MITx: Statistics, Computation & Applications

Statistics Refresher

Lecture 1: Observational Studies and Experiments

Mammography and breast cancer

- Breast cancer is one of the most common malignancies among women in the United States
- Mammography: screening women for breast cancer by X-rays

✶ Does mammography speed up detection by enough to matter?

✶ How would you approach this problem? What is important when setting up a study / experiment?

Mammography and breast cancer

- Breast cancer is one of the most common malignancies among women in the United States
- Mammography: screening women for breast cancer by X-rays

✶ Does mammography speed up detection by enough to matter?

✶ How would you approach this problem? What is important when setting up a study / experiment?

⇒ Perform a **randomized, controlled, double-blind experiment** to minimize the problem of **confounding**

HIP study: First large-scale randomized controlled experiment on mammography performed in 1960s

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

Reference: D. A. Freedman. *Statistical Models: Theory and Practice*, 2009.

HIP study

Which rates should be compared to show the efficacy of treatment?

- Seems natural to compare those who accepted screening to those who refused
- But this is an **observational** comparison!
- Becomes clear when comparing the death rates from all other causes
- Instead compare the whole treatment group against the whole control group

✶ **Intention-to-treat analysis**

Hypothesis testing

- Death rate from breast cancer in control group: 0.0020 ($= \frac{63}{31000}$)
- Death rate from breast cancer in treatment group: 0.0013 ($= \frac{39}{31000}$)

Is the difference in death rates between the treatment and control group sufficient to establish that mammography reduces the risk of death from breast cancer?

$\Rightarrow$ Perform a **hypothesis test**

Hypothesis testing

1. Determine a **model**:

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \mathrm{Bernoulli}(\pi)$ or $Y \sim \mathrm{Poisson}(\lambda)$

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \mathrm{Bernoulli}(\pi)$ or $Y \sim \mathrm{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \mathrm{Bernoulli}(\pi)$ or $Y \sim \mathrm{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$

Hypothesis testing

1. Determine a **model**:

   $X_1, \dots, X_{31'000} \sim \mathrm{Bernoulli}(\pi)$ or $Y \sim \mathrm{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$

3. Determine a **test statistic** (quantity that can differentiate between H_0 and H_A, and whose distribution under H_0 you can compute):

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$

3. Determine a **test statistic** (quantity that can differentiate between H_0 and H_A, and whose distribution under H_0 you can compute):

   $T :=$ Number of deaths under H_0:
   $T \sim \text{binomial}(31'000, 0.002)$ or $T \sim \text{Poisson}(63)$

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

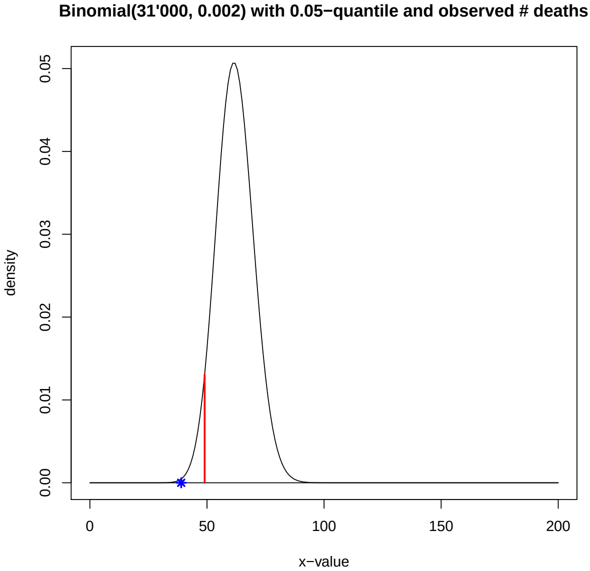
   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$

3. Determine a **test statistic** (quantity that can differentiate between H_0 and H_A, and whose distribution under H_0 you can compute):

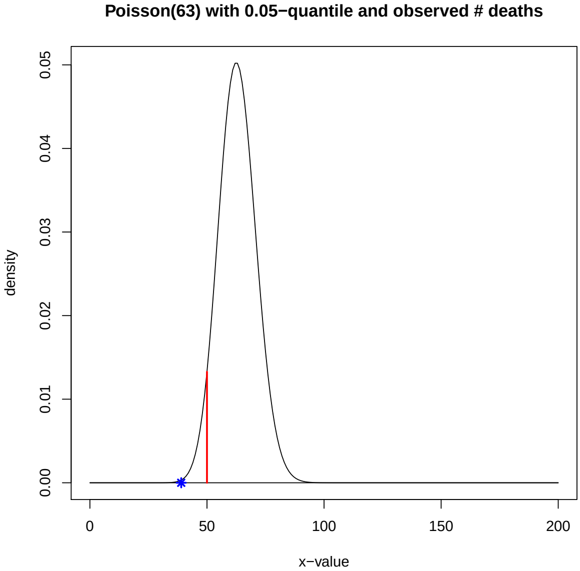
   $T :=$ Number of deaths under H_0:
   $T \sim \text{binomial}(31'000, 0.002)$ or $T \sim \text{Poisson}(63)$

4. Determine a **significance level** (α), i.e. the probability of rejecting H_0 when H_0 is true: e.g. $\alpha = 0.05$

Binomial distribution



Poisson distribution



P-value

- Probability under H_0 to obtain the observed value or a more extreme value of the test statistic

  $\Rightarrow$ p-value is always between 0 and 1!

  For mammography study: p-value is 0.0012 under binomial model and 0.0008 under Poisson model

- Smallest significance level for which H_0 just gets rejected

- Can be used for hypothesis testing: Reject H_0 if p-value $\leq \alpha$

- Quantifies significance of alternative

Power

	retain H_0	reject H_0
H_0 true	-	type I error
H_A true	type II error	-

Power

	retain H_0	reject H_0
H_0 true	-	type I error
H_A true	type II error	-

- Signficiance level bounds probability of type I error: $\mathbb{P}(\text{type I error}) \leq \alpha$
- Power $:= 1 - \mathbb{P}(\text{type II error})$

Power

	retain H_0	reject H_0
H_0 true	-	type I error
H_A true	type II error	-

- Signficiance level bounds probability of type I error: $\mathbb{P}(\text{type I error}) \leq \alpha$
- Power $:= 1 - \mathbb{P}(\text{type II error})$
- Note that there is a trade-off between the probability of making a type I error and the probability of making a type II error (Why?)
- Note that power of 1-sided test is usually higher than for 2-sided test (Why?)

  $\Rightarrow$ Perform 1-sided test if you are only interested in detecting deviations in one direction

Recap: Hypothesis testing

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

1. Model: $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$
2. Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$
3. Test statistic T = Number of deaths under H_0
   $T \sim \text{binomial}(31'000, 0.002)$ or $T \sim \text{Poisson}(63)$
4. Significance level: $\alpha = 0.05$

Any important assumption that we should relax?

Alternative test: assume no knowledge of π_{control}

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30’961	31’000
control	63 (0.0020)	30’937	31’000
total	102	61’898	62’000

Null hypothesis: $\pi_{\text{control}} = \pi_{\text{treatment}}$ Alternative: $\pi_{\text{control}} > \pi_{\text{treatment}}$

Alternative test: assume no knowledge of π_{control}

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

Null hypothesis: $\pi_{\text{control}} = \pi_{\text{treatment}}$ Alternative: $\pi_{\text{control}} > \pi_{\text{treatment}}$

Knowing that 102 subjects died and that number of treatments / controls is 31'000, what is probability that deaths are so unevenly distributed?

Alternative test: assume no knowledge of π_{control}

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

Null hypothesis: $\pi_{\text{control}} = \pi_{\text{treatment}}$ Alternative: $\pi_{\text{control}} > \pi_{\text{treatment}}$

Knowing that 102 subjects died and that number of treatments / controls is 31'000, what is probability that deaths are so unevenly distributed?

- Test statistic T: number of deaths among the treated individuals
- Model: **Hypergeometric distribution**:

$$\mathbb{P}_{H_0}(T = 39) = \frac{\binom{31'000}{39}\binom{31'000}{63}}{\binom{62'000}{102}}$$

- p-value $= 0.011$

Fisher's exact test

Test on previous slide is known as **Fisher's exact test**

Fisher's exact test

Test on previous slide is known as **Fisher's exact test**

Advantages:

- Does not assume knowledge about the true probability of dying due to breast cancer in the control population

Fisher's exact test

Test on previous slide is known as **Fisher's exact test**

Advantages:

- Does not assume knowledge about the true probability of dying due to breast cancer in the control population

Shortcomings:

- Assumes knowledge of the margins (i.e., row and column sums)
- Alternative is Bernard's test (estimates the margins)

  for more details on this test, see e.g. `http://www.nbi.dk/~petersen/Teaching/Stat2009/Barnard_ExactTest_TwoBinomials.pdf`
- Both tests are difficult to perform on large tables for computational reasons

References

- For a statistics review, including controlled experiments and observational studies (chapters 1 and 2) and hypothesis testing (chapter 26-29):

  D. Freedman, R. Pisani, R. Purves. *Statistics*. 2007.

- For how to perform hypothesis testing in `R` (chapter 4):

  P. Dalgaard. *Introductory Statistics with R*. 2002.

- For observational studies and experiments, including the HIP study (chapter 1):

  D. Freedman. *Statistical Models: Theory and Practice*. 2009.

MITx: Statistics, Computation & Applications

Statistics Refresher

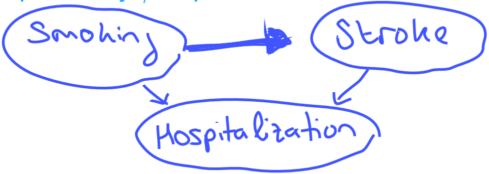
Lecture 1: Observational Studies and Experiments

Mammography and breast cancer

- Breast cancer is one of the most common malignancies among women in the United States
- Mammography: screening women for breast cancer by X-rays

✶ Does mammography speed up detection by enough to matter?

✶ How would you approach this problem? What is important when setting up a study / experiment?



Mammography and breast cancer

- Breast cancer is one of the most common malignancies among women in the United States
- Mammography: screening women for breast cancer by X-rays

✶ Does mammography speed up detection by enough to matter?

✶ How would you approach this problem? What is important when setting up a study / experiment?

⇒ Perform a **randomized, controlled, double-blind experiment** to minimize the problem of **confounding**

HIP study: First large-scale randomized controlled experiment on mammography performed in 1960s

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer		All other	
		No.	Rate	No.	Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

Reference: D. A. Freedman. *Statistical Models: Theory and Practice*, 2009.

HIP study

Which rates should be compared to show the efficacy of treatment?

- Seems natural to compare those who accepted screening to those who refused
- But this is an **observational** comparison!
- Becomes clear when comparing the death rates from all other causes
- Instead compare the whole treatment group against the whole control group

✶ **Intention-to-treat analysis**

Hypothesis testing

- Death rate from breast cancer in control group: 0.0020 ($= \frac{63}{31000}$)
- Death rate from breast cancer in treatment group: 0.0013 ($= \frac{39}{31000}$)

Is the difference in death rates between the treatment and control group sufficient to establish that mammography reduces the risk of death from breast cancer?

$\Rightarrow$ Perform a **hypothesis test**

Mammography and breast cancer

- Breast cancer is one of the most common malignancies among women in the United States
- Mammography: screening women for breast cancer by X-rays

✶ Does mammography speed up detection by enough to matter?

✶ How would you approach this problem? What is important when setting up a study / experiment?

⇒ Perform a **randomized, controlled, double-blind experiment** to minimize the problem of **confounding**

HIP study: First large-scale randomized controlled experiment on mammography performed in 1960s

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

Reference: D. A. Freedman. *Statistical Models: Theory and Practice*, 2009.

Hypothesis testing

- Death rate from breast cancer in control group: 0.0020 ($= \frac{63}{31000}$)
- Death rate from breast cancer in treatment group: 0.0013 ($= \frac{39}{31000}$)

Is the difference in death rates between the treatment and control group sufficient to establish that mammography reduces the risk of death from breast cancer?

$\Rightarrow$ Perform a **hypothesis test**

Hypothesis testing

1. Determine a **model**:

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

Hypothesis testing

1. Determine a **model**:

   $$X_1, \ldots, X_{31'000} \sim \mathrm{Bernoulli}(\pi) \quad \text{or} \quad Y \sim \mathrm{Poisson}(\lambda)$$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = \underline{0.002}$ or $\lambda = \underline{63}$

Hypothesis testing

1. Determine a **model**:

   $$X_1, \ldots, X_{31'000} \sim \mathrm{Bernoulli}(\pi) \quad \text{or} \quad Y \sim \mathrm{Poisson}(\lambda)$$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi \leq 0.002$ or $\lambda \leq 63$ $\pi \neq 0.002$, $\lambda \neq 63$

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \mathrm{Bernoulli}(\pi)$ or $Y \sim \mathrm{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$

3. Determine a **test statistic** (quantity that can differentiate between H_0 and H_A, and whose distribution under H_0 you can compute):

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$

3. Determine a **test statistic** (quantity that can differentiate between H_0 and H_A, and whose distribution under H_0 you can compute):

   $T :=$ Number of deaths under H_0:
   $T \sim \text{binomial}(31'000, 0.002)$ or $T \sim \text{Poisson}(63)$

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \mathrm{Bernoulli}(\pi)$ or $Y \sim \mathrm{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

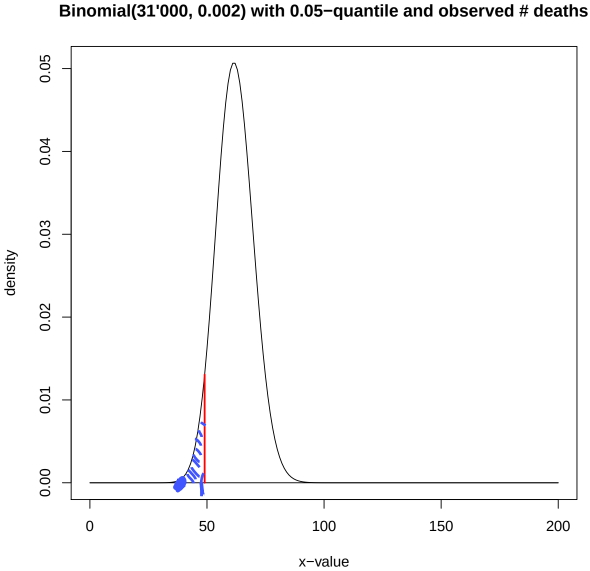
   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$

3. Determine a **test statistic** (quantity that can differentiate between H_0 and H_A, and whose distribution under H_0 you can compute):

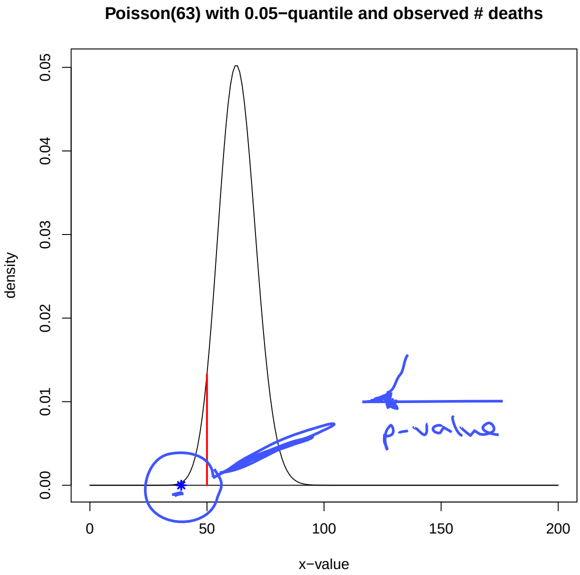
   $T :=$ Number of deaths under H_0:
   $T \sim \mathrm{binomial}(31'000, 0.002)$ or $T \sim \mathrm{Poisson}(63)$

4. Determine a **significance level** (α), i.e. the probability of rejecting H_0 when H_0 is true: e.g. $\alpha = 0.05$

Binomial distribution



Poisson distribution



P-value

- Probability under H_0 to obtain the observed value or a more extreme value of the test statistic

  $\Rightarrow$ p-value is always between 0 and 1!

  For mammography study: p-value is 0.0012 under binomial model and 0.0008 under Poisson model

- Smallest significance level for which H_0 just gets rejected

- Can be used for hypothesis testing: Reject H_0 if p-value $\leq \alpha$

- Quantifies significance of alternative

P-value

- Probability under H_0 to obtain the observed value or a more extreme value of the test statistic

  $\Rightarrow$ p-value is always between 0 and 1!

  For mammography study: p-value is 0.0012 under binomial model and 0.0008 under Poisson model

- Smallest significance level for which H_0 just gets rejected

- Can be used for hypothesis testing: Reject H_0 if p-value $\leq \alpha$

- Quantifies significance of alternative

Power

	retain H_0	reject H_0
H_0 true	-	type I error
H_A true	type II error	-

Power

	retain H_0	reject H_0
H_0 true	-	type I error
H_A true	type II error	-

- Signficiance level bounds probability of type I error: $\mathbb{P}(\text{type I error}) \leq \alpha$
- Power $:= 1 - \mathbb{P}(\text{type II error})$

Power

	retain H_0	reject H_0
H_0 true	-	type I error
H_A true	type II error	-

- Signficiance level bounds probability of type I error: $\mathbb{P}(\text{type I error}) \leq \alpha$
- Power $:= 1 - \mathbb{P}(\text{type II error})$
- Note that there is a trade-off between the probability of making a type I error and the probability of making a type II error (Why?)
- Note that power of 1-sided test is usually higher than for 2-sided test (Why?)

  $\Rightarrow$ Perform 1-sided test if you are only interested in detecting deviations in one direction

Power

	retain H_0	reject H_0
H_0 true	-	type I error
H_A true	type II error	-

- Signficiance level bounds probability of type I error: $\mathbb{P}(\text{type I error}) \leq \alpha$

- Power $:= 1 - \mathbb{P}(\text{type II error})$

- Note that there is a trade-off between the probability of making a type I error and the probability of making a type II error (Why?)

- Note that power of 1-sided test is usually higher than for 2-sided test (Why?)

  $\Rightarrow$ Perform 1-sided test if you are only interested in detecting deviations in one direction

Recap: Hypothesis testing

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

1. **Model**: $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$
2. **Null hypothesis** (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$
3. **Test statistic** T = Number of deaths under H_0
   $T \sim \text{binomial}(31'000, 0.002)$ or $T \sim \text{Poisson}(63)$
4. **Significance level**: $\alpha = 0.05$

Any important assumption that we should relax?

Alternative test: assume no knowledge of π_{control}

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

Null hypothesis: $\pi_{\text{control}} = \pi_{\text{treatment}}$ Alternative: $\pi_{\text{control}} > \pi_{\text{treatment}}$

Alternative test: assume no knowledge of π_{control}

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

Null hypothesis: $\pi_{\text{control}} = \pi_{\text{treatment}}$ Alternative: $\pi_{\text{control}} > \pi_{\text{treatment}}$

Knowing that 102 subjects died and that number of treatments / controls is 31'000, what is probability that deaths are so unevenly distributed?

Alternative test: assume no knowledge of π_{control}

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

Null hypothesis: $\pi_{\text{control}} = \pi_{\text{treatment}}$ Alternative: $\pi_{\text{control}} > \pi_{\text{treatment}}$

Knowing that 102 subjects died and that number of treatments / controls is 31'000, what is probability that deaths are so unevenly distributed?

- Test statistic T: number of deaths among the treated individuals
- Model: **Hypergeometric distribution**:

$$\mathbb{P}_{H_0}(T = 39) = \frac{\binom{31'000}{39}\binom{31'000}{63}}{\binom{62'000}{102}}$$

- p-value $= 0.011$ $= \sum_{i=0}^{39} \mathbb{P}_{H_0}(T = i)$

Fisher's exact test

Test on previous slide is known as **Fisher's exact test**

Fisher's exact test

Test on previous slide is known as **Fisher's exact test**

Advantages:

- Does not assume knowledge about the true probability of dying due to breast cancer in the control population

Fisher's exact test

Test on previous slide is known as **Fisher's exact test**

Advantages:

- Does not assume knowledge about the true probability of dying due to breast cancer in the control population

Shortcomings:

- Assumes knowledge of the margins (i.e., row and column sums)
- Alternative is Bernard's test (estimates the margins)

  for more details on this test, see e.g. `http://www.nbi.dk/~petersen/Teaching/Stat2009/Barnard_ExactTest_TwoBinomials.pdf`
- Both tests are difficult to perform on large tables for computational reasons

References

- For a statistics review, including controlled experiments and observational studies (chapters 1 and 2) and hypothesis testing (chapter 26-29):

  D. Freedman, R. Pisani, R. Purves. *Statistics*. 2007.

- For how to perform hypothesis testing in `R` (chapter 4):

  P. Dalgaard. *Introductory Statistics with R*. 2002.

- For observational studies and experiments, including the HIP study (chapter 1):

  D. Freedman. *Statistical Models: Theory and Practice*. 2009.