Testing Statistical Hypothesis (Section 9.4)

- Statistical inference provides methods for drawing conclusions about a population from sample data.
- Two of the most common types of statistical inference:
  1. Confidence intervals: Goal is to estimate a population parameter.
  2. Tests of significance: Goal is to assess the evidence provided by the data about some claim concerning the population parameter.
- Let's play a game to see how tests of significance works! My bag has 10 small balls (8 red, 2 blue).
- I will bet 3 people- Director, Dean and Registrar a candy bar that a blue ball will come up. I know my chances aren't very good, but I will take them anyway..

Introduction Continued

Game 1: Bhargab vs Director
Color drawn: Blue
Winner: Bhargab

Game 2: Bhargab vs Dean
Color drawn: Blue
Winner: Bhargab

Game 3: Bhargab vs Registrar
Color drawn: Blue
Winner: Bhargab

- Suppose I won in all three games. Does it seem reasonable if the bag really contains 8 red balls and 2 blue balls that I would win 3 times in 3 trials?

Basic Idea of Tests of Significance

- The actual chance of me winning 3 times (if the bag really contains what I claim) is 0.008.
- Moral of the story: An outcome that would rarely happen if a claim (hypothesis) were true is good evidence that the claim is not true.
- With this in mind, off we go to the Land of Significance Test
- Example: Every week, two teams (one made up of husbands, the other made up of escaped convicts — the "In-laws" and the "Outlaws") get together and play football. They decide who receives the kickoff by a flip of a fair coin provided by the In-laws (heads = outlaws receive, tails = in-laws receive).

Hypotheses

- Well, after losing the coin toss for a long time, the Outlaws have become suspicious!
- Being the well-read statisticians that they are, the Outlaws decide to see if the "fair" coin really is fair (They have little trouble stealing it from the In-laws ... but they got it eventually!)
- They then flip the coin 100 times on the sidelines, and get 35 heads.
- Outlaws expected .... heads but got 35 heads. They are furious!
- Two possible explanations:
  1. The coin is fair and this outcome just happened by chance (bad sample!)
  2. For a fair coin, this outcome is so unlikely (extreme) that we can conclude that the coin is not fair.

Hypotheses Contd.

So, we have two competing hypothesis:

- Null Hypothesis (H_0): The chance of heads = 50% (This is the statement of "status quo": It says the observed outcome is different from the expected one just by chance variation). Or, $p = 0.5 = p_0$
- Alternative Hypothesis (H_A): The chance of heads < 50%. (This is the statement we'd like to prove. It says chance variation is not enough to explain the outcome).

For a general scenario, 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 $H_1 : \theta < \theta_0$

Types of Error in Hypothesis Tests

- When we carry out a test, what types of errors we can make?

	Decision	
Truth (unknown) $\downarrow$	Reject H_0	Do not reject H_0
H_0		
H_A		

- Type I error: Reject H_0 when H_0 is actually true.
- Type II error: Not reject H_0 when H_A is actually true.

Probabilities of Type I and Type II Errors Contd.

- These errors are defined conditional on the true status (H_0 or H_1).
- We cannot eliminate the possibility of errors because our decision is based on a sample, and not the whole population.
- But we can have some control over the chances of making errors.
- Fact: P(Type I error) $= \alpha$(the significance level used).
- 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).

Probabilities of Type I and Type II Errors Contd.

Philosophy of Hypothesis Test: "Innocent until proven guilty". Assume null hypothesis to be true and see if there is convincing evidence in the data to prove otherwise.

- In a court trial, H_0: Person is Innocent vs. H_1: Person is Guilty
  Type I error:
  Type II error:
- If we try to decrease P(type I error) (by concluding persons under trial innocent more often), what happens to P(type II error)?
- Type I error is typically considered more severe and so it's chance is controlled by setting α to be a small.
- After setting α to be small, the studies are designed in such a way that type II error is minimized.

Test Statistic

- To decide between the null and alternative hypothesis, we find two things:
  1. how much different the observed outcome is from the expected outcome if the null hypothesis is true (test statistic)
  2. what is the chance of that type of difference occurring if null hypothesis is true (p-value).
- If such a difference is very unlikely to occur if null hypothesis is true, then we reject/do not reject the null hypothesis.
- Back to In-laws vs. Outlaws: As we are dealing with %, our test statistic under H_0:
  $Z = \frac{\text{observed(\%) - EV(\%)}}{\text{SE(\%)}} = -3$
- So, the observed outcome is -3 SD below the expected outcome. Is this an unlikely (extreme) outcome?

P-Value

- What is the chance of getting such an extreme (35% heads in 100 tosses) or more extreme outcome (less than 35%) if H_0 is true? Ans.0.00135.
- P-value: The chance of getting a test statistic as extreme or more extreme than what we observed if H_0 is true.
- Note: "Extreme" is in the direction of H_1.
- This p-values tell us: If H_0 is true (i.e., the coin is fair), then the chance of getting 35% or less heads in 100 tosses is 0.00135
- This tell us that if H_0 is true, the chance of getting a test statistic as extreme as -3 is very low.
- As our p-value is so small, we conclude that we have enough evidence to reject the null. That is, H_0 is not a reasonable explanation of the observed data.

P-Value Interpretation

P-value is NOT the probability that H_0 is true. H_0 is either true or not true. It does not vary from sample to sample.
P-value tells how likely it is to get the observed sample (or something more extreme) if H_0 is true (Note: H_0 is held fixed).
Smaller the P-value, stronger the evidence against H_0.

P-value Approach

- For p-value, how small is small enough to reject H_0? This cutoff is called "significance level" and is denoted by α.
- Typical choices of α are 0.05 or 0.01.
- Steps in Hypothesis Test:
  1. State H_0 and H_1 (H_1 is what we are interested in proving. It can be one-sided or two-sided).
  2. Calculate the appropriate test statistic (assuming H_0 to be true).
  3. Calculate the p-value.
  4. Use the p-value and the given significance level to draw the conclusion: If p-value $\leq \alpha$, reject H_0. If p-value $> \alpha$, do not reject H_0.

Outcome of a hypotheses test

We either 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 the court case.

Critical-Region approach

- 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:

$$T = \frac{\hat{\theta}-\theta}{\sqrt{V(\hat{\theta})}}$$

- 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.

Some specific tests

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

$\underbrace{X \sim N(\mu, \sigma^2)} \to x_1, x_2, \ldots, x_n$

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

Test statistic: $Z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Null distribution of Z : $Z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$

Critical point for the level α test:

One-sided alternative: $H_1: \mu > \mu_0$

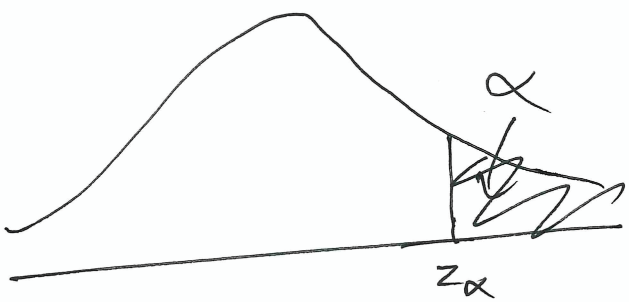
$H_1: \mu < \mu_0$

Two-sided alternative: $H_1: \mu \neq \mu_0$

p-value:

H_1	reject when	p-value	computing p-value
$\mu \neq \mu_0$	$\lvert z_{obs}\rvert > z_{\alpha/2}$	$P(Z > \lvert z_{obs}\rvert) + P(Z < -\lvert z_{obs}\rvert)$	2 pnorm $(-\lvert z_{obs}\rvert)$ $= 2[1-\text{pnorm}(\lvert z_{obs}\rvert)]$
$\mu > \mu_0$	$z_{obs} > z_{\alpha}$	$P[Z > z_{obs}]$	$1-$ pnorm $(z_{obs}) = \int_{z_{obs}}^{\infty} f(z)\,dz$
$\mu < \mu_0$	$z_{obs} < -z_{\alpha}$	$P[Z < z_{obs}]$	pnorm $(z_{obs}) = \int_{-\infty}^{z_{obs}} f(z)\,dz$

$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$



z_α: Critical point of $N(0,1)$

$P(Z > z_\alpha) = \alpha$

$P(Z < -z_\alpha) = \alpha$ $\qquad Z \sim N(0,1)$

$P(|Z| > z_{\alpha/2}) = \alpha$

z_{obs}: Observed value of $\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$.

$f(z)$: pdf of $N(0,1)$.

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

Test statistic: $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$, $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

$T \sim t_{n-1}$ for all n.

Critical point for the level α test:

One-sided alternative: $H_1 : \mu > \mu_0$

$H_1 : \mu < \mu_0$

Two-sided alternative: $H_1 : \mu \neq \mu_0$

p-value:

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

$F(x)$: pt $\leftarrow$ R function.

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

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim N(0,1), \text{ under } H_0$$

- 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 $E(X) = p$, $V(X) = p(1-p)$
- Use the z-test with test statistic $E(\bar{X}) = p$, $V(\bar{X}) = \frac{p(1-p)}{n}$.

$H_0: p = p_0$ ag $H_1: p \neq p_0$

$$Z = \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1), \text{ under } H_0.$$

- 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.

$n = 15$

X : Time between Keystrokes.

$\bar{x} = 0.3$, $n = 15$, $s = 0.12$

$H_0 : \mu = 0.2$ ag $H_1 : \mu \neq 0.2$, $\mu_0 = 0.2$

$$t = \frac{\sqrt{n}\,(\bar{x} - \mu_0)}{s}$$

Use t.test.

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.

$H_0: \mu = 5000$ against $H_1: \mu > 5000$

$\bar{x} = 5200$

$s = 800$

$n = 100$

1-sample Z-test

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.

$n = 1000$, $\alpha = 0.05$

$\hat{p} = 0.31$

$H_0: p = 0.30$ $\quad$ $H_1: p < 0.30$

~~1-samp~~ 1-prop Z-test.

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

Set up: $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$

- 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

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

$m = n$

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

$\sigma_D^2 = V[X-Y] = \sigma_X^2 + \sigma_Y^2 - 2\underbrace{\mathrm{cov}(X,Y)}_{\sigma_{XY}}$

$H_0: \mu_D = 0$ ag $H_1: \mu_D > 0$, < 0, $\neq 0$

Case 2: Independent samples with known variances σ_X^2 & σ_Y^2

$H_0: \mu_X - \mu_Y = \Delta$

Test statistic:

$$Z = \frac{\bar{X} - \bar{Y} - \cancel{\Delta}(\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0, 1)$$

obs. Z using $\bar{X}, \bar{Y}, \Delta, \sigma_X^2, \sigma_Y^2, m, n$.

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

Follow page 4 of HW #6

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

Test statistic: $H_0: \mu_X - \mu_Y = \Delta$

~~$n < 30, m < 30$~~

If at least one sample size being less than 30

$$T = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \sim t_\nu,$$

ν is computed by Satterthwaite's approximation.

- 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

Both $n, m \geq 30$

$$Z = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \sim N(0,1)$$

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

Estimation of common variance σ^2: we use pooled variance,

$$\cancel{\sigma^2}\ S_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

Test statistic: (Atleast one sample size < 30),

$$T = \frac{(\bar{X}-\bar{Y})-\Delta}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$$

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

If both sample sizes (>30),

$$Z = \frac{\bar{X}-\bar{Y}-\Delta}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1)$$

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:

See page 4 of HW #6

- 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 we the populations here follow Bernoulli distributions. Use *pooled sample proportion* in case of $H_0 : p_X = p_Y$.

$H_0: p_X - p_Y = \Delta$

Test statistic: Using sample, $\hat{p}_X$, $\hat{p}_Y$ ← based on m obs.

$\hat{p}_X$ ↓ based on n obs.

If $n, m \geq 30$

$$Z = \frac{\hat{p}_X - \hat{p}_Y - \Delta}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{m}}} \sim N(0,1)$$

- The level of the test is approximately α

$H_0: p_X = p_Y$. $(p_X = p_Y = p)$. p can be estimated by $\hat{p} = \frac{n\hat{p}_X + m\hat{p}_Y}{n+m}$

$$Z = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{p(1-p)\left(\frac{1}{n}+\frac{1}{m}\right)}} \sim N(0,1)$$

Test for normal variance, $H_0 : \sigma^2 = \sigma_0^2$

Test statistic: $c = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$

$H_0: \sigma^2 = \sigma_0^2$ ag $H_1: \sigma^2 > \sigma_0^2$

$H_1: \sigma^2 < \sigma_0^2$

$H_1: \sigma^2 \neq \sigma_0^2$

Critical point for the level α test:

One-sided alternative:

Two-sided alternative:

- **Chi-square test**

Page 6 of HW #6

Test for equality of two normal variances, $H_0 : \sigma_X^2 = \sigma_Y^2$

Test statistic: $F = \frac{S_X^2}{S_Y^2} \sim F_{n-1, m-1}$

$H_0: \sigma_X^2 = \sigma_Y^2$ og $H_1: \sigma_X^2 > \sigma_Y^2$

$\sigma_X^2 < \sigma_Y^2$

$\sigma_X^2 \neq \sigma_Y^2$

- Critical points and p-values are computed as in the F ~~chi-square~~ test.
- F-test for equality of two variances