Testing Statistical Hypothesis (Section 9.4)

Set up:

Testing hypotheses: Verifying claims regarding unknown θ based upon the evidence provided by the data.

Hypotheses: Two *mutually exclusive* statements about θ.

- **Null hypothesis** H_0: Value of θ corresponding to "status quo", "common belief", "no change", etc. Often, $H_0 : \theta = \theta_0$ (a given value)
- **Alternative hypothesis** H_1: The claim the researcher is hoping to prove.

Three possible hypotheses in this course:

- (**Two-sided**) $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$
- (**One-sided, right-tailed**)
  $H_0 : \theta = \theta_0$ (or $\theta \leq \theta_0$) against $H_1 : \theta > \theta_0$
- (**One-sided, left-tailed**)
  $H_0 : \theta = \theta_0$ (or $\theta \geq \theta_0$) against $H_1 : \theta < \theta_0$

Let us set up the hypotheses in the following examples.

Ex 1: A long-time authorized user of a computer account takes 0.2 seconds on average between keystrokes. One day, when a user typed in the username and password, 15 times between keystrokes were recorded. These data had mean of 0.3 seconds and standard deviation of 0.12 seconds. Do these data given evidence of an unauthorized login attempt?

Ex 2: The number of concurrent users for an ISP has historically averaged 5000. After a marketing campaign, the management would like to know if it has resulted in an increase in the number of concurrent users. To test this, data were collected by observing the number of concurrent users at 100 randomly selected moments of time. Suppose that the average and the standard deviation of the sample data are 5200 and 800, respectively. Is there evidence that the mean number of concurrent users has increased?

Ex 3: A recent poll of 1,000 American people estimated that the approval rating of the current congress is 31%. Do these data give evidence that less than 30% of the American people approve the performance of the congress?

Outcome of a hypotheses test: Accept H_0 or reject H_0 (i.e., accept H_1)

- We do not know the truth. (If we knew, there was no point in collecting data.)
- H_0 is rejected **only if** there is strong evidence against it, otherwise H_0 is accepted.
- Evidence is provided by the data.
- If H_0 is accepted, it doesn't mean that H_0 is true. It just means that there is not enough evidence in the data to reject it.
- If H_0 is rejected, it doesn't mean that H_1 is true. It simply means that the data strongly favors H_1.
- Analogous to a court case.

Two types of errors:

	Truth	
Test outcome	H_0 is true	H_1 is true
Accept H_0		
Reject H_0		

- Trade-off between the two error probabilities. As one decreases, the other increases. So, it may not be possible for a procedure with a given sample size n to have both probabilities to be small.
- Hypotheses are set up in a way that ensures *type I error is more serious than type II error.*

- Design a test procedure that guarantees that its type I error probability does not exceed a small prescribed value α, known as the **level of significance** or simply the α level of the test.
- In practice, $\alpha = 0.01, 0.05$ (most popular), or 0.10.
- No guarantee of P(type II error). We try to keep it small by choosing a large enough n.
- Power of test = 1−P(type II error).
- Typically, the error probabilities depend on the true θ.

Analogy with a court case

A suspect is brought to the court — "presume innocent until proven guilty."

H_0:

H_1:

H_0 is rejected (i.e., the suspect is convicted) only if there is strong evidence against his/her innocence. Otherwise, H_0 is accepted (i.e., the suspect is acquitted).

Type I error:

Type II error:

Basic premise:

Q: How can we make $P(\text{type I error}) = 0$? What happens to $P(\text{type II error})$?

Q: How can we make $P(\text{type II error}) = 0$? What happens to $P(\text{type I error})$?

A general approach for get a level α test

- Estimate θ by its point estimator $\hat{\theta}$
- Compute s.e.$(\hat{\theta})$ assuming $\theta = \theta_0$. Estimate it if it's unknown.
- Compute a **test statistic** T that measures measures how consistent the data are with H_0. Often, T has the form:
- Find the **null distribution** — the distribution of T assuming H_0 is true.
- Find the form of the **rejection region** $\mathcal{R}$ — the set of values of T for which H_0 is rejected.
- **Acceptance region** $\mathcal{A}$ = Complement of $\mathcal{R}$.
- Determine $\mathcal{R}$ by ensuring that the level of significance of the test is α, i.e., $P(\text{reject } H_0 | H_0 \text{ is true}) = \alpha$.

Some common rejection regions

Suppose $T =$

In this case, it is often easy to guess $\mathcal{R}$.

Case 1: $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$

Case 2: $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$

Case 3: $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$

Compute the critical point in a way that ensures that the level of the test equals the prescribed α.

The corresponding level α tests:

Suppose c_α is such that $P(T > c_\alpha | \theta = \theta_0) = \alpha$.

Case 1: $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$

$\mathcal{R} = \{|T| > c_{\alpha/2}\}$, i.e., reject H_0 when $|T| > c_{\alpha/2}$, otherwise accept it.

Case 2: $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$

$\mathcal{R} = \{T > c_\alpha\}$, i.e., reject H_0 when $T > c_\alpha$, otherwise accept it.

Case 3: $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$

$\mathcal{R} = \{T < -c_\alpha\}$, i.e., reject H_0 when $T < -c_\alpha$, otherwise accept it.

Hypothesis testing (continued)

We can perform a level α test by comparing T_{obs} with the critical point. But how strong is the evidence against the null? This is formally measured by *p-value*. Let's play a game to motivate its definition.

My bag has 10 small balls. I claim that 8 are red and 2 are blue. I will bet 3 people a candy bar that a blue ball will come up. My chances are not very good but I will take them anyway.

trial #	PKC vs ?	color drawn	winner
1			
2			
3			

Q. Does is it seem reasonable that I would win ... times in 3 trials if the bag contained 2 blue balls?

Let's cast this problem as a test of hypothesis.

Hypotheses:

T and T_{obs}

Null distribution T:

Q. What is the actual chance of getting T_{obs} if H_0 is true? What does it indicate about H_0?

p-value: The probability of getting a T that is *as extreme or more extreme* than T_{obs} assuming that H_0 is true.

- Small p-value implies

- Smaller the p-value, stronger the evidence against H_0.
- Level α test: Reject H_0 if p-value $\leq \alpha$.
- Another interpretation of p-value: The smallest level of significance at which H_0 is rejected.
- Advantage of p-value over critical point:

Q. Is p-value $= P(H_0$ is true)?

- H_0 is either true or not true, but we don't know the truth. Certainly, H_0 is not a random quantity.
- p-value tells us how likely our T_{obs} is (or something more extreme) if H_0 is true.

Summary of steps in a hypothesis test:

- Formulate H_0 and H_1
- Find a test statistic T and get its null distribution
- Compute T_{obs}
- Use the null distribution to compute either the critical point or the p-value for the test.
- State your conclusion.

Some specific tests
One-sample tests for μ where $X \sim N(\mu, \sigma^2)$

Case 1: z**-test (known** σ^2**):** $H_0 : \mu = \mu_0$

Test statistic:

Critical point for the level α **test:**
One-sided alternative:

Two-sided alternative:

p-value:

H_1	reject when	p-value	computing p-value
$\mu \neq \mu_0$			
$\mu > \mu_0$			
$\mu < \mu_0$			

Case 2: t**-test (unknown** σ^2**):** $H_0 : \mu = \mu_0$

Test statistic:

Critical point for the level α **test:**
One-sided alternative:

Two-sided alternative:

p**-value:**

H_1	reject when	p-value	computing p-value
$\mu \neq \mu_0$	$\lvert t_{\text{obs}}\rvert \geq t_{n-1,\alpha/2}$	$P(\lvert t\rvert \geq \lvert t_{\text{obs}}\rvert \,\lvert H_0)$	$2(1 - F(\lvert t_{\text{obs}}\rvert))$
$\mu > \mu_0$	$t_{\text{obs}} \geq t_{n-1,\alpha}$	$P(t \geq t_{\text{obs}} \,\lvert H_0)$	$1 - F(t_{\text{obs}})$
$\mu < \mu_0$	$t_{\text{obs}} \leq -t_{n-1,\alpha}$	$P(t \leq t_{\text{obs}} \,\lvert H_0)$	$F(t_{\text{obs}})$

One-sample test for μ when X is nonnormal

Large-sample z-test: $H_0 : \mu = \mu_0$

- Need large n but works for mean of any (non-normal) population
- Use the z-test with test statistic
- When n is large, the null distribution is approximately $N(0, 1)$ due to central limit theorem.
- This test has approximate level α.

One-sample test for population prop p

- The large-sample z-test works because in this case $X \sim$ Bernoulli(p) and $E(X) = p$.
- Use the z-test with test statistic
- This test has approximate level α.

Ex 1: A long-time authorized user of a computer account takes 0.2 seconds on average between keystrokes. One day, when a user typed in the username and password, 15 times between keystrokes were recorded. These data had mean of 0.3 seconds and standard deviation of 0.12 seconds. Do these data given evidence of an unauthorized login attempt? Assume normality for the times between keystrokes and 5% level of significance.

Ex 2: The number of concurrent users for an ISP has historically averaged 5000. After a marketing campaign, the management would like to know if it has resulted in an increase in the number of concurrent users. To test this, data were collected by observing the number of concurrent users at 100 randomly selected moments of time. Suppose that the average and the standard deviation of the sample data are 5200 and 800, respectively. Is there evidence that the mean number of concurrent users has increased? Assume 5% level of significance.

Ex 3: A recent poll of 1,000 American people estimated that the approval rating of the current congress is 31%. Do these data give evidence that less than 30% of the American people approve the performance of the congress? Assume 5% level of significance.

Two-sample tests for $\mu_X - \mu_Y$ for normal populations

<u>**Set up:**</u> $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$

- X sample: $X_1, \ldots, X_n$ — i.i.d. as X
- Y sample: $Y_1, \ldots, Y_m$ — i.i.d. as Y
- $H_0 : \mu_X - \mu_Y = \Delta$, where Δ is given and may be zero

<u>**Case 1:**</u> Paired samples, i.e., (X_i, Y_i) comes from subject $i = 1, \ldots, n$.

- $D = X - Y \sim N(\mu_D, \sigma_D^2)$ where
- Define the differences $D_i = X_i - Y_i$ — i.i.d. as
- Apply one-sample procedures to the differences — **paired z-test** or **paired t-test**

<u>Case 2:</u> Independent samples with known variances σ_X^2 & σ_Y^2

Test statistic:

- Know how to get critical points and p-values for z-test
- **Two-sample z-test**

Case 3: Independent samples with unknown variances σ_X^2 & σ_Y^2

Test statistic:

- Know how to get get critical points and p-values for a t-test
- Approximate **Two-sample t-test**
- No assumption regarding equality of variances

Case 4: Independent samples with unknown but equal variances $\sigma_X^2 = \sigma_Y^2$

Estimation of common variance σ^2:

Test statistic:

- Know how to get get critical points and p-values for a t-test
- **Two-sample t-test**

Two-sample tests for $\mu_X - \mu_Y$ for non-normal populations

Set up: Same as before but the populations are non-normal

Test statistic:

- Large-sample z-test. Its level is approximately α

Two-sample test for difference in proportions, $p_X - p_Y$

As before, apply large-sample z-test because the proportions can be interpreted as means of Bernoulli populations. Can also use *pooled sample proportion* in case of $H_0 : p_X = p_Y$ as suggested by the book.

Test statistic:

- The level of the test is approximately α