Chapter 7: One Sample Hypothesis Testing

- Two hypotheses

H_0 Null: Baseline, to be disproved

H_1 Alternative: Opposite of null (contradicts null)

- E.g.) Want to test if low socioeconomic status (SES) mothers have babies with lower birthweight
  Know national average is 120 oz
  For $n = 100$ Low SES mothers, observe
  $\bar{x} = 115$ oz, $s = 24$ oz

  Let μ = average birthweight of low SES mothers

  H_0: $\mu = 120$
  H_A: $\mu < 120$

- We use data (e.g. $\bar{x}$ and s^2) to make a decision (H_0 vs. H_1)

- Truth

Decision	H_0	H_1
Accept H_0	True negative	Type II error
Reject H_0	Type I error	true positive

- E.g. If $\mu = 120$, but we say $\mu < 120$, this is a Type I error.

- Typically, Null: parameter = some value

  Alternative: parameter $\neq$ value
  $>$ value
  $<$ value
  ↑
  one of these

- Probability of a Type I error

  $= Pr(\text{Reject } H_0 \mid H_0)$

  = <u>significance level</u>

  $= \alpha$

- Probability of a Type II error

  $= Pr(\text{Fail to reject } H_0 \mid H_1)$

  $= \beta$

- <u>Power</u> $= Pr(\text{reject } H_0 \mid H_1)$

  $= 1 - Pr(\text{fail to reject } H_0 \mid H_1)$

  $= 1 - \beta$

- These probabilities are all in terms of <u>repeated samples</u>.

- <u>Exercise</u>: New pain relief drug for osteoarthritis (OA) 50 OA patients take drug. Measure X_i = % decline in pain level reported. $X_i > 0 \Rightarrow$ less pain, $X_i < 0 \Rightarrow$ more pain. Drug is effective if average $X_i > 0$.
  i) What hypotheses are being tested?
  ii) What do type I, type II error, and power mean?

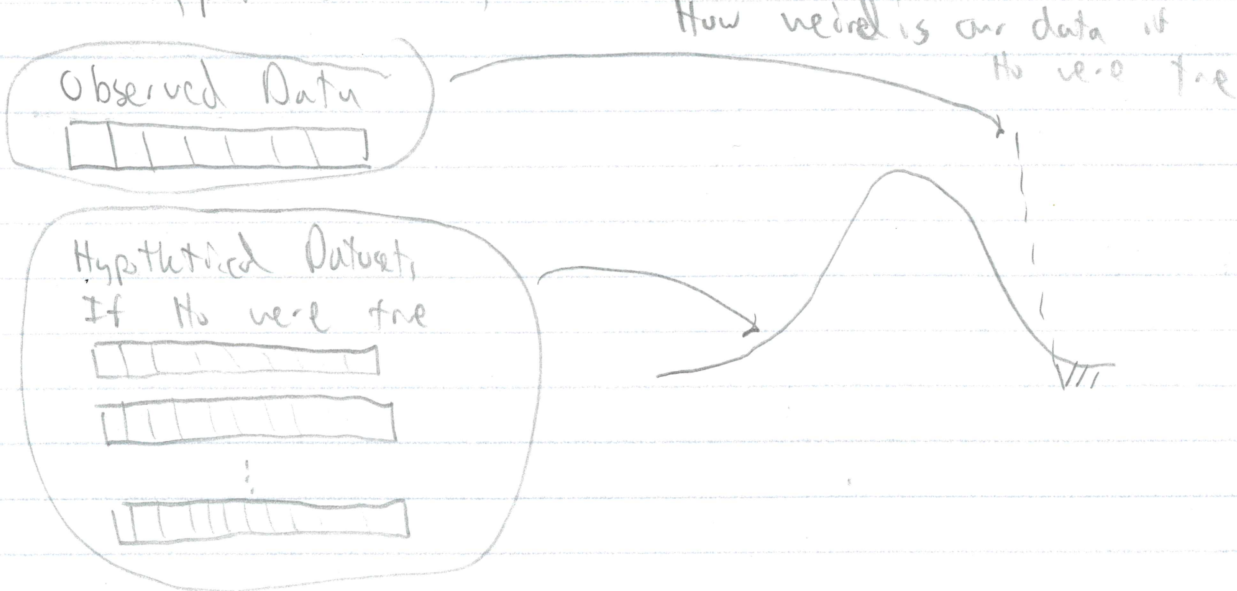
<u>Solution</u>: μ = mean % decline in pain
H_0: $\mu = 0$
H_A: $\mu > 0$
Type I: Say reduces pain on average but does not
Type II: Say does not reduce pain when it does
Power: Pr(say works | it works)

- Goal is to make α and β as small as possible

But, typically as $\alpha \searrow$, $\beta \nearrow$ and as $\beta \searrow$, $\alpha \nearrow$

Strategy: fix $\alpha = 0.05$ (or similar) and use a test that has small β.

- All Hypothesis testing



- If data are very weird, reject H_0

  If data not very weird, fail to reject H_0.

- Back to our example:

  $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

  $H_0: \mu = \mu_0$
  $H_1: \mu < \mu_0$

  This is a <u>one-sided</u> test because the alternative is of the form $>$ or $<$ (as opposed to $\mu \neq \mu_0$)

- Idea: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

  use this as a measure of weird

- If $\mu = \mu_0$, then t should be close to 0, since $\bar{x}$ will be close to μ_0

- In fact: $t \sim t_{n-1}$

  but only if H_0 is true

- So compare t to a t_{n-1} distribution

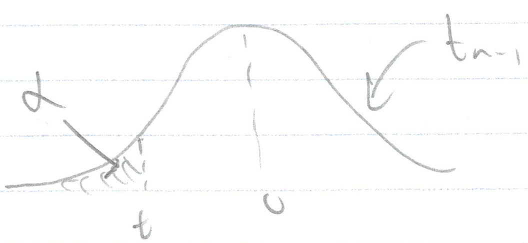
- Reject H_0 if $t < a$

  some value

  since provides evidence for H_1

- Want to fix Type I error at α

$Pr(t < a) = \alpha$



Choose $a = t_{n-1,\alpha} = qt(\alpha, n-1)$

- For $H_0: \mu = \mu_0$
  $H_1: \mu < \mu_0$
  given α

Let $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Reject H_0 if $t < qt(\alpha, n-1)$

Fail to reject if $t \geq qt(\alpha, n-1)$

- A test statistic is a statistic which measures strength of evidence against H_0

  e.g. $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

- A critical value is the value beyond which we reject H_0

  E.g. $t_{n-1, \alpha} = qt(\alpha, n-1)$

- Rejection Region: Values of data for which we reject H_0

- Acceptance Region: values of data for which we accept H_0

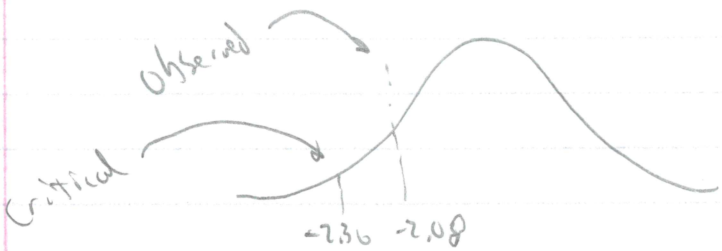
Ex.) Lets choose $\alpha = 0.1$ and run the test for the birthweight example.

$H_0: \mu = 120$ $\quad\bar{x} = 115\quad n = 100$
$H_1: \mu < 120$ $\quad s = 24$

$$t = \frac{115 - 120}{24/\sqrt{100}} = -2.08$$

$$t_{100-1,\,0.1} = qt(0.1, 99) = -2.36$$

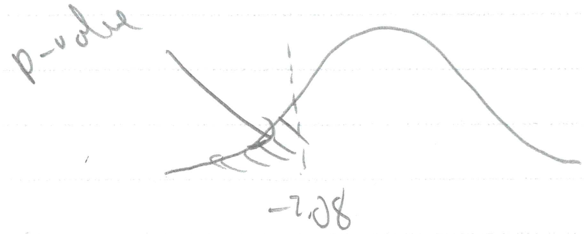
Since $-2.08 > -2.36$, not weird, so fail to to reject H_0.



- P-value: smallest α at which we would reject H_0

  = Probability we would see data as extreme or more extreme than what we saw

Ex.)



$$= pt(-2.08, n-1)$$

- Note: p-value < α $\iff$ $t < t_{n-1,\alpha}$

- So typically, we only report p-value and reject if $p < \alpha$

- Exercise.) Mean "infarct size" is 25 (ck-g-EQ/m²) (size of dead tissue in heart)
  8 patients treated with drug have average infarct size of 16 w/ SD. 10. Is drug effective in reducing infarct size?

$H_0: \mu = 25$ $\qquad H_1: \mu < 25$

$$t = \frac{16 - 25}{10/\sqrt{8}} = -2.55$$

$pt(-2.55, 8-1) = 0.01905$

$\Rightarrow$ have significant results.

- $0.1 < p < 0.5 \Rightarrow$ significant
  $0.001 < p < 0.01 \Rightarrow$ highly significant
  $p < 0.001 \Rightarrow$ very highly significant.

- **Note:** p-value tells you strength of evidence that there is an effect

↑ It it tells you nothing about the size of an effect.

- Large sample size, tiny effect ⇒ still small p-value

- Ex.) $n = 10{,}000 \quad \bar{x} = 119, \quad s = 24$

$$t = \frac{119 - 120}{54/\sqrt{10000}} = -4.17$$

p-value = pt(-4.17, 9999) = 0.00001536

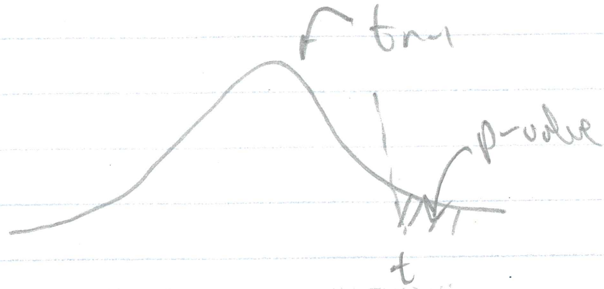
↑ very sure $\mu < 120$, but 119 is such a tiny difference from 119

↑ statistically significant, not practically significant

- If $x_1, x_2, \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$H_0: \mu = \mu_0, \quad H_1: \mu > \mu_0$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$



p-value $= 1 - pt(t, n-1)$

Reject if t is large (supports alt)

- Exercise: Mean cholesterol in population is 175 mg/dL
10 children of fathers who died from heart disease
$\bar{x} = 200, \quad s = 50$
Is there evidence that these children have higher cholesterol?

$H_0: \mu = 175, \quad H_1: \mu > 175$

$$t = \frac{200 - 175}{50/\sqrt{10}} = 1.58$$

$1 - pt(1.58, 9) = 0.07428$

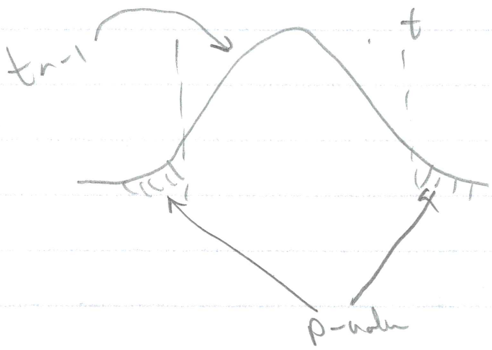
[weak evidence that they have higher cholesterol

- If $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

p-value = $2\,pt(-abs(t), n-1)$



- This is a <u>two-tailed test</u>.

- Critical value technique:

Reject if $|t| > qt(1-\frac{\alpha}{2}, n-1)$

- <u>Exercise.</u>) Cholesterol levels of US women are 190 mg/dL on average. Want to compare cholesterol levels of recent Asian immigrants. 100 female Asian immigrants, $\bar{x} = 181.52$ mg/dL, $s = 40$ mg/dL

$H_0: \mu = 190$ $\quad$ $H_1: \mu \neq 190$

$$t = \frac{181.52 - 190}{40/\sqrt{100}} = -2.12$$

p-value = $2\,pt(-2.12, 99) = 0.037$

- Use 2-sided tests by default unless only one direction is of interest

↑ Eg. drug efficacy.

$\boxed{\text{One-sample t-tests in R}}$

- Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$

Then reject H_0 at level α if and only if $100(1-\alpha)\%$ CI <u>does not</u> contain μ_0

- I.e., if μ_0 is outside CI, reject H_0 at level α, and p-value is less than α

- General result: All tests correspond to some confidence interval and vice versa.

- So a $100(1-\alpha)\%$ CI contains all values of μ_0 that would fail to reject $H_0: \mu = \mu_0$

- One sided CI's correspond to one sided tests.

- Power Calculations:

Power = $Pr(\text{Reject } H_0 \mid H_1) = 1 - \beta$

- In experiments/surveys, common to guess power prior to collecting data.
  - Calculate n needed
  - See what effect sizes we can detect
  - How likely will our study be successful?
- Need:
  - i) Guess of effect size $\mu_0 - \mu_1$] from pilot study or
  - ii) Guess of standard deviation σ] wild guess
  - iii.) Sample size n] provided by
  - iv.) Significance level α] researcher

- Since σ is known, use z-test instead of t-test to calculate power

- $H_0: \mu = \mu_0$ Reject if $\underbrace{\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}_{N(0,1) \text{ if } H_0 \text{ true}} < z_\alpha$

  $H_A: \mu < \mu_0$

$Pr(\text{Reject } H_0 \mid \mu = \mu_1)$

$= Pr\left(\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha \mid \mu = \mu_1\right)$

$= Pr\left(\dfrac{\bar{X} - \mu_1 + \mu_1 - \mu_0}{\sigma/\sqrt{n}} < z_\alpha \mid \mu = \mu_1\right)$

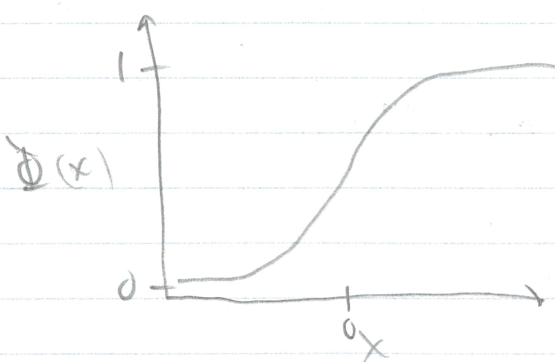
$$= \Pr\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} - \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} < z_\alpha \,\middle|\, \mu = \mu_1 \right)$$

$$= \Pr\Bigg(\underbrace{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}_{\substack{\sim N(0,1) \\ \text{if } \mu = \mu_1}} < \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_\alpha \,\Bigg|\, \mu = \mu_1 \Bigg)$$

$$= \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_\alpha \right)$$ ← CDF of $N(0,1)$

$$= \boxed{\Phi\left(z_\alpha + \frac{\sqrt{n}\,(\mu_0 - \mu_1)}{\sigma} \right)}$$ (re-writing)

estimate of power

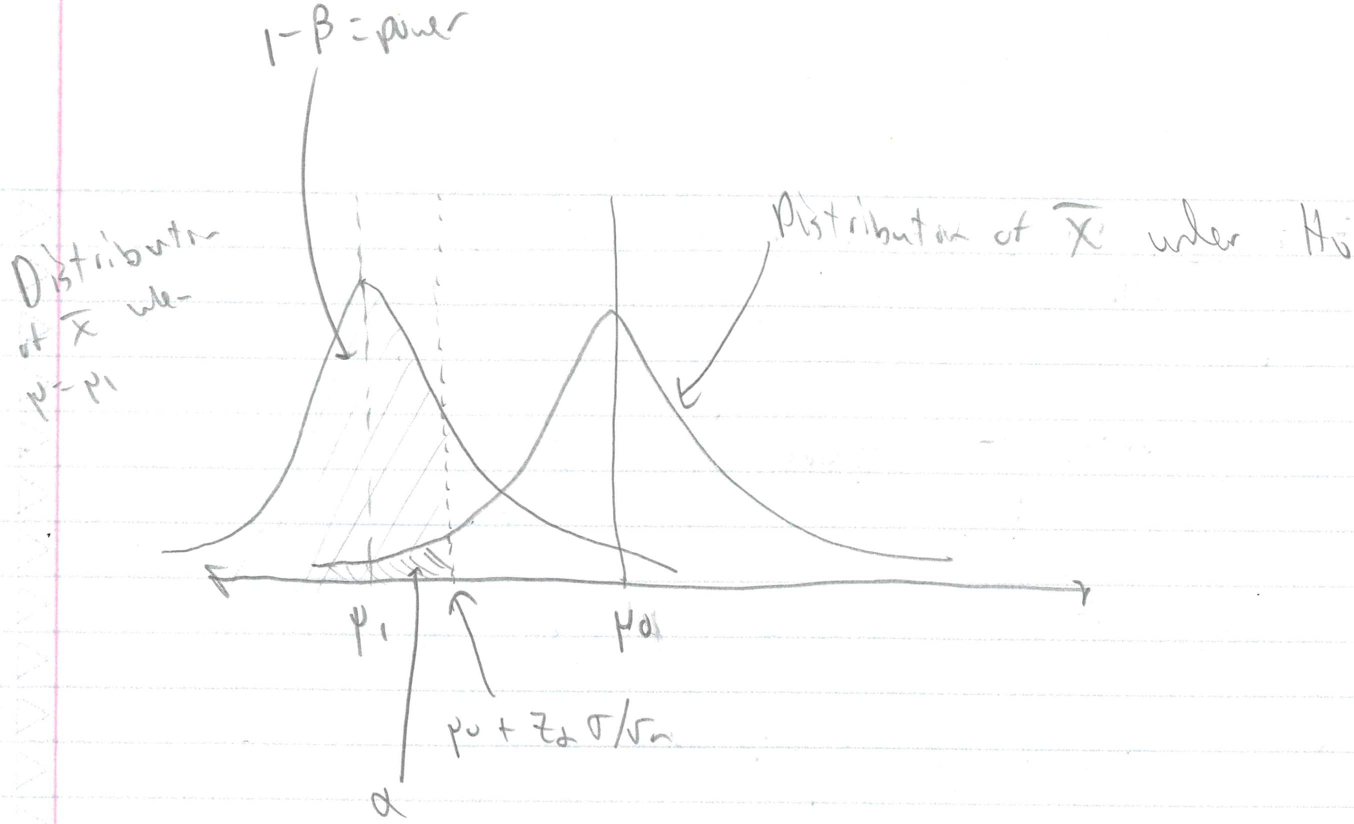


$\alpha \searrow \Rightarrow z_\alpha \searrow \Rightarrow$ Power $\searrow$ (more stringent test)

$\sqrt{n} \searrow \Rightarrow$ Power $\searrow$ (less data)

$(\mu_0 - \mu_1) \searrow \Rightarrow$ Power $\searrow$ (smaller effect size)

$\sigma \searrow \Rightarrow$ Power $\nearrow$ (more precise data)



- Example.) Pilot study with 10 individuals get $\bar{X} = -5$, $s = 10$. $\mu = 0$ is no effect. Proposed new study of $n = 30$. At $\alpha = 0.05$, what is power?

$\mu_1 = -5$, $\sigma = 10$, $n = 30$, $\alpha = 0.05$ $\Rightarrow z_{0.05} = -1.645$
qnorm(0.05)

$$\Phi\left(-1.645 + \frac{0-(-5)}{10}\sqrt{30}\right) = \text{pnorm}(1.094)$$

$$= 0.863$$

Exercise.) Mean birthweight in US is 120 oz. What is the power to detect low birthweight in a study with $n=100$, $\mu_1 = 115$, $\alpha = 0.05$, $\sigma = 24$

$$\Phi\left(-1.645 + \frac{120-115}{24}\sqrt{100}\right) = \text{pnorm}(0.438) = 0.669$$

- If $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$

$$\text{Power} = \Phi\left(z_\alpha + \sqrt{n}\frac{(\mu_1 - \mu_0)}{\sigma}\right)$$

just swap μ_1 and μ_0

- If $H_0: \mu = \mu_0$ vs. $\mu \neq \mu_0$

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

add both but use $\alpha/2$ quantile.

- Sample size calculation

- Typically, you want a power of at least 0.8

- So, what n will give us power of 0.8?

- Assume
  - i.) Guess of effect size $\mu_1 - \mu_0$] Pilot study or
  - ii.) Guess of standard deviation σ] wild guess
  - iii.) Power $1-\beta$] chosen by
  - iv.) significance level] researcher

- If $H_0: \mu = \mu_0$ vs. $H_1: \mu < \mu_0$

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

just solve for n
(either numerically or by hand)

quantile function

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

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

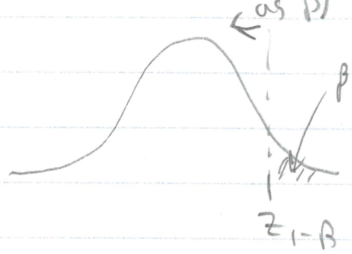
$$\Rightarrow n = \frac{(z_{1-\beta} - z_\alpha)^2\,\sigma^2}{(\mu_0-\mu_1)^2} = \boxed{\frac{(z_{1-\beta} + z_{1-\alpha})^2\,\sigma^2}{(\mu_0-\mu_1)^2}}$$

same n for $H_0: \mu = \mu_0$ vs. $H_1: \mu > \mu_0$

Exercise.) What is the effect on n of

(i) $\sigma \uparrow$, (ii) $\beta \uparrow$, (iii) $\alpha \uparrow$, (iv) $|\mu_0 - \mu_1| \uparrow$

i.) $\sigma \uparrow \Rightarrow n \uparrow$ (less precise so need more data)
ii.) $\beta \uparrow \Rightarrow 1-\beta \downarrow \Rightarrow Z_{1-\beta} \downarrow \Rightarrow n \downarrow$ (less power needed)
iii.) $\alpha \uparrow \Rightarrow 1-\alpha \downarrow \Rightarrow Z_{1-\alpha} \downarrow \Rightarrow n \downarrow$ (less stringent test)
iv.) $|\mu_0 - \mu_1| \uparrow \Rightarrow n \downarrow$ (larger effect size easier



- For $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$, also solve for n in power calculation)

$$n \approx \frac{\sigma^2 (Z_{1-\beta} + Z_{1-\alpha/2})^2}{(\mu_0 - \mu_1)^2}$$ only difference is $\alpha/2$

$\boxed{\text{Power Calculations in R}}$

- Skip §7.8

One-Sample Inference for Binomial

- **Example:** Prevalence of breast cancer is 2%. Of 10,000 women whose mothers had breast cancer, 400 of them got breast cancer in their lives. Are they at higher risk?

- Let X = # of those 10,000 who got breast cancer

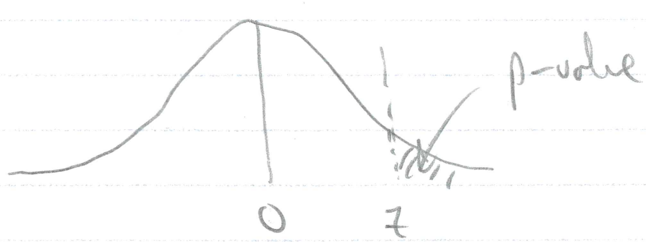
$$X \sim \text{Binom}(10000, p)$$

$$H_0: p = 0.02$$
$$H_1: p > 0.02$$

- One way to run this test, is to use the normal approximation

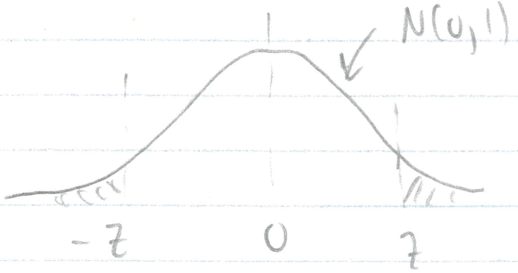
- $$\hat{p} = \frac{X}{n} \sim N\left(p_0, \frac{p_0(1-p_0)}{n}\right) \quad \text{if } H_0 \text{ were true}$$

$$\Rightarrow \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$$



$$\text{p-value} = 1 - \Phi\left(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right)$$

- If $H_0: p = p_0$ vs. $H_1: p \neq p_0$, then calculate area in both tails.



$$\text{p-value} = 2\,\Phi\left(-\frac{|\hat{p} - p_0|}{\sqrt{p_0(1-p_0)/n}}\right)$$

- Continuity correction: Subtract $\frac{1}{2n}$ in numerator if calculating upper tail. Add $\frac{1}{2n}$ in numerator if calculating lower tail.

- Ex.) Back to breast cancer example

$$\hat{p} = \frac{400.5}{10000} = 0.04005$$

$$z = \frac{0.04005 - 0.02}{\sqrt{0.02(1-0.02)/10000}} = 14.32$$

$$\text{p-value} = \Phi(-14.32) = 8.4 \cdot 10^{-47} < 0.0001$$

↑ very highly significant

Exact test:

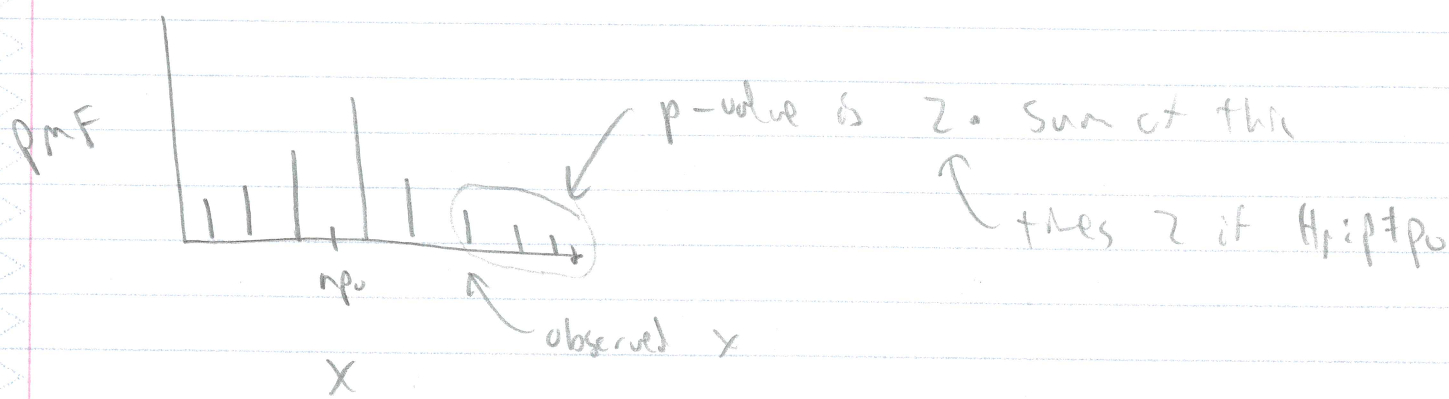
- Still $X \sim Bin(n, p)$, $H_0: p = p_0$, $H_1: p \neq p_0$ ($>$, $<$)

- Normal method only works for large n

Rule: $n p_0 (1-p_0) \geq 5$

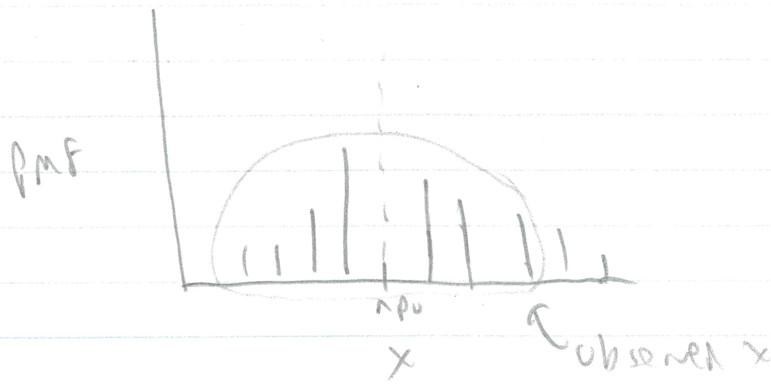
- If n is not large enough, use an exact test

↑ Controls for α for all n (not just large n)

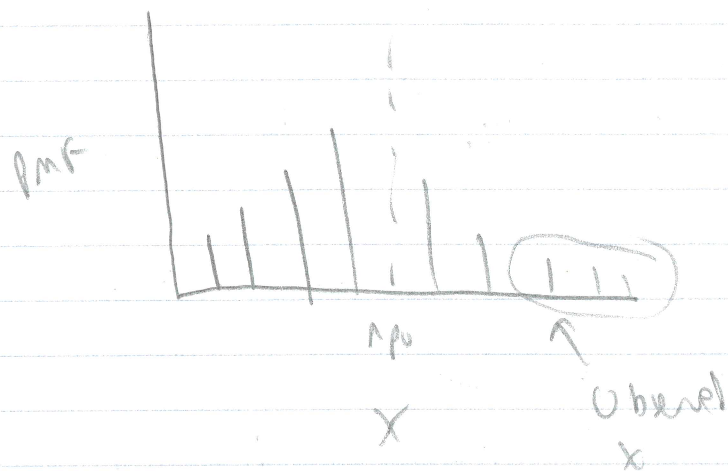


$$\text{p-value} = \begin{cases} 2 \sum_{k=0}^{x} \binom{n}{k} p_0^k (1-p_0)^{n-k} & \text{if } x \leq n p_0 \\ 2 \sum_{k=x}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} & \text{if } x > n p_0 \end{cases}$$

$H_1: p < p_0$



$H_1: p > p_0$



↳ p-value = Pr(As or more extreme X | H_0)

↳ can compute this probability exactly for binomial tests.

- Ex.) In a nuclear facility, 13 deaths, of which 5 were caused by cancer. Cancer causes 20% deaths in same age group. Are there actually more cancer deaths than would be expected?

- $H_0: p = 0.2$ $\quad X \sim \text{Binom}(13, p)$
  $H_1: p > 0.2$

$$\text{p-value} = \sum_{k=5}^{13} \binom{n}{k} 0.2^{k} \, 0.8^{13-k}$$

$$= 1 - \text{pbinom}(4, \text{size} = 13, \text{prob} = 0.2)$$

$$= 0.09913$$

$\hookrightarrow$ No evidence more cancer deaths than typical

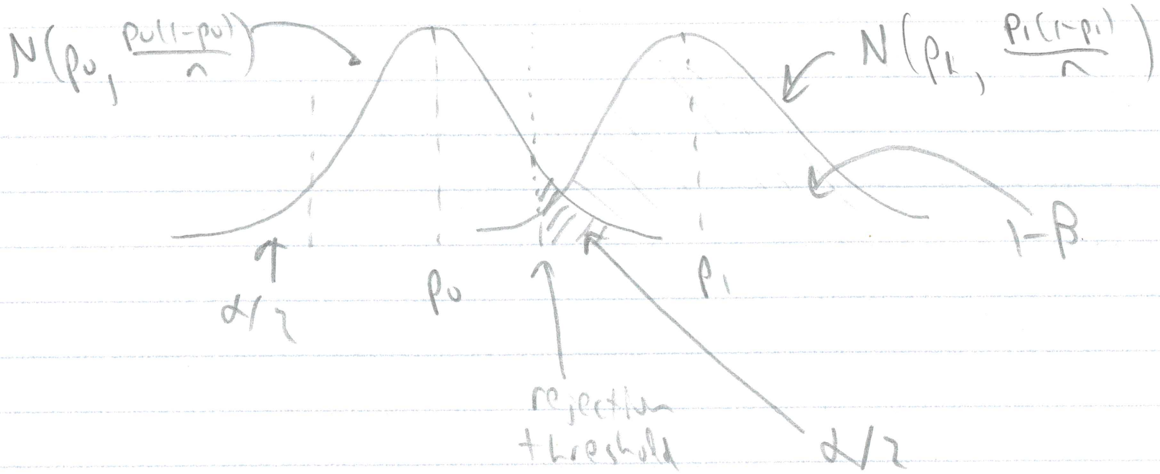
$\boxed{\text{Binomial tests in R}}$

Power Calculations for binomial tests

$Pr(\text{Reject } H_0 \mid H_1)$

Need:
- true alternative value, p_1
- Sample size n
- Significance level α

Skip derivation, but using normal approximation



$H_a: p \neq p_0$, $\quad \text{Power} = \Phi\left[\frac{p_0(1-p_0)}{p_1(1-p_1)}\left(z_{\alpha/2} + \frac{|p_0-p_1|\sqrt{n}}{\sqrt{p_0(1-p_0)}}\right)\right]$

$n \uparrow \Rightarrow$ power $\uparrow$ (more data)
$\alpha \uparrow \Rightarrow$ power $\uparrow$ (less stringent)
$|p_0-p_1| \uparrow \Rightarrow$ power $\uparrow$ (more signal)

Sample Size Calculation

Need

1.) true alternative p_1
2.) Power desired $1-\beta$
3.) significance level α

Set $1-\beta = f(n, p_1, \alpha)$

solve for n

$\boxed{\text{Power Calculations in R}}$

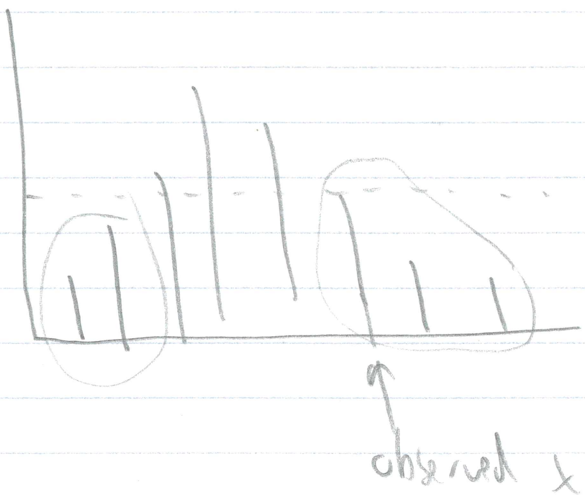
Poisson Test

$X \sim Poi(\mu)$

$H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0, \quad \mu > \mu_0, \quad \mu < \mu_0$

$$\text{p-value} = \sum_{k \text{ s.t. } P(k) < Pr(X)} Pr(k) = \sum_{k \text{ s.t. } Pr(k) < Pr(X)} \frac{1}{k!} e^{-\mu_0} \mu_0^k$$



- Ex.) 8418 rubber workers, ages 40-84
  4 deaths due to Hodgkins disease
  About 3.3 deaths expected based on US mortality rates

$X = \#$ with Hodgkins $\sim Poi(\mu)$

$H_0: \mu = 3.3$ vs. $H_1: \mu \neq 3.3$

poisson.test(x=4, r=3.3)

$\hookrightarrow$ p-value = 0.578

Note: Binomial not appropriate here because each individual has a different p_i. The μ_0 was based on these expected p_i's for each individual.

- Standardized Mortality Ratio: (SMR)

$$100\% \cdot \frac{\text{observed}}{\text{expected}}$$

- SMR > 100% ⇒ more deaths in study population
  SMR < 100% ⇒ fewer deaths
  SMR = 100% ⇒ same # deaths

- Ex.) $SMR = \frac{4}{3.3} \cdot 100\% \approx 121\%$

Exercises 7.1 – 7.8

12 patients, 24 hours, mean serum-creatine is 1.2 mg/dL

7.1) General pop $\mu = 1$, $\sigma = 0.4$

$H_0: \mu = 1$ $\quad$ $H_1: \mu \neq 1$

$$z = \frac{1.2 - 1}{0.4/\sqrt{12}} = 1.732$$

under H_0, $z \sim N(0,1)$

$2 \cdot \text{pnorm}(-1.73) = 0.8326$

fail to reject H_0

If instead $s = 0.4$, $t = 1.73 \sim t_{11}$

$2 \cdot \text{pt}(-1.73, df = 11) = 0.1112$

7.2) p-value = 0.8326

7.3 $s = 0.6$, σ unknown

$$t = \frac{1.2 - 1}{0.6/\sqrt{12}} = 1.155$$

$2 \cdot \text{pt}(-1.155, df = 11) = 0.2727$

Fail to reject H_0

7.4) $1.2 \pm \underbrace{t_{11, 0.975}}_{\substack{qt(0.975,\, df=11) \\ 2.201}} \cdot 0.6/\sqrt{12}$

$= (0.82, 1.58)$

7.5) 1 included in 95% CI $\Rightarrow$ p-value > 0.05

7.6) $2 \cdot pt(-1.52, 6) = 0.1793$

7.7) $1 - pt(2.5, df = 36) = 0.008557$

7.8) $qt(0.1, df = 54) = -1.297$

• 7.52 – 7.55 $n = 200$, X = # who develop cataracts
observe $x = 4$

7.52) $X \sim Binom(200, p)$ $H_0: p = 0.01$ vs. $H_1: p \neq 0.01$

Use exact Binomial test since $0.01 \cdot 0.99 \cdot 200 = 1.98 < 5$

7.53) Rosner: $2 \cdot \sum_{k=4}^{200} \binom{n}{k} 0.01^k \, 0.99^{n-k}$

$= 2\left(1 - \underbrace{\sum_{k=0}^{3} \binom{n}{k} 0.01^k \, 0.99^{n-k}}\right)$

$= 0.2839$

= dbinom(0:3, size = 200, prob = 0.01)

R-way: 0.142

binom.test(x=4, n=200, p=0.01)

- Extend to 5-year period. $X = 20$, $n = 200$
5-year incidence rate is 0.05

(7.54) $H_0: p = 0.05$, $H_1: p \neq 0.05$

$0.05 \cdot 0.95 \cdot 200 = 9.5$, so can use normal way

$$z = \frac{0.1 - 0.05 - \frac{1}{400}}{\sqrt{0.05 \cdot 0.95 / 200}} = 3.082$$

$$2 \cdot (1 - \text{pnorm}(3.082)) = 0.002055$$

Reject H_0

(7.55) $\hat{p} \pm z_{0.975} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ qnorm(0.975) = 1.96

$$= 0.1 \pm 1.96 \sqrt{\frac{0.1 \cdot 0.9}{200}}$$

$$= (0.05842, 0.1416)$$

prop.test(x = 20, n = 200, p = 0.05)