Statistics: Hypothesis tests

1. Reminder: point estimation
2. Hypothesis tests
3. Significance tests

Reminder: point estimation

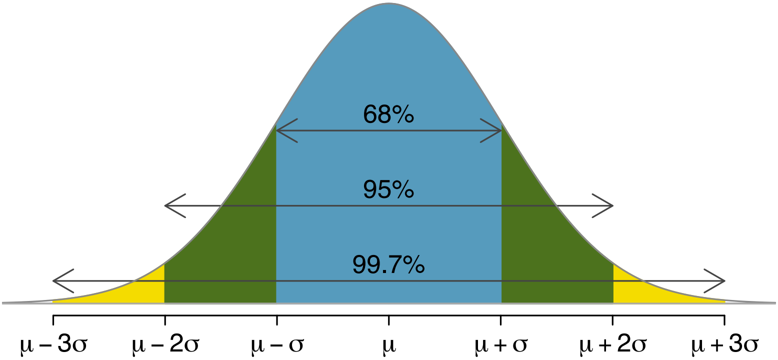
Sample definitions

- the population mean μ is a population parameter
- the sample mean $\bar{X}$ is a point estimate of μ
- we know the sample n and its mean $\bar{X}$, but we do not know μ and might not know the true population N

Sampling error

- sampling variation causes $\bar{X}$ to vary
- the values of $\bar{X}$ form a sampling distribution
- its standard deviation $\frac{\sigma}{\sqrt{N}}$ is the standard error of the mean (SEM), which is estimated from the sample by $\frac{s}{\sqrt{n}}$.

Standard normal distribution



Source: Diez *et al.* 2011

Standard normal distribution

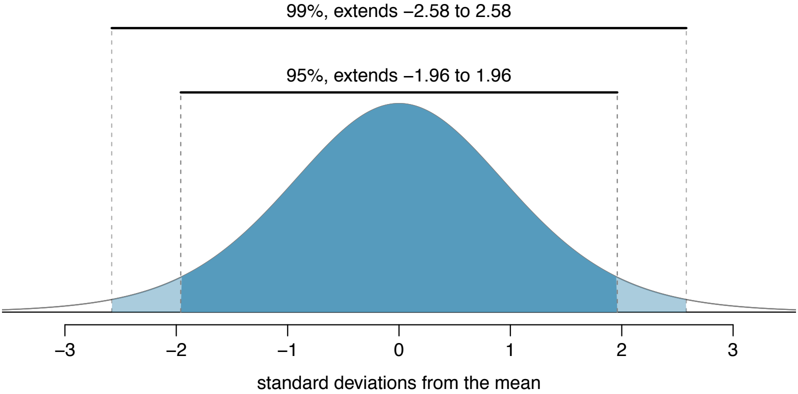


Figure 4.10: The area between -$z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

Generalization of the standard normal distribution

Central Limit Theorem (CLT)

For ‘iid’ (independent and identically distributed) random variables $X_1, X_2, ..., X_n$, the sampling distribution of the mean approximates a normal distribution as $n > 30$ increases.

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\bar{X}_i - \mu\right) \xrightarrow{d} \mathcal{N}(0,\ \sigma^2)$$

Law of Large Numbers (LLN)

$$\frac{X_1 + X_2 + ... + X_n}{n} = \mu$$

Estimation of confidence intervals

Confidence intervals

If the sampling distribution is approximately normal, fractions of the point estimates are contained within the mean $\bar{X} \pm Z$-scores:

- For a 95% CI: $\bar{X} - 1.96 \cdot SEM, \bar{X} + 1.96 \cdot SEM$
- For a 99% CI: $\bar{X} - 2.58 \cdot SEM, \bar{X} + 2.58 \cdot SEM$

The margin of error of the interval is $Z \cdot SEM$.

Accuracy trade-off

Wider intervals trade precision for additional confidence:

- A 95% CI is smaller but less reliable than a 99% CI.
- A 99% CI is larger but more reliable than a 95% CI.

Neither level of confidence can ensure that $\mu \in$ CI.

Hypothesis tests with 95% CIs, H_0 and H_a

Comparison of means x and y

For two *independent* groups:

1. Hypothesize that $\bar{x} \neq \bar{y}$ at a given level of confidence
2. Compute the 95% CI for $\bar{y}$ to assess the difference

The hypothesis will be tested against $H_0 : \bar{x} = \bar{y}$.

General logic of testing the null hypothesis H_0

- Proof by contradiction: the goal is the test is to reject H_0
- Type I error: rejecting H_0 while it is actually true
- Type II error: accepting H_0 while it is actually true

The alternative hypothesis H_a is tested against $H_0 : \bar{x} = \bar{y}$.

Working examples

Average donations to political parties

- Group 1 (2008): $\bar{x} = 35$, other parameters unobserved
- Group 2 (2012): $\bar{y} = 45, s = 5, n = 100$

Working examples

Average donations to political parties

- Group 1 (2008): $\bar{x} = 35$, other parameters unobserved
- Group 2 (2012): $\bar{y} = 45, s = 5, n = 100$
- 95% CI for $\bar{y} = 45 \pm 1.96 \cdot \frac{5}{\sqrt{100}} = [44.02, 45.98]$

There is strong evidence that $\bar{x} \neq \bar{y}$: we can reject H_0.

Working examples

Average donations to political parties

- Group 1 (2008): $\bar{x} = 35$, other parameters unobserved
- Group 2 (2012): $\bar{y} = 45, s = 5, n = 100$
- 95% CI for $\bar{y} = 45 \pm 1.96 \cdot \frac{5}{\sqrt{100}} = [44.02, 45.98]$

There is strong evidence that $\bar{x} \neq \bar{y}$: we can reject H_0.

Average support for income equality

- Group 1 (Whites): $\bar{X}_1 = 75, s = 10, N = 900$
- Group 2 (Blacks): $\bar{X}_2 = 85, s = 45, N = 81$

Working examples

Average donations to political parties

- Group 1 (2008): $\bar{x} = 35$, other parameters unobserved
- Group 2 (2012): $\bar{y} = 45, s = 5, n = 100$
- 95% CI for $\bar{y} = 45 \pm 1.96 \cdot \frac{5}{\sqrt{100}} = [44.02, 45.98]$

There is strong evidence that $\bar{x} \neq \bar{y}$: we can reject H_0.

Average support for income equality

- Group 1 (Whites): $\bar{X}_1 = 75, s = 10, N = 900$
- Group 2 (Blacks): $\bar{X}_2 = 85, s = 45, N = 81$
- 95% CI for $\bar{X}_1 \approx [74.3, 75.6]$; 95% CI for $\bar{X}_2 \approx [75, 95]$

There is no strong evidence that $\bar{X}_1 \neq \bar{X}_2$: we should retain H_0.

Contextual Type I and II Errors

Type I Error in judicial trials

"Last year executed man proven innocent by DNA evidence."

- H_0: presumption of innocence
- H_a: ... until proven guilty (H_0 wrongly rejected)

Type II Error in child protection

"Violent father beats children after being released from custody."

- H_0: parents considered responsible
- H_a: ... until proven abusive (H_0 wrongly retained)

Formal significance tests

1. Write up the null hypothesis as an equality

- $H_0 : \mu = 0$, or $\bar{x} - \bar{y} = 0$, or $\bar{X} = k$, ...
- One-sided hypothesis test: $H_a : \mu > 0$, or $H_a : \mu < 0$, ...
- Two-sided hypothesis test: $H_a : \mu \neq 0$, or $\bar{X} \neq k$, ...

2. Declare a level of statistical significance

- $\alpha = 0.05$ for a 95% CI ($Z = 1.96 \approx 2$)
- $\alpha = 0.01$ for a 99% CI ($Z = 2.58 \approx 2.5$)
- Verify whether $Pr(H_0) < \alpha$ so that you can reject H_0

$Pr(H_0)$ is the *p*-value of the test.

One-sided significance test

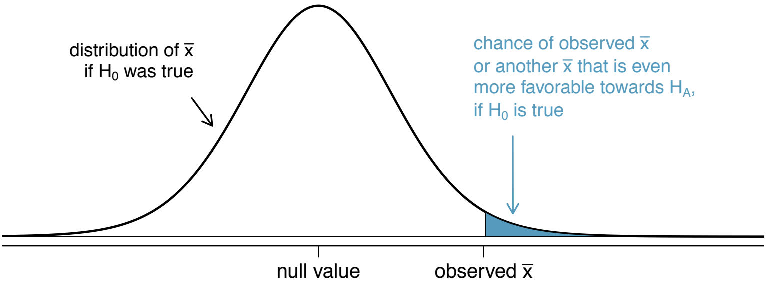


Figure 4.16: To identify the p-value, the distribution of the sample mean is considered as if the null hypothesis was true. Then the p-value is defined and computed as the probability of the observed $\bar{x}$ or an $\bar{x}$ even more favorable to H_A under this distribution.

Source: Diez *et al.* (2011)

Two-sided significance test

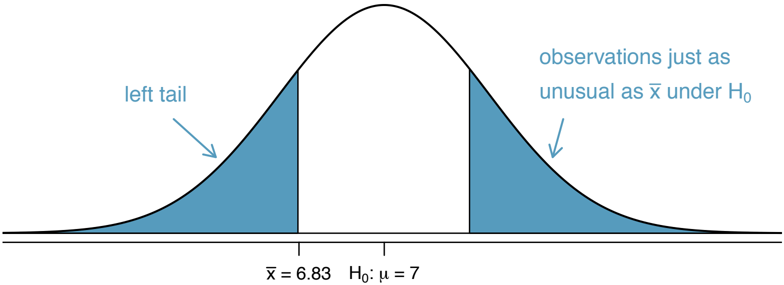


Figure 4.18: H_A is two-sided, so *both* tails must be counted for the p-value.

Source: Diez *et al.* (2011)

Working examples

Income gap

What is H_0 for a comparison of average income between males and females?

Working examples

Income gap

What is H_0 for a comparison of average income between males and females? $H_0 : \bar{X}_{males} = \bar{X}_{females}$

Working examples

Income gap

What is H_0 for a comparison of average income between males and females? $H_0 : \bar{X}_{males} = \bar{X}_{females}$

Sleeping time

What is H_a for a comparison of average sleeping time between night workers, who suffer from broken sleeping patterns, and day workers?

Working examples

Income gap

What is H_0 for a comparison of average income between males and females? $H_0 : \bar{X}_{males} = \bar{X}_{females}$

Sleeping time

What is H_a for a comparison of average sleeping time between night workers, who suffer from broken sleeping patterns, and day workers? $H_a : \bar{X}_{nightworkers} < \bar{X}_{dayworkers}$

Working examples

Income gap

What is H_0 for a comparison of average income between males and females? $H_0 : \bar{X}_{males} = \bar{X}_{females}$

Sleeping time

What is H_a for a comparison of average sleeping time between night workers, who suffer from broken sleeping patterns, and day workers? $H_a : \bar{X}_{nightworkers} < \bar{X}_{dayworkers}$

Work fun

What is H_a for a comparison of average employee productivity with and without a firewall to block Facebook access?

Working examples

Income gap

What is H_0 for a comparison of average income between males and females? $H_0 : \bar{X}_{males} = \bar{X}_{females}$

Sleeping time

What is H_a for a comparison of average sleeping time between night workers, who suffer from broken sleeping patterns, and day workers? $H_a : \bar{X}_{nightworkers} < \bar{X}_{dayworkers}$

Work fun

What is H_a for a comparison of average employee productivity with and without a firewall to block Facebook access?
$H_a : \bar{X}_{withoutFacebook} > \bar{X}_{withFacebook}$

Sanity checks

Sample requirements

- The data must come from a simple random sample
- The observations should be independent
- The observations should be normally distributed
- Sample size should be approximately at least $N = 30$

Strong skewness or clear outliers will violate normality at low N. Randomness and independence might fail for other reasons.

Rejecting H_0

- $Pr(H_a) \neq 1 - Pr(H_0)$: you never get to measure $Pr(H_a)$
- Statistical significance does not imply substantive significance

Sample size

Determining a margin of error

$$Z \cdot \frac{\sigma}{\sqrt{N}} \leq ME$$

Example: average number of sexual partners

How many people should we sample from a population where the number of sexual partners has an unknown mean and a standard deviation of 5 if we want a margin of error around 2 partners at 95% confidence?

Sample size

Determining a margin of error

$$Z \cdot \frac{\sigma}{\sqrt{N}} \leq ME$$

Example: average number of sexual partners

How many people should we sample from a population where the number of sexual partners has an unknown mean and a standard deviation of 5 if we want a margin of error around 2 partners at 95% confidence?

$$1.96 \cdot \frac{\sigma}{\sqrt{N}} \leq 2 \quad 1.96 \cdot \frac{5}{2} \leq \sqrt{N} \quad (1.96 \cdot \frac{5}{2})^2 \leq N \quad N \geq 25$$

Reading significance tests

What the test is about

- Comparison of means: $\bar{X}_{females} - \bar{X}_{males}$
- Comparison of proportions: $\hat{p}_{Blacks} - \hat{p}_{White}$
- H_0 and H_a: equality, increase or decrease among groups
- Means and difference given with confidence intervals

How to read the *p*-value

- Comparison of means: t-test
- Comparison of proportions: proportions test
- Two-tailed test: $H_a : difference \neq 0$
- One-tailed test: $H_a : difference > 0$ or $H_a : difference < 0$

```
. ttest bmi if raceb==3 & age < 35, by(female)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Male |     731    27.60214    .1622325    4.386283    27.28364    27.92064
  Female |     906    26.70284    .1740403    5.238583    26.36127    27.04441
---------+--------------------------------------------------------------------
combined |    1637    27.10442    .1209956    4.895466    26.86709    27.34174
---------+--------------------------------------------------------------------
    diff |             .899302    .2424424                .4237716    1.374832
------------------------------------------------------------------------------
    diff = mean(Male) - mean(Female)                              t =   3.7093
Ho: diff = 0                                     degrees of freedom =     1635

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9999         Pr(|T| > |t|) = 0.0002          Pr(T > t) = 0.0001
```

```
. ttest bmi if raceb==3 & age < 35, by(female)
```

difference in means
p-value of the null hypothesis

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Male	731	27.60214	.1622325	4.386283	27.28364	27.92064
Female	906	26.70284	.1740403	5.238583	26.36127	27.04441
combined	1637	27.10442	.1209956	4.895466	26.86709	27.34174
diff		.899302	.2424424		.4237716	1.374832

diff = mean(Male) - mean(Female) t = 3.7093
Ho: diff = 0 degrees of freedom = 1635

Ha: diff < 0	Ha: diff != 0	Ha: diff > 0
Pr(T < t) = 0.9999	Pr(|T| > |t|) = 0.0002	Pr(T > t) = 0.0001

Reading guide for a t-test (with Stata software)

Top table

- Group 1 ($N = 731$ males) has a mean BMI of 27.6
- Group 2 ($N = 906$ females) has a mean BMI of 26.7
- `diff` is the difference in means $\delta_{males-females} = .89$
- Columns show the standard error (`Std. Err.`), standard deviation (`Std. Dev.`) and 95% confidence intervals

How to read the p-value

- Central p-value for $H_0 : \delta = 0$: `0.0002`
- H_0 is very unlikely ($p < 0.01$): reject the null hypothesis

```
. prtest torture in 2000/2800, by(female)

Two-sample test of proportions                     Male: Number of obs =      360
                                                 Female: Number of obs =      341
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Male |       .725   .0235334                      .6788754    .7711246
      Female |    .771261   .0227454                      .7266808    .8158412
-------------+----------------------------------------------------------------
        diff |   -.046261   .0327288                     -.1104082    .0178862
             |  under Ho:   .0328295    -1.41   0.159
------------------------------------------------------------------------------
        diff = prop(Male) - prop(Female)                          z =  -1.4091
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0794         Pr(|Z| < |z|) = 0.1588          Pr(Z > z) = 0.9206
```

```
. prtest torture in 2000/2800, by(female)

Two-sample test of proportions                     Male: Number of obs =      360
                                                 Female: Number of obs =      341

------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Male |       .725   .0235334                          .6788754    .7711246
      Female |    .771261   .0227454                          .7266808    .8158412
-------------+----------------------------------------------------------------
        diff |   -.046261   .0327288                         -.1104082    .0178862
             |  under Ho:   .0328295    -1.41   0.159
------------------------------------------------------------------------------
        diff = prop(Male) - prop(Female)                          z =  -1.4091
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0794         Pr(|Z| < |z|) = 0.1588          Pr(Z > z) = 0.9206
```

difference ~ 4 percentage points

p-value above alpha:
retain null hypothesis

Homework

Read CK-12 handbook ch. 8 for the exam

and enjoy the rest of your semester.

Note: final stats exam will cover confidence intervals (Ch. 7) and hypothesis tests (Ch. 8). Histograms are part of the topic.