6. Statistical Inference: Significance Tests

Goal: Use statistical methods to check hypotheses such as

"Mental health tends to be better at higher levels of socioeconomic status (SES)" (an effect)

"For treating anorexia, cognitive behavioral and family therapies have same effect" (*no* effect)

Hypotheses: Predictions about a population expressed in terms of parameters for certain variables

A **significance test** uses data to summarize evidence about a hypothesis by comparing sample estimates of parameters to values predicted by the hypothesis.

We answer a question such as, “If the hypothesis were true, would it be unlikely to get estimates such as we obtained?”

Five Parts of a Significance Test

- **Assumptions** about type of data (quantitative, categorical), sampling method (random), population distribution (binary, normal), sample size (large?)

- **Hypotheses**:

Null hypothesis (H_0): A statement that parameter(s) take specific value(s) (Usually: “no effect”)

Alternative hypothesis (H_a): states that parameter value(s) falls in some alternative range of values

(an “effect”)

p. 1 examples?

- **Test Statistic**: Compares data to what null hypo. H_0 predicts, often by finding the *number of standard errors* between sample estimate and H_0 value of parameter
- ***P*-value (*P*):** A probability measure of evidence about H_0, giving the probability (under presumption that H_0 true) that the test statistic equals observed value or value even more extreme in direction predicted by H_a.
  - The smaller the *P*-value, the stronger the evidence against H_0.
- **Conclusion**:
  - If no decision needed, report and interpret P-value
  - If decision needed, select a cutoff point (such as 0.05 or 0.01) and reject H_0 if P-value = that value

– The most widely accepted minimum level is 0.05, and the test is said to be *significant at the .05 level* if the P-value = 0.05.

– If the *P*-value is not sufficiently small, we *fail to reject* H_0 (then, H_0 is not necessarily true, but it is plausible)

– Process is analogous to American judicial system
  - H_0: Defendant is *innocent*
  - H_a: Defendant is *guilty*

Significance Test for Mean

- *Assumptions*: Randomization, quantitative variable, normal population distribution
- *Null Hypothesis*: H_0: $\mu = \mu_0$ where μ_0 is particular value for population mean

  (typically "no effect" or "no change" from a standard)
- *Alternative Hypothesis*: H_a: $\mu \neq \mu_0$

  2-sided alternative includes both $>$ and $<$ null value
- *Test Statistic*: The number of standard errors that the sample mean falls from the H_0 value

$$t = \frac{\bar{y} - m_0}{se} \quad \text{where} \quad se = s/\sqrt{n}$$

When H_0 is true, the sampling dist of the *t* test statistic is the *t* distribution with $df = n - 1$.

- *P-value*: Under presumption that H_0 true, probability the test statistic equals observed value or even more extreme (i.e., larger in absolute value), providing stronger evidence against H_0
  - This is a *two-tail* probability, for the two-sided H_a

- *Conclusion*: Report and interpret *P*-value. If needed, make decision about H_0

Example: Anorexia study (revisited)

- Weight measured before and after period of treatment

- y = weight at end – weight at beginning

- In previous chapter, we found CI for population mean of y based on n=17 girls receiving “family therapy,” with data

y = 11.4, 11.0, 5.5, 9.4, 13.6, -2.9, -0.1, 7.4, 21.5, -5.3,
-3.8, 13.4, 13.1, 9.0, 3.9, 5.7, 10.7

Is there evidence that family therapy has an effect?

- Let μ = population mean weight change
- Test H_0: μ = 0 (no effect) against H_a: μ ≠ 0.
- Data have

Variable	N	Mean	Std.Dev.	Std. Error Mean
weight_change	17	7.265	7.157	1.736

Recall that *se* obtained as

$$se = s/\sqrt{n} = 7.157/\sqrt{17} = 1.736$$

- **Test Statistic** *(df = 16):*

$$t = \frac{\bar{y} - m_0}{se} = \frac{7.265 - 0}{1.736} = 4.2$$

- ***P*-Value**: $P = 2\text{P}(t > 4.2) = 0.0007$

Note *t* table (Table B, p. 593) tells us $P(t > 3.686) = 0.001$, so test statistic $t = 3.686$ (or -3.686) would have P-value = 0.002

Interpretation: If H_0 were true, prob. would = 0.0007 of getting sample mean at least 4.2 standard errors from null value of 0.

•**Conclusion**: Very strong evidence that the population mean differs from 0. (Specifically, it seems that $\mu > 0$, as also suggested by 95% CI of (3.6, 10.9) we found in Chap. 5 notes)

SPSS output

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
weight_change	17	7.265	7.1574	1.7359

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
weight_change	4.185	16	.001	7.2647	3.58	10.945

Equivalence between result of significance test and result of CI

- When *P*-value = 0.05 in two-sided test, 95% CI for μ does not contain H_0 value of μ (such as 0)
- When *P*-value > 0.05 in two-sided test, 95% CI necessarily contains H_0 value of μ

(This is true for “two-sided” tests)

- CI has more information about actual value of μ

Example: Suppose sample mean = 7.265, s = 7.16, based on n = 4 (instead of n = 17)

Then,

$$se = s/\sqrt{n} = 7.16/\sqrt{4} = 3.58$$

$$\text{and } t = (7.265 - 0)/3.58 = 2.0$$

With *df* = *3*, this has two-sided *P*-value = 0.14.

Not very strong evidence against null hypo.

It is plausible that μ = 0.

Margin of error = 3.182(3.58) = 11.4, and 95% CI is (-4.1, 18.7), which contains 0 (in agreement with test result)

One-sided test about mean

Example: If study predicts family therapy has *positive* effect, could use $H_{a:}\ \mu > 0$

Data support this hypothesis if t far out in right tail, so P-value = right-tail prob.

- P-Value: $P = P(t > 2.0) = 0.07$ (for $n = 4$ case)

For $H_{a:}\ \mu < 0$, P-value = left-tail probability

- P-value: $P = P(t < 2.0) = 0.93$

In practice, two-sided tests are more common

Making a decision:

The α-level is a fixed number, also called the *significance level,* such that

if *P*-value = α, we "reject H_0"
If *P*-value > α, we "do not reject H_0"

Note: We say "Do not reject H_0" rather than "Accept H_0" because H_0 value is only one of many plausible values.

Example (*n* = 4, two-sided)*:* Suppose α = 0.05. Since *P*-value = 0.14, we do not reject H_0. But 0 is only one of a range of plausible values exhibited in 95% CI of (-4.1, 18.7).

Effect of sample size on tests

- With large *n* (say, *n* > 30), assumption of normal population distribution not important because of Central Limit Theorem.
- For small *n,* the *two-sided t* test is robust against violations of that assumption. One-sided test is *not* robust.
- For a given observed sample mean and standard deviation, the larger the sample size *n*, the larger the test statistic (because *se* in denominator is smaller) and the smaller the *P*-value. (i.e., we have more evidence with more data)
- We're more likely to reject a false H_0 when we have a larger sample size (the test then has more "power")
- With large *n*, "statistical significance" not the same as "practical significance."

Example: Suppose anorexia study for weight change had

$$\bar{y} = 1.0, s = 2.0, \text{ for } n = 400$$

$$\text{Then } se = 2.0/\sqrt{400} = 0.1,$$

Test stat

$$t = (1.0 - 0)/0.1 = 10.0,$$

$$P - \text{value} = 0.000000.....$$

95% CI is 1.0 ± 1.96(0.1), or (0.8, 1.2).

This shows there is a positive effect, but it is *very small in practical terms.*

Significance Test for a Proportion π

- Assumptions:
  - Categorical variable
  - Randomization
  - Large sample (but two-sided ok for nearly all *n*)

- Hypotheses:
  - Null hypothesis: H_0: $\pi = p_0$
  - Alternative hypothesis: H_a: $\pi \neq p_0$ (2-sided)
  - H_a: $\pi > p_0$ H_a: $\pi < p_0$ (1-sided)
  - Set up hypotheses before getting the data

- Test statistic:

$$z = \frac{\hat{p} - p_0}{s_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Note $s_{\hat{p}} = se_0 = \sqrt{p_0(1-p_0)/n}$, not $se = \sqrt{\hat{p}(1-\hat{p})/n}$ as in a CI

As in test for mean, test statistic has form

(estimate of parameter – H_0 value)/(standard error)

= no. of standard errors the estimate falls from H_0 value

- *P*-value:

  H_a: $\pi \neq p_0$ *P* = 2-tail prob. from standard normal dist.
  H_a: $\pi > p_0$ *P* = right-tail prob. from standard normal dist.
  H_a: $\pi < p_0$ *P* = left-tail prob. from standard normal dist.

- Conclusion: As in test for mean (e.g., reject H_0 if *P*-value = α)

Example: Can dogs smell cancer? (*British Medical Journal*, Sept. 25, 2004)

- Each trial, one bladder cancer urine sample placed among six control urine samples
- Do dogs make the correct selection better than with random guessing?
- In 54 trials, dogs made correct selection 22 times.

Let π = probability of correct guess, for particular trial

H_0: $\pi = 1/7$ (= 0.143, no effect), H_a: $\pi > 1/7$

Sample proportion = 22/54 = 0.407

Standard error

$$se_0 = \sqrt{p_0(1-p_0)/n} = \sqrt{(1/7)(6/7)/54} = 0.0476$$

Test statistic

z = (sample – null)/se_0 = [0.407 – (1/7)]/0.0476 = 5.6

P-value = right-tail probability from standard normal

= 0.00000001

There is extremely strong evidence that dogs' selections are better than random guessing (for the conceptual population this sample represents)

For standard α cut-off such as 0.05, we reject H_0 and conclude that $\pi > 1/7$.

Caveat: As in most medical studies, subjects were a *convenience sample.* We can not realistically randomly sample bladder cancer patients or dogs for the experiment.

Even though samples not random, important to employ randomization in experiment, in placement of bladder cancer patient's urine specimen among the 6 control specimens.

Decisions in Tests

- a-level (significance level): Pre-specified "hurdle" for which one rejects H_0 if the P-value falls **below** it (Typically 0.05 or 0.01)

P-Value	H_0 Conclusion	H_a Conclusion
$\leq .05$	Reject	Accept
$> .05$	Do not Reject	Do not Accept

- Rejection Region: Values of the test statistic for which we reject the null hypothesis
  - For 2-sided tests with $a = 0.05$, we reject H_0 if $|z| \geq 1.96$

Error Types

- Type I Error: Reject H_0 when it is true
- Type II Error: Do not reject H_0 when it is false

Test Result –	Reject H_0	Don't Reject H_0
True State H_0 True	Type I Error	Correct
H_0 False	Correct	Type II Error

P(Type I error)

- Suppose *a*-level = 0.05. Then, P(Type I error) = P(reject null, given it is true) = $P(|z| > 1.96) = 0.05$
- i.e., *the α-level is the P(Type I error).*
- Since we "give benefit of doubt to null" in doing test, it's traditional to take α small, usually 0.05 but 0.01 to be very cautious not to reject null when it may be true.
- As in CIs, don't make α *too* small, since as α goes down, β = P(Type II error) goes up

  (Think of analogy with courtroom trial)
- Better to report *P*-value than merely whether reject H_0

  (Are *P* = 0.049 and 0.051 really substantively different?)

  See homework 6.24

P(Type II error)

- P(Type II error) = b depends on the true value of the parameter (from the range of values in H_a).

- The farther the true parameter value falls from the null value, the easier it is to reject null, and P(Type II error) goes down. (see graph of null and alternative dist's)

- Power of test = 1 - β = P(reject null, given it is false)

- In practice, you want a large enough *n* for your study so that *P*(Type II error) is small for the size of effect you expect.

Example: Testing new treatment for anorexia

For a new treatment, we expect mean weight change = about 10 pounds, with std. dev. about 10. If our study will have $n = 20$, what is P(Type II error) if we plan to test H_0: $\mu = 0$ against H_a: $\mu > 0$, using $\alpha = 0.05$?

- We fail to reject H_0: $\mu = 0$ if we get P-value > 0.05
- We get P-value = 0.05 if test statistic $t = 1.729$
  (i.e., with $df = 19$, 0.05 is right-tail prob. above 1.729, so "rejection region" is values of $t > 1.729$)
- With $n = 20$, we expect a standard error of about

$$se = 10/\sqrt{20} = 2.24$$

- We get $t = 1.729$ if the sample mean is about
  $1.729(2.24) = 3.87$. i.e., $t = (3.87 - 0)/2.24 = 1.729$.

- So, we'll get $t < 1.729$ and P-value > 0.05 (and make a Type II error) if the sample mean < 3.87.

- But, if actually $\mu = 10$, a sample mean of 3.87 is about $(3.87 - 10)/2.24 = -2.74$ standard errors from μ
  (i.e., 2.74 std. errors below $\mu = 10$)

- When $df = 19$, the probability falling at least 2.74 standard errors below the mean is about 0.007. So, there's little chance of making a Type II error.

- But what if μ actually is only 5? (exercise; > 0.007 or < 0.007?)

Limitations of significance tests

- *Statistical significance* does not mean *practical significance* (Recall example on p. 17 of these notes)

- Significance tests don't tell us about the *size* of the effect (like a CI does)

- Some tests may be "statistically significant" just by chance

(and some journals only report "significant" results!)

Example: Are many medical "discoveries" actually Type I errors?

Reality: Most medical studies are "non-significant," not finding an effect.

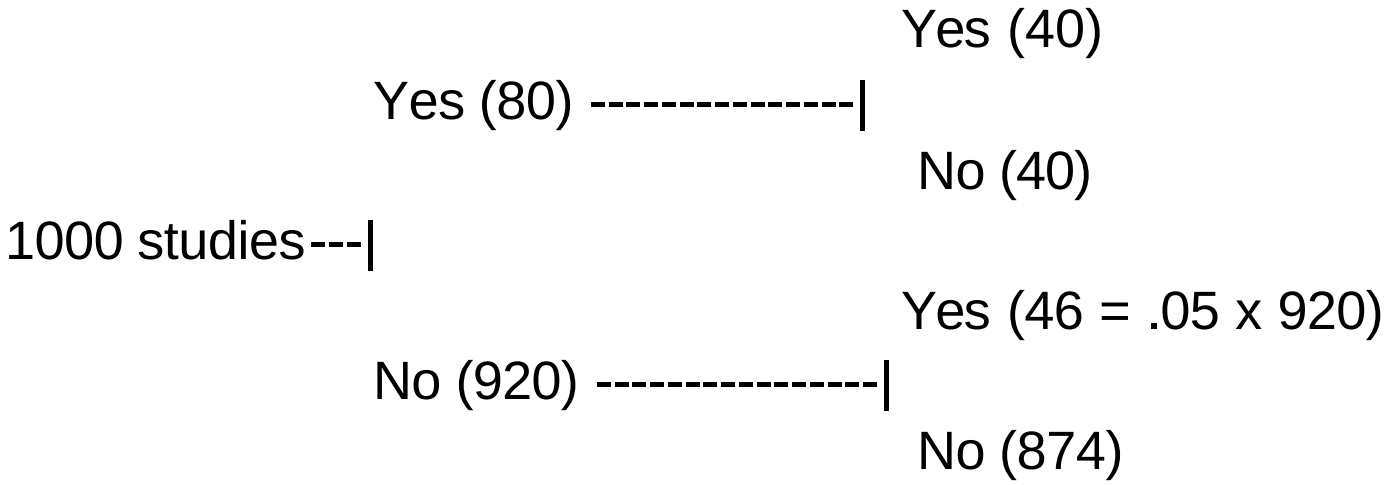
In medical research, when effects exist but are not strong, they may not be detected with the sample size practical for a study.

(A *British Medical Journal* article in 2001 estimated that when an effect truly exists, P(Type II error) = 0.50!)

In medical studies, suppose an effect actually exists 8% of the time. Could a substantial percentage of medical "discoveries" (i.e., significant results) actually be Type I errors?

Simple-minded solution: Draw a ***tree diagram*** to show what we'd expect to happen with many studies (say, 1000)

True effect? ***Decision Reject null?***



Of studies with rejected null hypo's, Type I error rate = 46/(46+40) = 0.53

Moral of the story: Be skeptical when you hear reports of new medical advances.

There may be *no* actual effect

(i.e. the entire study may merely be a Type I error!)

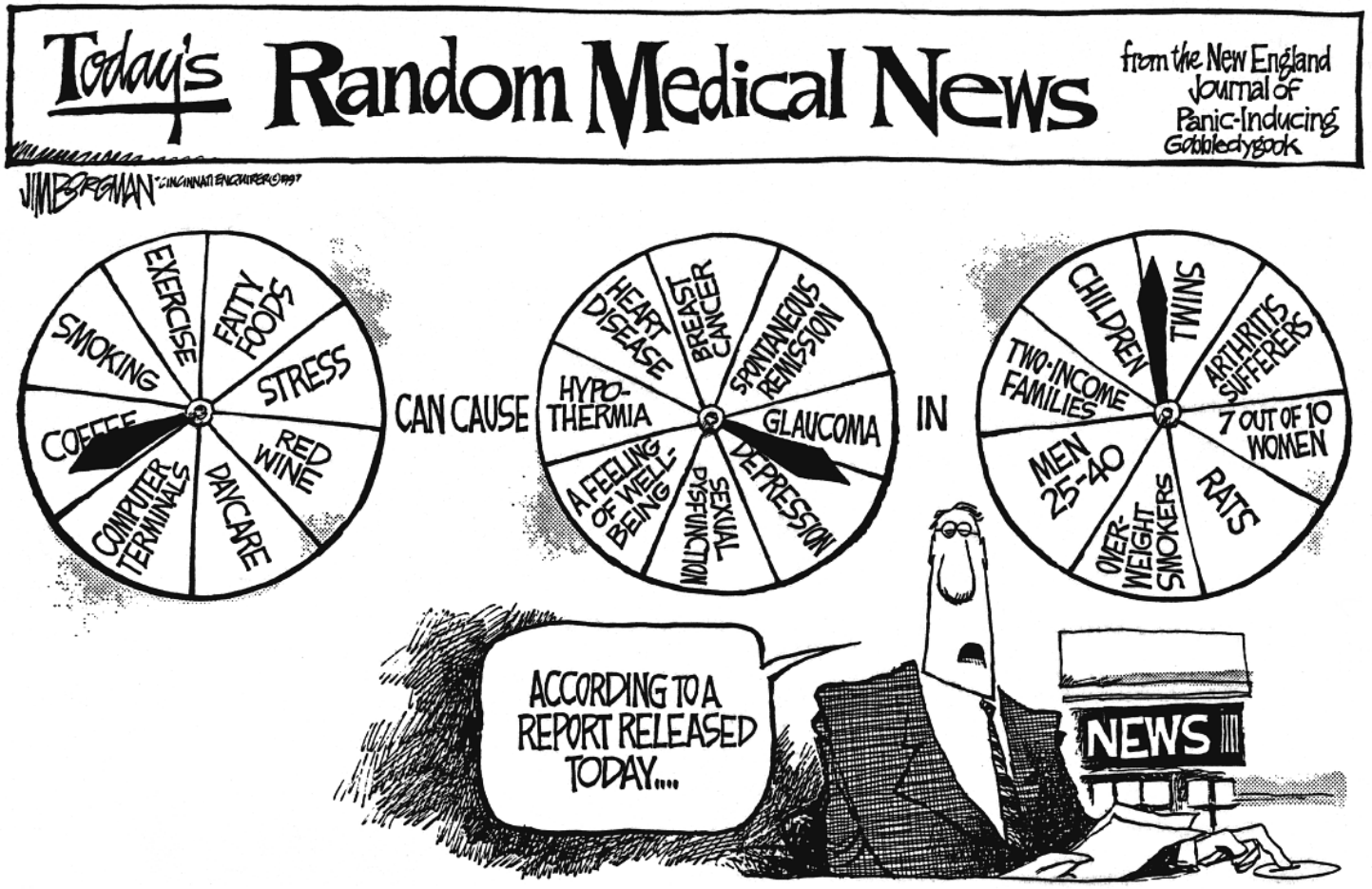
If an actual effect exists, we may be seeing a sample outcome in right-hand tail of sampling distribution of possible sample effects, and the actual effect may be *much weaker* than reported.

(picture of what I mean by this)

Actual case: A 1993 study estimated that injections of magnesium could double the chance of surviving a major heart attack.

A much larger later study of 58,000 heart attack patients found no effect at all.

Figure from Agresti and Franklin, *Statistics: The Art and Science of Learning from Data* (p. 468)



The binomial distribution

If

- Each *observation* is binary (one of two categories)
- Probabilities for each observation π for category 1,
  1-π for category 2
- Observations are independent

then for *n* observations, the number *x* in category 1 has

$$P(x) = \frac{n!}{x!(n-x)!} p^{x}(1-p)^{n-x}, \quad x = 0,1,...,n$$

This can be used to conduct tests about π when *n* is too small to rely on large-sample methods (e.g., when expected no. observations in either category < about 10)

Example: Exercise 6.33 (ESP)

- Person claims to be able to guess outcome of coin flip in another room correctly more often than not

- π = probability of correct guess (for any one flip)

- H_0: $\pi = 0.50$ (random guessing)
- H_a: $\pi > 0.50$ (better than random guessing)

- Experiment: $n = 5$ trials, $x = 4$ correct

Find the *P*-value, and interpret it.

The binomial distribution for $n = 5$, $\pi = 0.50$

$$P(0) = \frac{n!}{x!(n-x)!} p^{x}(1-p)^{n-x} = \frac{5!}{0!5!}(0.50)^{0}(0.50)^{5} = (0.50)^{5} = 1/32$$

$$P(1) = \frac{n!}{x!(n-x)!} p^{x}(1-p)^{n-x} = \frac{5!}{1!4!}(0.50)^{1}(0.50)^{4} = 5(0.50)^{5} = 5/32$$

$$P(2) = \frac{n!}{x!(n-x)!} p^{x}(1-p)^{n-x} = \frac{5!}{2!3!}(0.50)^{2}(0.50)^{3} = 10(0.50)^{5} = 10/32$$

$$P(3) = \frac{5!}{3!2!} 0.5^{3}(1-0.5)^{2} = 10/32$$

$$P(4) = \frac{5!}{4!1!} 0.5^{4}(1-0.5)^{1} = 5/32$$

$$P(5) = \frac{5!}{5!0!} 0.5^{5}(1-0.5)^{0} = 1/32$$

- For $H_a : \pi > 0.50$,

  P-value is probability of observed result or result even more extreme in right-hand tail

  $= P(4) + P(5) = 6/32 = 0.19$

There is not much evidence to support the claim.

We'd need to observe $x = 5$ in $n = 5$ trials to reject null at 0.05 level

(Then, *P*-value = $1/32 < 0.05$)

Notes about binomial distribution

- Binomial is the most important probability distribution for categorical data

- Binomial dist for *x* = number in category 1 has

$$m = E(x) = np, \quad s = \sqrt{np(1-p)}$$

whereas sample proportion $\hat{p} = x/n$ has

$$E(\hat{p}) = p, \quad s_{\hat{p}} = \sqrt{p(1-p)/n}$$

Example: Poll results for proportion with $n = 1000$, $\pi = 0.50$

- x = *number* in category of interest has

$$m = E(x) = np = 1000(0.50) = 500, \quad s = \sqrt{np(1-p)} = \sqrt{1000(0.50)(0.50)} = 15.8$$

- *proportion* $\hat{p}$ in category of interest has

$$E(\hat{p}) = p = 0.50, \quad s_{\hat{p}} = \sqrt{p(1-p)/n} = \sqrt{(0.50)(0.50)/1000} = 0.0158$$

(effect of n? Compare to $n = 10$)

Review significance test questions

Do a minority of Americans believe that same-sex marriage should be legal? Which is the appropriate alternative hypothesis?

a. $H_a: \pi = 0.50$

b. $H_a: \hat{p} > 0.50$

c. $H_a: \pi = 0.00$

d. $H_a: \pi < 0.50$

e. $H_a: \pi \neq 0.50$

What happens to P(Type II error)

1. when the actual population proportion gets closer to the null hypothesis value?
2. when you decrease the P(Type I error) from 0.05 to 0.01 for making the decision?

a. Decreases
b. Increases
c. Stays the same

Let's practice with one more problem (optional HW exercise 6.21)

Multiple-choice question, 4 choices. Test whether more correct answers than expected just due to chance (with random guessing of answer).

a. Set up hypotheses.

b. For 400 students, 125 get correct answer. Find *P*-value and interpret.