggested the following strategy: If the minimal acceptable response rate is π_0 , then choose the first stage with n_0 patients such that

$$(1 - \pi_0)^{n_0} = .05; \quad n_0 = \frac{\ln(.05)}{\ln(1 - \pi_0)};$$

if there are 0 responses among the first n_0 patients then stop and declare the treatment a failure; otherwise, continue with additional patients that will ensure a certain degree of predetermined accuracy in the 95% confidence interval.

If, for example, we wanted the 95% confidence interval for the response rate to be within $\pm 15\%$ when a treatment is considered minimally effective at $\pi_0 = 20\%$, then the sample size necessary for this degree of precision is

$$1.96 \left(\frac{.2 \times .8}{n} \right)^{1/2} = .15, \quad \text{or } n = 28.$$

In this example, Gehan's design would treat 14 patients initially. If none responded, the treatment would be declared a failure and the study stopped. If there was at least one response, then another 14 patients would be treated and a 95% confidence interval for π would be computed using the data from all 28 patients.

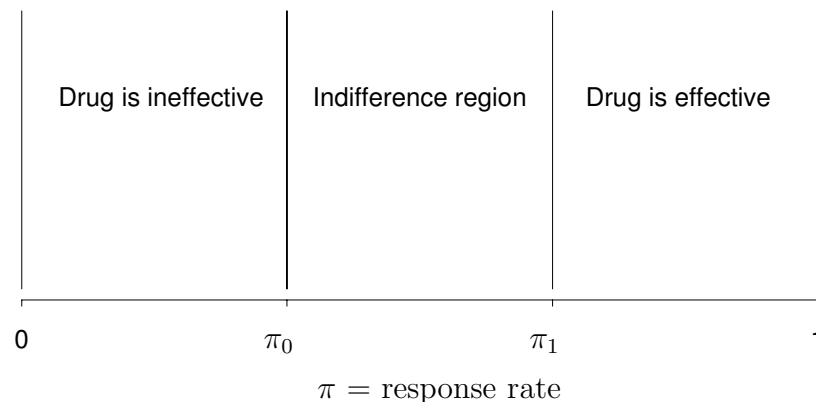
2.2.3 Simon's Two-Stage Design

Another way of using two-stage designs was proposed by Richard Simon. Here, the investigators must decide on values π_0 , and π_1 , with $\pi_0 < \pi_1$ for the probability of response so that

- If $\pi \leq \pi_0$, then we want to declare the drug ineffective with high probability, say $1 - \alpha$, where α is taken to be small.
- If $\pi \geq \pi_1$, then we want to consider this drug for further investigation with high probability, say $1 - \beta$, where β is taken to be small.

The values of α and β are generally taken to be between .05 and .20.

The region of the parameter space $\pi_0 < \pi < \pi_1$ is the indifference region.



A two-stage design would proceed as follows: Integers n_1 , n , r_1 , r , with $n_1 < n$, $r_1 < n_1$, and $r < n$ are chosen (to be described later) and

- n_1 patients are given treatment in the first stage. If r_1 or less respond, then declare the treatment a failure and stop.
- If more than r_1 respond, then add $(n - n_1)$ additional patients for a total of n patients.
- At the second stage, if the total number that respond among all n patients is greater than r , then declare the treatment a success; otherwise, declare it a failure.

Statistically, this decision rule is the following: Let X_1 denote the number of responses in the first stage (among the n_1 patients) and X_2 the number of responses in the second stage (among the $n - n_1$ patients). X_1 and X_2 are assumed to be independent binomially distributed random variables, $X_1 \sim b(n_1, \pi)$ and $X_2 \sim b(n_2, \pi)$, where $n_2 = n - n_1$ and π denotes the probability of response. Declare the treatment a failure if

$$(X_1 \leq r_1) \text{ or } \{(X_1 > r_1) \text{ and } (X_1 + X_2 \leq r)\},$$

otherwise, the treatment is declared a success if

$$\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r)\}.$$

Note: If $n_1 > r$ and if the number of patients responding in the first stage is greater than r , then there is no need to proceed to the second stage to declare the treatment a success.

According to the constraints of the problem we want

$$P(\text{declaring treatment success} | \pi \leq \pi_0) \leq \alpha,$$

or equivalently

$$P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r) | \pi = \pi_0\} \leq \alpha; \quad (2.1)$$

Note: If the above inequality is true when $\pi = \pi_0$, then it is true when $\pi < \pi_0$.

Also, we want

$$P(\text{declaring treatment failure} | \pi \geq \pi_1) \leq \beta,$$

or equivalently

$$P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r) | \pi = \pi_1\} \geq 1 - \beta. \quad (2.2)$$

Question: How are probabilities such as $P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r) | \pi\}$ computed?

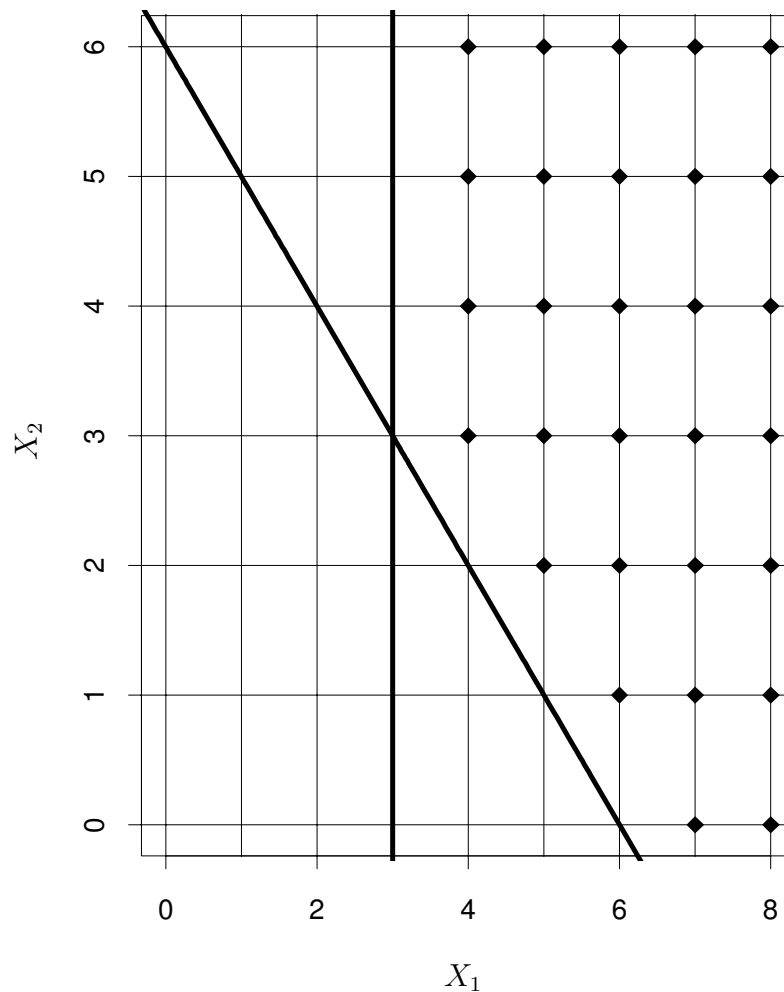
Since X_1 and X_2 are independent binomial random variables, then for any integer $0 \leq m_1 \leq n_1$ and integer $0 \leq m_2 \leq n_2$, the

$$\begin{aligned} P(X_1 = m_1, X_2 = m_2 | \pi) &= P(X_1 = m_1 | \pi) \times P(X_2 = m_2 | \pi) \\ &= \left\{ \binom{n_1}{m_1} \pi^{m_1} (1 - \pi)^{n_1 - m_1} \right\} \left\{ \binom{n_2}{m_2} \pi^{m_2} (1 - \pi)^{n_2 - m_2} \right\}. \end{aligned}$$

We then have to identify the pairs (m_1, m_2) where $(m_1 > r_1)$ and $(m_1 + m_2) > r$, find the probability for each such (m_1, m_2) pair using the equation above, and then add all the appropriate probabilities.

We illustrate this in the following figure:

Figure 2.2: *Example: $n_1 = 8, n = 14, X_1 > 3,$ and $X_1 + X_2 > 6$*



As it turns out there are many combinations of (r_1, n_1, r, n) that satisfy the constraints (??) and (??) for specified $(\pi_0, \pi_1, \alpha, \beta)$. Through a computer search one can find the “**optimal design**” among these possibilities, where the optimal design is defined as the combination (r_1, n_1, r, n) , satisfying the constraints (??) and (??), which gives the smallest expected sample size when

$$\pi = \pi_0.$$

The expected sample size for a two stage design is defined as

$$n_1 P(\text{stopping at the first stage}) + n P(\text{stopping at the second stage}).$$

For our problem, the expected sample size is given by

$$n_1 \{P(X_1 \leq r_1 | \pi = \pi_0) + P(X_1 > r | \pi = \pi_0)\} + n P(r_1 + 1 \leq X_1 \leq r | \pi = \pi_0).$$

Optimal two-stage designs have been tabulated for a variety of $(\pi_0, \pi_1, \alpha, \beta)$ in the article

Simon, R. (1989). Optimal two-stage designs for Phase II clinical trials. *Controlled Clinical Trials*. 10: 1-10.

The tables are given on the next two pages.

Table 1 Designs for $p_1 - p_0 = 0.20^a$

p_0	p_1	Optimal Design				Minimax Design			
		Reject Drug if Response Rate		EN(p_0)	PET(p_0)	Reject Drug if Response Rate		EN(p_0)	PET(p_0)
		$\leq r_1/n_1$	$\leq r/n$			$\leq r_1/n_1$	$\leq r/n$		
0.05	0.25	0/9	2/24	14.5	0.63	0/13	2/20	16.4	0.51
		0/9	2/17	12.0	0.63	0/12	2/16	13.8	0.54
		0/9	3/30	16.8	0.63	0/15	3/25	20.4	0.46
0.10	0.30	1/12	5/35	19.8	0.65	1/16	4/25	20.4	0.51
		1/10	5/29	15.0	0.74	1/15	5/25	19.5	0.55
		2/18	6/35	22.5	0.71	2/22	6/33	26.2	0.62
0.20	0.40	3/17	10/37	26.0	0.55	3/19	10/36	28.3	0.46
		3/13	12/43	20.6	0.75	4/18	10/33	22.3	0.50
		4/19	15/54	30.4	0.67	5/24	13/45	31.2	0.66
0.30	0.50	7/22	17/46	29.9	0.67	7/28	15/39	35.0	0.36
		5/15	18/46	23.6	0.72	6/19	16/39	25.7	0.48
		8/24	24/63	34.7	0.73	7/24	21/53	36.6	0.56
0.40	0.60	7/18	22/46	30.2	0.56	11/28	20/41	33.8	0.55
		7/16	23/46	24.5	0.72	17/34	20/39	34.4	0.91
		11/25	32/66	36.0	0.73	12/29	27/54	38.1	0.64
0.50	0.70	11/21	26/45	29.0	0.67	11/23	23/39	31.0	0.50
		8/15	26/43	23.5	0.70	12/23	23/37	27.7	0.66
		13/24	36/61	34.0	0.73	14/27	32/53	36.1	0.65
0.60	0.80	6/11	26/38	25.4	0.47	18/27	24/35	28.5	0.82
		7/11	30/43	20.5	0.70	8/13	25/35	20.8	0.65
		12/19	37/53	29.5	0.69	15/26	32/45	35.9	0.48
0.70	0.90	6/9	22/28	17.8	0.54	11/16	20/25	20.1	0.55
		4/6	22/27	14.8	0.58	19/23	21/26	23.2	0.95
		11/15	29/36	21.2	0.70	13/18	26/32	22.7	0.67

^aFor each value of (p_0, p_1) , designs are given for three sets of error probabilities (α, β) . The first, second and third rows correspond to error probability limits $(0.10, 0.10)$, $(0.05, 0.20)$, and $(0.05, 0.10)$ respectively. For each design, EN(p_0) and PET(p_0) denote the expected sample size and the probability of early termination when the true response probability is p_0 .

Table 2 Designs for $p_1 - p_0 = 0.15^a$

p_0	p_1	Optimal Design				Minimax Design			
		Reject Drug if Response Rate		$EN(p_0)$	$PET(p_0)$	Reject Drug if Response Rate		$EN(p_0)$	$PET(p_0)$
$\leq r_1/n_1$	$\leq r/n$	$\leq r_1/n_1$	$\leq r/n$						
0.05	0.20	0/12	3/37	23.5	0.54	0/18	3/32	26.4	0.40
		0/10	3/29	17.6	0.60	0/13	3/27	19.8	0.51
		1/21	4/41	26.7	0.72	1/29	4/38	32.9	0.57
0.10	0.25	2/21	7/50	31.2	0.65	2/27	6/40	33.7	0.48
		2/18	7/43	24.7	0.73	2/22	7/40	28.8	0.62
		2/21	10/66	36.8	0.65	3/31	9/55	40.0	0.62
0.20	0.35	5/27	16/63	43.6	0.54	6/33	15/58	45.5	0.50
		5/22	19/72	35.4	0.73	6/31	15/53	40.4	0.57
		8/37	22/83	51.4	0.69	8/42	21/77	58.4	0.53
0.30	0.45	9/30	29/82	51.4	0.59	16/50	25/69	56.0	0.68
		9/27	30/81	41.7	0.73	16/46	25/65	49.6	0.81
		13/40	40/110	60.8	0.70	27/77	33/88	78.5	0.86
0.40	0.55	16/38	40/88	54.5	0.67	18/45	34/73	57.2	0.56
		11/26	40/84	44.9	0.67	28/59	34/70	60.1	0.90
		19/45	49/104	64.0	0.68	24/62	45/94	78.9	0.47
0.50	0.65	18/35	47/84	53.0	0.63	19/40	41/72	58.0	0.44
		15/28	48/83	43.7	0.71	39/66	40/68	66.1	0.95
		22/42	60/105	62.3	0.68	28/57	54/93	75.0	0.50
0.60	0.75	21/34	47/71	47.1	0.65	25/43	43/64	54.4	0.46
		17/27	46/67	39.4	0.69	18/30	43/62	43.8	0.57
		21/34	64/95	55.6	0.65	48/72	57/84	73.2	0.90
0.70	0.85	14/20	45/59	36.2	0.58	15/22	40/52	36.8	0.51
		14/19	46/59	30.3	0.72	16/23	39/49	34.4	0.56
		18/25	61/79	43.4	0.66	33/44	53/68	48.5	0.81
0.80	0.95	5/7	27/31	20.8	0.42	5/7	27/31	20.8	0.42
		7/9	26/29	17.7	0.56	7/9	26/29	17.7	0.56
		16/19	37/42	24.4	0.76	31/35	35/40	35.3	0.94

^aFor each value of (p_0, p_1) , designs are given for three sets of error probabilities (α, β) . The first, second, and third rows correspond to error probability limits $(0.10, 0.10)$, $(0.05, 0.20)$, and $(0.05, 0.10)$ respectively. For each design, $EN(p_0)$ and $PET(p_0)$ denote the expected sample size and the probability of early termination when the true response probability is p_0 .

3 Phase III Clinical Trials

3.1 Why are clinical trials needed

A clinical trial is the clearest method of determining whether an intervention has the postulated effect. It is very easy for anecdotal information about the benefit of a therapy to be accepted and become standard of care. The consequence of not conducting appropriate clinical trials can be serious and costly. As we discussed earlier, because of anecdotal information, blood-letting was common practice for a very long time. Other examples include

- It was believed that high concentrations of oxygen was useful for therapy in premature infants until a clinical trial demonstrated its harm
- Intermittent positive pressure breathing became an established therapy for chronic obstructive pulmonary disease (COPD). Much later, a clinical trial suggested no major benefit for this very expensive procedure
- Laetrile (a drug extracted from grapefruit seeds) was rumored to be the wonder drug for Cancer patients even though there was no scientific evidence that this drug had any biological activity. People were so convinced that there was a conspiracy by the medical profession to withhold this drug that they would get it illegally from “quacks” or go to other countries such as Mexico to get treatment. The use of this drug became so prevalent that the National Institutes of Health finally conducted a clinical trial where they proved once and for all that Laetrile had no effect. You no longer hear about this issue any more.
- The Cardiac Antiarrhythmia Suppression Trial (CAST) documented that commonly used antiarrhythmia drugs were harmful in patients with myocardial infarction
- More recently, against common belief, it was shown that prolonged use of Hormone Replacement Therapy for women following menopause may have deleterious effects.

3.2 Issues to consider before designing a clinical trial

David Sackett gives the following six prerequisites

1. The trial needs to be done
 - (i) the disease must have either high incidence and/or serious course and poor prognosis
 - (ii) existing treatment must be unavailable or somehow lacking
 - (iii) The intervention must have promise of efficacy (pre-clinical as well as phase I-II evidence)
2. The trial question posed must be appropriate and unambiguous
3. The trial architecture is valid. Random allocation is one of the best ways that treatment comparisons made in the trial are valid. Other methods such as blinding and placebos should be considered when appropriate
4. The inclusion/exclusion criteria should strike a balance between efficiency and generalizability. Entering patients at high risk who are believed to have the best chance of response will result in an efficient study. This subset may however represent only a small segment of the population of individuals with disease that the treatment is intended for and thus reduce the study's generalizability
5. The trial **protocol** is feasible
 - (i) The protocol must be attractive to potential investigators
 - (ii) Appropriate types and numbers of patients must be available
6. The trial administration is effective.

Other issues that also need to be considered

- Applicability: Is the intervention likely to be implemented in practice?
- Expected size of effect: Is the intervention “strong enough” to have a good chance of producing a detectable effect?

- Obsolescence: Will changes in patient management render the results of a trial obsolete before they are available?

Objectives and Outcome Assessment

- Primary objective: What is the primary question to be answered?
 - ideally just one
 - important, relevant to care of future patients
 - capable of being answered
- Primary outcome (endpoint)
 - ideally just one
 - relatively simple to analyze and report
 - should be well defined; objective measurement is preferred to a subjective one. For example, clinical and laboratory measurements are more objective than say clinical and patient impression
- Secondary Questions
 - other outcomes or endpoints of interest
 - subgroup analyses
 - secondary questions should be viewed as exploratory
 - * trial may lack power to address them
 - * multiple comparisons will increase the chance of finding “statistically significant” differences even if there is no effect
 - avoid excessive evaluations; as well as problem with multiple comparisons, this may affect data quality and patient support

Choice of Primary Endpoint

Example: Suppose we are considering a study to compare various treatments for patients with HIV disease, then what might be the appropriate primary endpoint for such a study? Let us look at some options and discuss them.

The HIV virus destroys the immune system; thus individuals infected are susceptible to various opportunistic infections which ultimately leads to death. Many of the current treatments are designed to target the virus either trying to destroy it or, at least, slow down its replication. Other treatments may target specific opportunistic infections.

Suppose we have a treatment intended to attack the virus directly, Here are some possibilities for the primary endpoint that we may consider.

1. Increase in CD4 count. Since CD4 count is a direct measure of the immune function and CD4 cells are destroyed by the virus, we might expect that a good treatment will increase CD4 count.
2. Viral RNA reduction. Measures the amount of virus in the body
3. Time to the first opportunistic infection
4. Time to death from any cause
5. Time to death or first opportunistic infection, whichever comes first

Outcomes 1 and 2 may be appropriate as the primary outcome in a phase II trial where we want to measure the activity of the treatment as quickly as possible.

Outcome 4 may be of ultimate interest in a phase III trial, but may not be practical for studies where patients have a long expected survival and new treatments are being introduced all the time. (Obsolescence)

Outcome 5 may be the most appropriate endpoint in a phase III trial. However, the other outcomes may be reasonable for secondary analyses.

3.3 Ethical Issues

A clinical trial involves human subjects. As such, we must be aware of ethical issues in the design and conduct of such experiments. Some ethical issues that need to be considered include the following:

- No alternative which is superior to any trial intervention is available for each subject
- Equipoise—There should be genuine uncertainty about which trial intervention may be superior for each individual subject before a physician is willing to allow their patients to participate in such a trial
- Exclude patients for whom risk/benefit ratio is likely to be unfavorable
 - pregnant women if possibility of harmful effect to the fetus
 - too sick to benefit
 - if prognosis is good without interventions

Justice Considerations

- Should not exclude a class of patients for non medical reasons nor unfairly recruit patients from poorer or less educated groups

This last issue is a bit tricky as “equal access” may hamper the evaluation of interventions. For example

- Elderly people may die from diseases other than that being studied
- IV drug users are more difficult to follow in AIDS clinical trials

3.4 The Randomized Clinical Trial

The objective of a clinical trial is to evaluate the effects of an intervention. Evaluation implies that there must be some comparison either to

- no intervention
- placebo
- best therapy available

Fundamental Principle in Comparing Treatment Groups

Groups must be alike in all important aspects and only differ in the treatment which each group receives. Otherwise, differences in response between the groups may not be due to the treatments under study, but can be attributed to the particular characteristics of the groups.

How should the control group be chosen

Here are some examples:

- Literature controls
- Historical controls
- Patient as his/her own control (cross-over design)
- Concurrent control (non-randomized)
- Randomized concurrent control

The difficulty in non-randomized clinical trials is that the control group may be different prognostically from the intervention group. Therefore, comparisons between the intervention and control groups may be biased. That is, differences between the two groups may be due to factors other than the treatment.

Attempts to correct the bias that may be induced by these confounding factors either by design (matching) or by analysis (adjustment through stratified analysis or regression analysis) may not be satisfactory.

To illustrate the difficulty with non-randomized controls, we present results from 12 different studies, all using the same treatment of 5-FU on patients with advanced carcinoma of the large bowel.

Table 3.1: *Results of Rapid Injection of 5-FU for Treatment of Advanced Carcinoma of the Large Bowel*

Group	# of Patients	% Objective Response
1. Sharp and Benefiel	13	85
2. Rochlin et al.	47	55
3. Cornell et al.	13	46
4. Field	37	41
5. Weiss and Jackson	37	35
6. Hurley	150	31
7. ECOG	48	27
8. Brennan et al.	183	23
9. Ansfield	141	17
10. Ellison	87	12
11. Knoepp et al.	11	9
12. Olson and Greene	12	8

Suppose there is a new treatment for advanced carcinoma of the large bowel that we want to compare to 5-FU. We decide to conduct a new study where we treat patients only with the new drug and compare the response rate to the historical controls. At first glance, it looks as if the response rates in the above table vary tremendously from study to study even though all these used the same treatment 5-FU. If this is indeed the case, then what comparison can possibly be made if we want to evaluate the new treatment against 5-FU? It may be possible, however, that the response rates from study to study are consistent with each other and the differences we are seeing come from random sampling fluctuations. This is important because if we believe there is no study to study variation, then we may feel confident in conducting a new study using only

the new treatment and comparing the response rate to the pooled response rate from the studies above. How can we assess whether these differences are random sampling fluctuations or real study to study differences?

Hierarchical Models

To address the question of whether the results from the different studies are random samples from underlying groups with a common response rate or from groups with different underlying response rates, we introduce the notion of a hierarchical model. In a hierarchical model, we assume that each of the N studies that were conducted were from possibly N different study groups each of which have possibly different underlying response rates π_1, \dots, π_N . In a sense, we now think of the world as being made of many different study groups (or a population of study groups), each with its own response rate, and that the studies that were conducted correspond to choosing a small sample of these population study groups. As such, we imagine π_1, \dots, π_N to be a random sample of study-specific response rates from a larger population of study groups. Since π_i , the response rate from the i -th study group, is a random variable, it has a mean and a variance which we will denote by μ_π and σ_π^2 . Since we are imagining a super-population of study groups, each with its own response rate, that we are sampling from, we conceptualize μ_π and σ_π^2 to be the average and variance of these response rates from this super-population. Thus π_1, \dots, π_N will correspond to an iid (independent and identically distributed) sample from a population with mean μ_π and variance σ_π^2 . I.e.

$$\pi_1, \dots, \pi_N, \text{ are iid with } E(\pi_i) = \mu_\pi, \text{var}(\pi_i) = \sigma_\pi^2, i = 1, \dots, N.$$

This is the first level of the hierarchy.

The second level of the hierarchy corresponds now to envisioning that the data collected from the i -th study (n_i, X_i) , where n_i is the number of patients treated in the i -th study and X_i is the number of complete responses among the n_i treated, is itself a random sample from the i -th study group whose response rate is π_i . That is, conditional on n_i and π_i , X_i is assumed to follow a binomial distribution, which we denote as

$$X_i | n_i, \pi_i \sim b(n_i, \pi_i).$$

This hierarchical model now allows us to distinguish between random sampling fluctuation and real study to study differences. If all the different study groups were homogeneous, then there

should be no study to study variation, in which case $\sigma_\pi^2 = 0$. Thus we can evaluate the degree of study to study differences by estimating the parameter σ_π^2 .

In order to obtain estimates for σ_π^2 , we shall use some classical results of conditional expectation and conditional variance. Namely, if X and Y denote random variables for some probability experiment then the following is true

$$E(X) = E\{E(X|Y)\}$$

and

$$\text{var}(X) = E\{\text{var}(X|Y)\} + \text{var}\{E(X|Y)\}.$$

Although these results are known to many of you in the class; for completeness, I will sketch out the arguments why the two equalities above are true.

3.5 Review of Conditional Expectation and Conditional Variance

For simplicity, I will limit myself to probability experiments with a finite number of outcomes. For random variables that are continuous one needs more complicated measure theory for a rigorous treatment.

Probability Experiment

Denote the result of an experiment by one of the outcomes in the sample space $\Omega = \{\omega_1, \dots, \omega_k\}$. For example, if the experiment is to choose one person at random from a population of size N with a particular disease, then the result of the experiment is $\Omega = \{A_1, \dots, A_N\}$ where the different A 's uniquely identify the individuals in the population, If the experiment were to sample n individuals from the population then the outcomes would be all possible n -tuple combinations of these N individuals; for example $\Omega = \{(A_{i_1}, \dots, A_{i_n}), \text{ for all } i_1, \dots, i_n = 1, \dots, N$. With replacement there are $k = N^n$ combinations; without replacement there are $k = N \times (N-1) \times \dots \times (N-n+1)$ combinations of outcomes if order of subjects in the sample is important, and $k = \binom{N}{n}$ combinations of outcomes if order is not important.

Denote by $p(\omega)$ the probability of outcome ω occurring, where $\sum_{\omega \in \Omega} p(\omega) = 1$.

Random variable

A random variable, usually denoted by a capital Roman letter such as X, Y, \dots is a function that assigns a number to each outcome in the sample space. For example, in the experiment where we sample one individual from the population

$X(\omega)$ = survival time for person ω

$Y(\omega)$ = blood pressure for person ω

$Z(\omega)$ = height of person ω

The **probability distribution** of a random variable X is just a list of all different possible values that X can take together with the corresponding probabilities.

i.e. $\{(x, P(X = x))\}$, for all possible x , where $P(X = x) = \sum_{\omega: X(\omega)=x} p(\omega)$.

The **mean** or **expectation** of X is

$$E(X) = \sum_{\omega \in \Omega} X(\omega)p(\omega) = \sum_x xP(X = x),$$

and the **variance** of X is

$$\begin{aligned} \text{var}(X) &= \sum_{\omega \in \Omega} \{X(\omega) - E(X)\}^2 p(\omega) = \sum_x \{x - E(X)\}^2 P(X = x) \\ &= E\{X - E(X)\}^2 = E(X^2) - \{E(X)\}^2. \end{aligned}$$

Conditional Expectation

Suppose we have two random variables X and Y defined for the same probability experiment, then we denote the conditional expectation of X , conditional on knowing that $Y = y$, by $E(X|Y = y)$ and this is computed as

$$E(X|Y = y) = \sum_{\omega: Y(\omega)=y} X(\omega) \frac{p(\omega)}{P(Y = y)}.$$

The conditional expectation of X given Y , denoted by $E(X|Y)$ is itself a random variable which assigns the value $E(X|Y = y)$ to every outcome ω for which $Y(\omega) = y$. Specifically, we note that $E(X|Y)$ is a function of Y .

Since $E(X|Y)$ is itself a random variable, it also has an expectation given by $E\{E(X|Y)\}$. By the definition of expectation this equals

$$E\{E(X|Y)\} = \sum_{\omega \in \Omega} E(X|Y)(\omega)p(\omega).$$

By rearranging this sum, first within the partition $\{\omega : Y(\omega) = y\}$, and then across the partitions for different values of y , we get

$$\begin{aligned} E\{E(X|Y)\} &= \sum_y \left\{ \frac{\sum_{\omega: Y(\omega)=y} X(\omega)p(\omega)}{P(Y=y)} \right\} P(Y=y) \\ &= \sum_{\omega \in \Omega} X(\omega)p(\omega) = E(X). \end{aligned}$$

Thus we have proved the very important result that

$$E\{E(X|Y)\} = E(X).$$

Conditional Variance

There is also a very important relationship involving conditional variance. Just like conditional expectation, the conditional variance of X given Y , denoted as $\text{var}(X|Y)$, is a random variable, which assigns the value $\text{var}(X|Y=y)$ to each outcome ω , where $Y(\omega) = y$, and

$$\text{var}(X|Y=y) = E[\{X - E(X|Y=y)\}^2|Y=y] = \sum_{\omega: Y(\omega)=y} \{X(\omega) - E(X|Y=y)\}^2 \frac{p(\omega)}{p(Y=y)}.$$

Equivalently,

$$\text{var}(X|Y=y) = E(X^2|Y=y) - \{E(X|Y=y)\}^2.$$

It turns out that the variance of a random variable X equals

$$\text{var}(X) = E\{\text{var}(X|Y)\} + \text{var}\{E(X|Y)\}.$$

This follows because

$$E\{\text{var}(X|Y)\} = E[E(X^2|Y) - \{E(X|Y)\}^2] = E(X^2) - E[\{E(X|Y)\}^2] \quad (3.1)$$

and

$$\text{var}\{E(X|Y)\} = E[\{E(X|Y)\}^2] - [E\{E(X|Y)\}]^2 = E[\{E(X|Y)\}^2] - \{E(X)\}^2 \quad (3.2)$$

Adding (??) and (??) together yields

$$E\{var(X|Y)\} + var\{E(X|Y)\} = E(X^2) - \{E(X)\}^2 = var(X),$$

as desired.

If we think of partitioning the sample space into regions $\{\omega : Y(\omega) = y\}$ for different values of y , then the formula above can be interpreted in words as

“the variance of X is equal to the mean of the within partition variances of X plus the variance of the within partition means of X ”. This kind of partitioning of variances is often carried out in ANOVA models.

Return to Hierarchical Models

Recall

$$X_i | n_i, \pi_i \sim b(n_i, \pi_i), \quad i = 1, \dots, N$$

and

$$\pi_1, \dots, \pi_N \text{ are iid } (\mu_\pi, \sigma_\pi^2).$$

Let $p_i = X_i/n_i$ denote the sample proportion that respond from the i -th study. We know from properties of a binomial distribution that

$$E(p_i | \pi_i, n_i) = \pi_i$$

and

$$var(p_i | \pi_i, n_i) = \frac{\pi_i(1 - \pi_i)}{n_i}.$$

Note: In our conceptualization of this problem the probability experiment consists of

1. Conducting N studies from a population of studies
2. For each study i we sample n_i individuals at random from the i -th study group and count the number of responses X_i
3. Let us also assume that the sample sizes n_1, \dots, n_N are random variables from some distribution.

4. The results of this experiment can be summarized by the iid random vectors

$$(\pi_i, n_i, X_i), \quad i = 1, \dots, N.$$

In actuality, we don't get to see the values $\pi_i, i = 1, \dots, N$. They are implicitly defined, yet very important in the description of the model. Often, the values π_i are referred to as random effects. Thus, the observable data we get to work with are

$$(n_i, X_i), \quad i = 1, \dots, N.$$

Using the laws of iterated conditional expectation and variance just derived, we get the following results:

$$E(p_i) = E\{E(p_i|n_i, \pi_i)\} = E(\pi_i) = \mu_\pi, \quad (3.3)$$

$$\begin{aligned} \text{var}(p_i) &= E\{\text{var}(p_i|n_i, \pi_i)\} + \text{var}\{E(p_i|n_i, \pi_i)\} \\ &= E\left\{\frac{\pi_i(1-\pi_i)}{n_i}\right\} + \text{var}(\pi_i) \\ &= E\left\{\frac{\pi_i(1-\pi_i)}{n_i}\right\} + \sigma_\pi^2. \end{aligned} \quad (3.4)$$

Since the random variables $p_i, i = 1, \dots, N$ are iid, an unbiased estimator for $E(p_i) = \mu_\pi$ is given by the sample mean

$$\bar{p} = N^{-1} \sum_{i=1}^N p_i,$$

and an unbiased estimator of the variance $\text{var}(p_i)$ is the sample variance

$$s_p^2 = \frac{\sum_{i=1}^N (p_i - \bar{p})^2}{N-1}.$$

One can also show, using properties of a binomial distribution, that a conditionally unbiased estimator for $\frac{\pi_i(1-\pi_i)}{n_i}$, conditional on n_i and π_i , is given by $\frac{p_i(1-p_i)}{n_i-1}$; that is

$$E\left\{\frac{p_i(1-p_i)}{n_i-1} \middle| n_i, \pi_i\right\} = \frac{\pi_i(1-\pi_i)}{n_i}.$$

I will leave this as a homework exercise for you to prove.

Since $\frac{p_i(1-p_i)}{n_i-1}, i = 1, N$ are iid random variables with mean

$$E\left\{\frac{p_i(1-p_i)}{n_i-1}\right\} = E\left[E\left\{\frac{p_i(1-p_i)}{n_i-1} \middle| n_i, \pi_i\right\}\right]$$

$$= E \left\{ \frac{\pi_i(1 - \pi_i)}{n_i} \right\},$$

this means that we can obtain an unbiased estimator for $E \left\{ \frac{\pi_i(1 - \pi_i)}{n_i} \right\}$ by using

$$N^{-1} \sum_{i=1}^N \frac{p_i(1 - p_i)}{n_i - 1}.$$

Summarizing these results, we have shown that

- $s_p^2 = \frac{\sum_{i=1}^N (p_i - \bar{p})^2}{N-1}$ is an unbiased estimator for $var(p_i)$ which by (??) equals

$$E \left\{ \frac{\pi_i(1 - \pi_i)}{n_i} \right\} + \sigma_\pi^2$$

- We have also shown that $N^{-1} \sum_{i=1}^N \frac{p_i(1 - p_i)}{n_i - 1}$ is an unbiased estimator for

$$E \left\{ \frac{\pi_i(1 - \pi_i)}{n_i} \right\}$$

Consequently, by subtraction, we get that the estimator

$$\hat{\sigma}_\pi^2 = \left\{ \frac{\sum_{i=1}^N (p_i - \bar{p})^2}{N - 1} \right\} - \left\{ N^{-1} \sum_{i=1}^N \frac{p_i(1 - p_i)}{n_i - 1} \right\}$$

is an unbiased estimator for σ_π^2 .

Going back to the example given in Table 3.1, we obtain the following:

-

$$\frac{\sum_{i=1}^N (p_i - \bar{p})^2}{N - 1} = .0496$$

-

$$N^{-1} \sum_{i=1}^N \frac{p_i(1 - p_i)}{n_i - 1} = .0061$$

- Hence

$$\hat{\sigma}_\pi^2 = .0496 - .0061 = .0435$$

Thus the estimate for study to study standard deviation in the probability of response is given by

$$\hat{\sigma}_\pi = \sqrt{.0435} = .21.$$

This is an enormous variation clearly indicating substantial study to study variation.