MTH220 Summary Sheet

Tags

Created by Ho Han Sheng

▼ Central Limit Theorem Applications (CLT)

If $X_1, X_2, ..., X_n$ are independent identically distributed random variables

With mean μ and variance σ^2, then the sample mean

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

for large n

▼ Mean

$$E(\overline{X}) = \mu$$

▼ Variance

$$V(\overline{X}) = \frac{\sigma^2}{n}$$

▼ CLT Properties

If random samples are drawn from a population on different occasions, then the individual observations and the sample means will very probably be different. The sample mean is itself a random variable and its distribution is called the sampling distribution of the mean.

CLT states that: For large n, the sampling distribution of the mean for samples of size n from a population with mean μ and variance σ^2 is approximately normal with mean μ and variance σ^2/n . The approximation improves as the sample size increases.

▼ Sample Total

The distribution of the sample total has mean $n\mu$ and variance $n\sigma^2$and is also approximately normal. This means that the distribution of the sum of a large number n of independent identically distributed random variables is approximately normal.

▼ Expected value of sample total and sample mean

$$E(T_n) = n\mu$$

$$E(\overline{X}_n) = \mu$$

▼ Variance of sample total and sample mean

$$V(T_n) = n\sigma^2$$

$$V(\overline{X}_n) = \frac{\sigma^2}{n}$$

▼ Standard error of the mean

This is the standard deviation of the sampling distribution of the mean

$$\sigma/\sqrt{n}$$

▼ Normal approximation for Binomial Distribution

Binomial distribution $B(n,p)$ may be usefully approximated

By a normal distribution with same mean and variance, $N(np, npq)$

When both np and nq are at least 5

$(q = 1 - p)$

$$X \sim B(n,p) \Rightarrow X \sim N(\mu = np, \sigma^2 = np(1-p))$$

▼ Continuity Correction

When we approximate discrete random variable X by a normal random variable Y, we need to apply this continuity correction

Number line visualisation

$$P(X \leq x) \approx P(Y \leq x + 0.5)$$

$$P(X = x) \approx P(x - 0.5 \leq Y \leq x + 0.5)$$

Desired information	With continuity correction
$P(X = x)$	$P(x - 0.5 \leq X \leq x + 0.5)$
$P(X \leq x)$	$P(X \leq x + 0.5)$
$P(X < x) = P(X \leq x$ - 1)	$P(X \leq x - 1 + 0.5)$
$P(X \geq x)$	$P(X \geq x - 0.5)$
$P(X > x) = P(X \geq x + 1)$	$P(X \geq x +$ 1- 0.5)
$P(a \leq X \leq b)$	$P(a - 0.5 \leq X \leq b + 0.5)$

▼ Normal approximation for Poisson Distribution

Poisson distribution: $Poisson(\mu)$ may be usefully approximated by

A normal distribution with same mean and variance, $N(\mu, \mu)$

When μ is at least 30

$$X \sim Poisson(\mu) \Rightarrow X \sim N(\mu, \mu)$$

Same continuity corrections must be applied here

▼ Confidence interval for the mean of a normal distribution with known variance

Suppose X with a normal distribution $N(\mu, \sigma^2)$

The variance σ^2 is known, μ is the unknown parameter

Let $X_1, X_2, ..., X_n$ be the data collected from X, in a random sample size n

$\overline{X}$ is the sample mean

Sample mean provides an estimate of the population mean

However, different samples drawn from the same population will usually produce different estimates

Confidence interval for the parameter μ with confidence coefficient $(1 - \alpha)$

or A $(1 - \alpha)100$ percent confidence interval

$$(\mu^-, \mu^+) = \left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Generally, population standard deviation σ is not known

We will have to use sample standard deviation s to estimate

This gives rise to the approximate large-sample confidence intervals

▼ Z-value

e.g 95% confidence interval = $(1 - 0.05)100$ percent confidence interval

Where $\alpha = 0.05$

$z_{\alpha/2} = z_{0.05/2} = z_{0.025}$

From normal distribution table (z-value table),

Find z-value for probability of $(1 - 0.025) = 0.975$

Hence $z_{\alpha/2} = 1.96$

▼ Confidence interval for the mean of a population with unknown distribution

We can use central limit theorem (CLT) for large sample approximation

Replacing population variance by sample variance

$100(1 - \alpha)\%$ confidence interval for population mean:

$$(\mu^-, \mu^+) = \left(\overline{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

Where s = sample standard deviation

▼ Confidence interval for the mean of a normal distribution with unknown variance

Let $X_1, X_2, ..., X_n$ be a random sample from a normal distribution

With mean μ not known and unknown variance σ^2

To construct a $100(1-\alpha)\%$ confidence interval for parameter μ,

We must find $t_{\alpha/2}$ from t-table

Such that $P(T > t_{\alpha/2}) = \alpha/2$

Corresponding to $n-1$ degrees of freedom

Given random sample of size n, sample mean $\overline{x}$ and sample standard deviation s

From a normal distribution with mean μ

$100(1-\alpha)\%$ confidence interval for μ

$$(\mu^-, \mu^+) = \left(\overline{x} - t\frac{s}{\sqrt{n}}, \overline{x} + t\frac{s}{\sqrt{n}}\right)$$

Where t is the $(1-\alpha/2)$ quantile of t-distribution with $(n-1)$ degrees of freedom

▼ t-value

e.g 95% confidence interval = $100(1-0.05)\%$ confidence interval

Where $\alpha = 0.05$

$t_{\alpha/2} = t_{0.05/2} = t_{0.025}$

From t-table,

Find t-value for quantile of $(1-0.025) = 0.975$

Which is α column of 0.025

v is degrees of freedom

e.g. sample size $5-1=4$ degrees of freedom

Hence $t(4) = 2.776$

▼ CI for large sample size and unknown variance

Where degree of freedom for t is large (not in the table), can approximate results with the Z-test instead

i.e. replace t value with z value

▼ An approximate large-sample confidence interval for a proportion

Approximate $100(1-\alpha)\%$ confidence interval for a proportion p

Obtained by observing x successes in a sequence of n independent Bernoulli trials

Each with probability of success p

e.g. proportion/ percentage of people in a population that own a smartphone

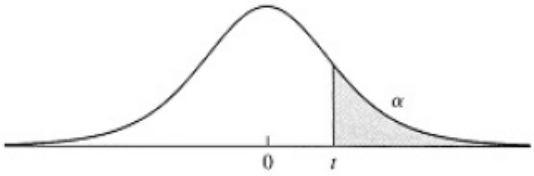
$$(p^-, p^+) = \left(\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Where $\hat{p} = x/n$ is the estimate of p

and z is the $(1-\alpha/2)$ quantile of the standard normal distribution

This confidence interval is valid when both np and $n(1-p)$ are least 5

Table 2.1 Upper percentage points for the Student's t distribution



	α								
ν	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	0.255	0.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291

▼ One-way ANOVA test

Total variability in the observations (Total sum of squares; SST)

is partitioned into 2 components:

- The variation among the treatment means (treatment sum of squares; SSTR)
- The variation among the experimental units within treatments (error sum of squares; SSE)

i.e.

- The variation (variance) among the means of the 4 labs
- The variation (variance) among the independent observations within a lab

Essentially,

SST = SSTR + SSE

Error of sum of squares (SSE) = SST - SSTR

Source	df	SS	MS	*F* statistic
Treatment	k -1	$SS_{treatment}$ ($SSTR$)	$MS_{treatment}$ ($MSTR$)	$F_0 = \frac{MS_{treatment}}{MS_{error}}$
Error	N - k	SS_{error} (SSE)	MS_{error} (MSE)	
Total	N -1	SS_{total} (SST)		

$$SSTR = SS_{treatment} = \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2$$

$$SSE = SS_{error} = \sum_{i=1}^{k}(n_i - 1)s_i^2$$

▼ Mean square for treatments

Found by dividing the sum of squares by the corresponding number of degrees of freedom

$$MSTR = \frac{SSTR}{k-1}$$

where k = number of treatments

▼ Mean square error

An unbiased estimator of the common population variance σ^2 within each of the k treatments

$$MSE = \frac{SSE}{N-k}$$

where $N = \sum_{i=1}^{k} n_i = kn$, if $n_i = n$ for all i

▼ F-statistic

We test the equality of the population means by comparing 2 variability components

- Within Groups Variability (SSE)

  Variability about the individual sample means within the k groups of observations

- Between Groups Variability (SSTR)

  Variability among the k group means

We are interested to test the null hypothesis that the treatment means are equal

Hence our hypotheses are

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

$$H_1 : \text{at least one } \mu_i \text{ is different from the others}$$

The F-statistic is the ratio of the mean squares for treatment to error:

$$F_o = \frac{MSTR}{MSE}$$

H_0 : No difference in treatment means

If $F_0 > F_{critical}$, the null hypothesis is rejected and we can conclude that there are differences in the treatment means at the chosen level of significance

▼ Explain the difference between type I and type II errors

▼ Type I error

Rejecting the null hypothesis when null hypothesis is true

Probability of type I error is denoted by α

▼ Type II error

Accepting the null hypothesis when null hypothesis is false

Probability of type II error is denoted by β

▼ Define the critical regions

This is the rejection region for H_0

Size of the critical regions = probability of type I error = α

Where α is the level of significance of the test

▼ Z Test

This is to test on the mean of a normal distribution, with known variance

i.e. mean μ of a single normal population where variance of population σ^2 is known

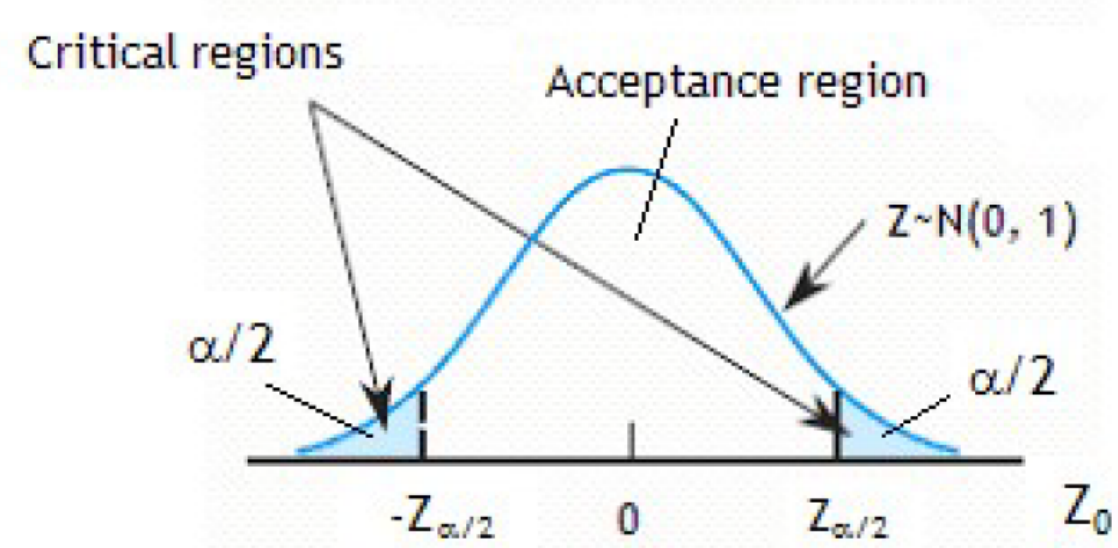
The test statistic is given by:

$$Z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

If the null hypothesis $H_0 : \mu = \mu_0$ is true

The probability that the test statistic Z_0 will fall between $-z_{a/2}$ and $+z_{a/2}$ is $1 - \alpha$

i.e. the probability of Z_0 falling in the regions $Z_0 < -z_{a/2}$ or $Z_0 > +z_{a/2}$ is $\alpha/2$



This implies that if null hypothesis is true, it will be unusual and rare to encounter a sample with an observed test statistic that falls in the tails of the Z distribution

The sample value of $\overline{X}$ is considerably different from μ_0

Hence H_0 is false and should be rejected

▼ One-tailed

$H_0: \mu \leq \mu_0$

$H_1: \mu > \mu_0$

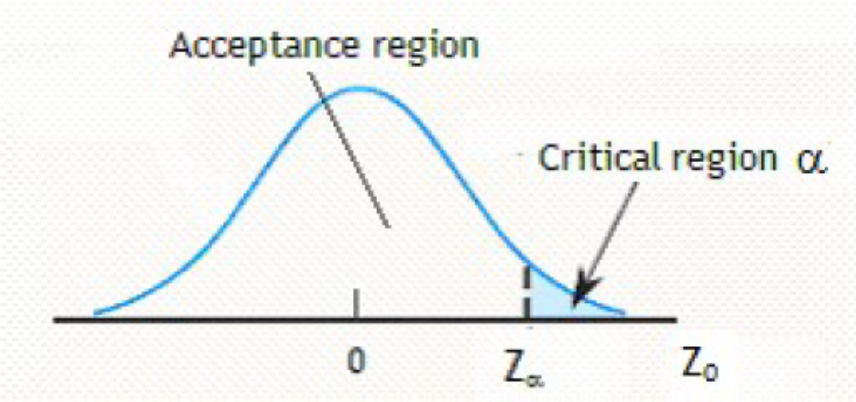


Figure 3.2 The critical region for the one-sided alternative hypothesis $H_1: \mu > \mu_0$ (Z test)

(Source: Page 374, Chapter 8 of textbook Ho, Xie & Goh, 2011)

In this case, if value of Z_0 exceeds $+z_\alpha$

Null hypothesis H_0 will be rejected

▼ Large sample testing of hypothesis about a population mean

Given a large sample size $n\ (> 30)$ from a population with mean μ

Central limit theorem (CLT) can provide an approximate test of the null hypothesis

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

▼ T Test

Test on the mean of a normal distribution, with unknown variance

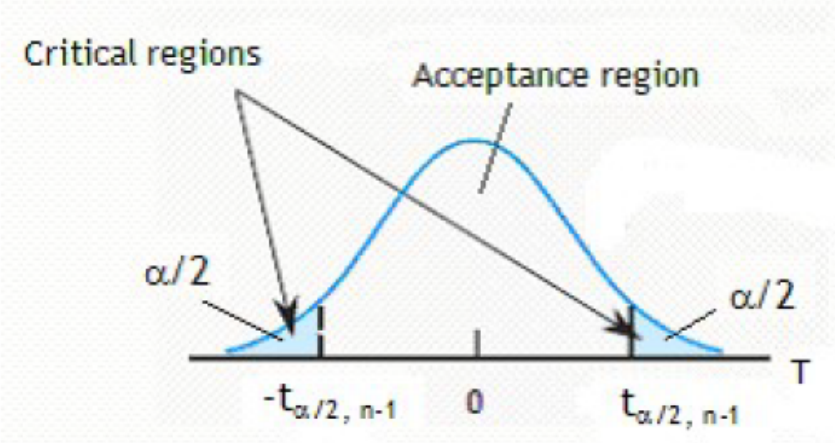
Given a sample of size n from a normal distribution with mean μ and unknown variance σ^2

The test statistic is:

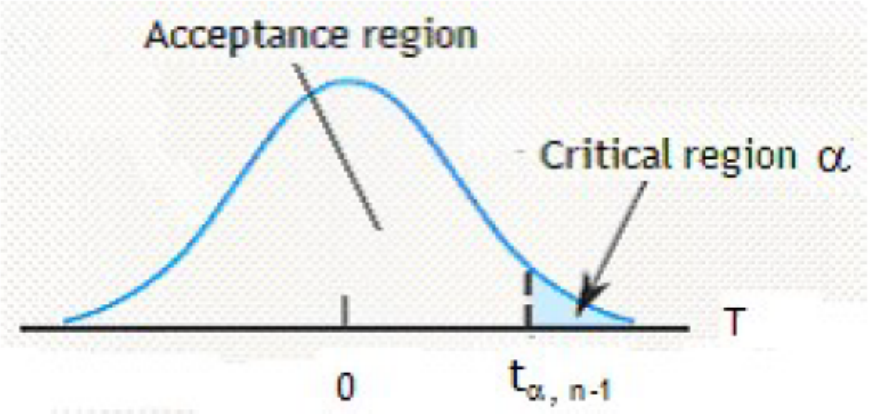
$$t_0 = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

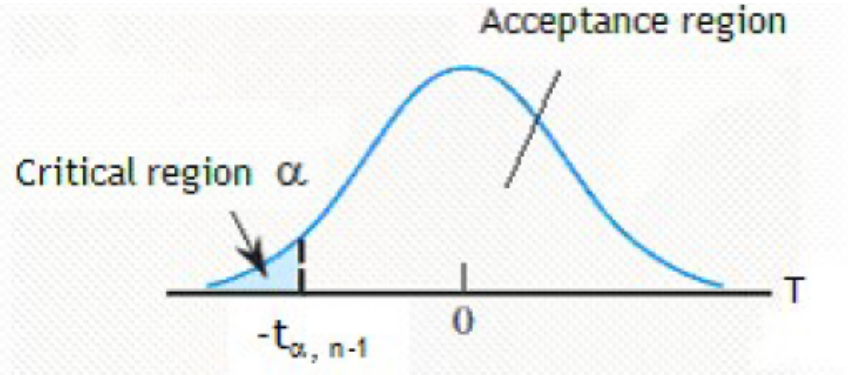
Where $\overline{X}$ is sample mean and s is the sample standard deviation

▼ Two tailed



▼ One tailed





▼ Using R

```
# mu argument gives value which you want to compare
# with the sample mean
t.test(dataset, mu=k)
```

By default, R performs a two-tailed test

For one-tailed, alternative argument must be set as "greater" or "less"

```
t.test(dataset, mu=k, alternative="less", conf.level=0.95)
```

Addn.: default 95% confidence interval for population mean is included with output

To adjust size of CI, use conf.level argument

▼ Example

```
dataset = c(171.6, 191.8, 178.3, 184.9, 189.1)
t.test(dataset, mu=185, alternative="less", conf.level=0.95)
```

▼ Construct confidence intervals in association with hypothesis testing

Null hypothesis $H_0 : \theta = \theta_0$ can be tested against two sided alternative hypothesis $H_1 : \theta \neq \theta$

Through using data to construct a confidence interval (θ^-, θ^+) for parameter θ

For a test at α significance level, a $100(1-\alpha)\%$ confidence interval should be used

If θ_0 is not inside the confidence interval, the null hypothesis can be rejected in favour of the alternative hypothesis

If θ_0 is inside the CI, then there is insufficient evidence to reject the null hypothesis

▼ Testing population proportions

Sometimes, instead of population means, we are interested in estimating the percentage (or proportion) of some group with a certain characteristics

$$(p^-, p^+) = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Where $\hat{p} = X/n$

▼ Z as test statistic

$$Z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}}$$

Where $\sigma_{\hat{p}}^2 = p(1-p)/n$

Hence,

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

If null hypothesis $H_0 : p = p_0$ is true,

Then X is approximately normal with mean np_0 and variance $np_0(1-p_0)$

Then,

$$Z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

If random variable X follows a binomial model with parameters n and p

Then approximate two-sided $(1-\alpha)100\%$ confidence interval for p is:

$$\hat{p} - z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

▼ Formulate the generic null and alternative hypotheses

Given two populations X_1 and X_2 with unknown means μ_1 and μ_2

We want to test the difference between the 2 means

Population variances are assumed to be known with values σ_1^2 and σ_2^2

Also assumed random variables X_1 and X_2 are normally distributed

If non-normal, assume conditions of central limit theorem applies

Sample size of n_1 drawn from X_1 with sample mean $\overline{X}_1$

Sample size of n_2 drawn from X_2 with sample mean $\overline{X}_2$

Assume both random samples are independent

and data within each sample are independently distributed with means μ_1 and μ_2

A two-sided hypothesis test on the difference in population means as follows:

$H_0 : \mu_1 - \mu_2 = \mu_0$

$H_1 : \mu_1 - \mu_2 \neq \mu_0$

Where μ_0 is a specified difference

Test procedure is based on the distribution of the difference in sample means $\overline{X}_1 - \overline{X}_2$

Since,

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

The appropriate test statistic is:

$$Z_0 = \frac{\overline{X}_1 - \overline{X}_2 - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The denominator, i.e.:

$$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Represents the standard error of the point estimate $\overline{X}_1 - \overline{X}_2$

If objective is to test equality of the 2 means, then $\mu_0 = 0$

▼ Possible scenarios for a hypothesis test

Alternative hypothesis	Rejection criteria
$H_1: \mu_1 - \mu_2 \neq \mu_0$	$Z_0 > Z_{\alpha/2}$ or $Z_0 < -Z_{\alpha/2}$
$H_1: \mu_1 - \mu_2 < \mu_0$	$Z_0 < -Z_\alpha$
$H_1: \mu_1 - \mu_2 > \mu_0$	$Z_0 > Z_\alpha$

▼ Confidence interval on the difference in means (variance known)

$100(1-\alpha)\%$ confidence interval on $\mu_1 - \mu_2$ can be constructed as follows:

$$(L^-, L^+) = \left(\overline{x}_1 - \overline{x}_2 - (z_{\alpha/2})\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, \overline{x}_1 - \overline{x}_2 + (z_{\alpha/2})\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

Where L^- and L^+ are lower and upper confidence limits

$z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the standard normal distribution

$-z_{\alpha/2}$ is the lower $100\alpha/2$ percentage point of the standard normal distribution

▼ Perform a two-sample t test

To test the difference in means of 2 normal distribution with unknown but equal variances

Useful in investigating whether the mean values of population are significantly different or not

Or whether the difference in means $\mu_1 - \mu_2$ is equal to a specified value μ_0

2 normal populations with mean values $\mu_1 \text{ and } \mu_2$

Both populations have unknown and equal variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Random samples of size n_1 drawn from pop 1 and size n_2 from pop 2

Let $\{\overline{X}_1, \overline{X}_2\}$ and $\{s_1^2, s_2^2\}$ be sample means and sample variances respectively

▼ Assumptions of the test

- That the variation in the 1st and 2nd population may be adequately modelled by a normal distribution with means μ_1 and μ_2 respectively and variance σ^2 (equal variance)
  - The adequacy of the normal model can be checked by graphical methods or statistical tests
  - Moderate departure from normality does not adversely affect the test procedure using t-statistics
- That the observations on the two populations are independent of one another.
- That the variance is the same in the two populations** (important)

▼ Hypothesis test that population variances are equal

It is very unlikely that sample variances s_1^1 and s_2^2 are equal

The questions is: How pronounced must the difference between the 2 sample variances be before the assumption of equal underlying variances is thrown into doubt

Hence, we can perform a hypothesis test that the population variances are equal

before carrying out t test for the equality of the means

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

▼ Weighted average of the two sample variances s_p^2

This is an estimate of the common variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

▼ Test procedure

Under $H_0 : \mu_1 - \mu_2 = \mu_0$

The following test statistic has t-distribution with $n_1 + n_2 - 2$ degrees of freedom

$$t_0 = \frac{\overline{X}_1 - \overline{X}_2 - \mu_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

When sample size is large, i.e. both n_1 and n_2 exceed 30

Normal distribution procedure based on conditions of central limit theorem can be used

▼ Decision making criteria

Alternative hypothesis	Rejection criteria
$H_1: \mu_1 - \mu_2 \neq \mu_0$	$t_0 > t_{\alpha/2,\ n_1+n_2-2}$ or $t_0 < -t_{\alpha/2,\ n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 < \mu_0$	$t_0 < -t_{\alpha,\ n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 > \mu_0$	$t_0 > t_{\alpha,\ n_1+n_2-2}$

▼ Two-sample t-interval

If n_1 and n_2 are the same sample sizes, and $\overline{X}_1$ and $\overline{X}_2$ are the sample means of two independent samples from normal distributions with means μ_1 and μ_2 and common variance

Then $100(1-\alpha)\%$ confidence interval for the difference between means ($\mu_1 - \mu_2$) is given by:

$$(d^-, d^+) = \left(\overline{X}_1 - \overline{X}_2 - (t_{\alpha/2})(s_p)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ \overline{X}_1 - \overline{X}_2 - (t_{\alpha/2})(s_p)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

Where $t_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of $t_{(n_1+n_2-2)}$

and s_p is the pooled estimate of the common standard deviation

▼ Perform a paired t test

To test for the mean difference in a matched pair of observations

e.g. observations before patient takes a pill vs observations after patient takes a pill

D_i is the difference in the i^{th} matched pair of observation (X_i, Y_i)

$D_i = X_i - Y_i$

Where,

X_i → the i^{th} observation before an entity is subjected to the treatment

Y_i → the i^{th} observation after the entity is subjected to the treatment

Investigate the null hypothesis that the mean difference of the matched pair of observation is 0 or any specified difference value

Assuming that the observed differences $d_i, i = 1, 2, ..., n$ are independent observations on a normal random variable

With unknown mean μ and unknown variance σ^2

Let $H_0 : \mu_d = \mu_0$

Then the following statistic follows a t-distribution with d.f. = $n-1$

$$\frac{\overline{d}-\mu_0}{s/\sqrt{n}}$$

One-tailed or two-tailed t test can be carried out depending on the alternative hypothesis

▼ Confidence interval for the mean difference d

Suppose that the differences are normally distributed with mean d

Given a random sample of n pairs of observations with differences $d_1, d_2, ..., d_n$

Sample mean d and sample standard deviation s

The $100(1-\alpha)\%$ confidence interval for the mean difference d is given by:

$$(d^-, d^+) = \left(\overline{d} - t_{\alpha/2}\frac{s_d}{\sqrt{n}}, \overline{d} + t_{\alpha/2}\frac{s_d}{\sqrt{n}}\right)$$

Where $t_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of $t(n-1)$

This confidence interval is exact (?)

▼ Explain the concept and interpretation of p-values

The p-value is a number (p)

Also called the significance probability (SP)

Definition:

Corresponding to an observed value of a test statistic, the SP is the lowest level of significance at which the null hypothesis could have been rejected.

This quantifies the extent to which the data cast doubt on the null hypothesis

The lower the SP, the more evidence the data provides against the null hypothesis

The higher the SP, the more the data supports the null hypothesis

▼ Procedure

1. Determine the null and alternative hypothesis.
2. Decide what data to collect.
3. Determine a suitable test statistic and its null distribution.
4. Collect data and calculate the observed value of the test statistic.
5. Identify all other values of the test statistic that are at least as extreme, in relation to the null hypothesis, as the value that was actually observed.

6. Calculate the significance probability, which is the probability, under the null hypothesis, of those values of the test statistic identified in Step 5.
7. Interpret the significance probability.
8. Report clearly the conclusion to be drawn from your test

▼ Rough interpretations

Significance Probability p	*Rough interpretation*
$p > 0.10$	little evidence against H_0
$0.10 \geq p > 0.05$	weak evidence against H_0
$0.05 \geq p > 0.01$	moderate evidence against H_0
$p \leq 0.01$	strong evidence against H_0

▼ Comparing Two Proportions

For observations made on 2 independent binomial random variables

$$X_1 \sim B(n_1, p_1) \qquad X_2 \sim B(n_2, p_2)$$

We can test for the equality of two proportions

Where,

$H_0 : p_1 = p_2$

The test statistic is:

$$D = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

For large values of n_1 and n_2,

An approximate null distribution of D is:

$$N\left(0, \hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Where

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Hence,

$$N\left(0, \frac{X_1 + X_2}{n_1 + n_2}\left(1 - \frac{X_1 + X_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

▼ Explain linear regression model

Statistical models that reflect the way in which variation in an observed variable (e.g. height) changes with one or more other variables (e.g. age) are called regression models.

Regression analysis is the process of determining how a variable Y is related to one, or more variables $X_1, X_2, X_3, ..., X_k$

There are several terminologies for

X : inputs / factors / regressors / predictors / explanatory variables

Y : output / response / dependent variable

▼ Usefulness of linear regression model

We can study the relationship between variables

and try to explain or predict the behaviour of the response variable in terms of the behaviour of one or more other explanatory variables

▼ Regression models

▼ General regression model

When the distribution of r.v. Y is related to the value taken by some associated variable X

Then the relationship can be represented by a general regression model:

$$Y_i = h(X_i) + W_i$$

Where $h()$ represents some function

and (W_i) s are independent r.v. with zero mean

▼ Regression Curve

$$Y = h(X)$$

▼ Linear Regression Model

Where Y depends linearly on x

$$Y_i = \alpha + \beta x_i + W_i$$

Random variables W_i are called residuals

They are independent with zero mean and constant variance σ^2

▼ Regression line

$$y = \alpha + \beta x$$

▼ Explain the method of least squares

Suppose a linear relation between variables X and Y in the form of:

$$Y_i = \alpha + \beta x + \omega$$

Parameters α and β are unknown and need to be estimated

The error term ω is assumed to be independent, normally distributed with mean 0 and variance σ^2

α is the y-intercept of the line

β is the gradient of the line

β is the amount of increase (or decrease) in the deterministic component of Y for every 1-unit increase in X.

The task is to select suitable values for the parameters α and β that will best fit the given set of n observations

▼ Method of least squares

Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be n-observations of the random variables (X, Y)

Least square estimate of the regression line is: $y = \hat{\alpha} + \hat{\beta} x$

Least square estimate $\hat{\beta}$ of the gradient parameter β is:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

Least square estimate $\hat{\alpha}$ of the constant term α is given by:

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

▼ Sum of Squares of deviations

$$S_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$S_{yy} = \sum (y_i - \overline{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

$$S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

Use equations from above to construct least squares line

$$y = \hat{\alpha} + \hat{\beta}x$$

▼ Compute Pearson correlation coefficient

A measure of strength of linear association between X and Y is given by the Pearson correlation coefficient, r

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

r is dimensionless

Sign of r indicates whether the relationship between 2 variables is positive or negative

Absolute value of r gives a measure of strength of linear association between the variables

Further r is from zero, the stronger the relationship

i.e. closer to +1 or -1

▼ R code

```
cor(x, y, method = "")
```

▼ Coefficient of determination

Another way to measure the usefulness of the model is to measure the contribution of X in predicting Y

Interpreted as "proportion of total sample variability explained by the linear relationship"

We can calculate how much the errors of prediction of Y were reduced by using the information provided by X

This is the square of Pearson correlation coefficient

$$r^2$$

▼ Interpret various strengths of correlation

Correlation does not equal causation

Although there might be a reasonably strong association between 2 variables

The data does not say anything about why they correlate to each other

There may be a multitude of explanations as to why

e.g.

- Changes in X cause changes in Y.

- Changes in Y cause changes in X.
- Changes in some third variable, Z, independently cause changes in X and Y.
- The observed relationship between X and Y is just coincidence, with no causal explanation at all.

A scatterplot cannot determine which of these explanations is valid

▼ Explain the difference between Pearson correlation coefficient and Spearman rank correlation coefficient

The spearman method replaces original data by their ranks

Then calculates the Pearson correlation coefficient for the ranks

The Pearson correlation coefficient is a measure of strength of linear association,

While the Spearman rank correlation coefficient is a measure of monotonic association (increasing or decreasing relationship)

Advantage of spearman's is that it requires only the rank of the data

▼ Analyse results of Wilcoxon signed rank test

Used to compare 2 probability distributions when a paired difference design is used

e.g. consumer preferences for two competing products are often compared by having each of a sample of consumers rate both products

Ratings are paired on each consumer

Formulated hypotheses are:

H_0 : The probability distributions of the ratings for products A and B are identical.

H_1 : The probability distributions of the ratings differ for the two products.

We replace the individual differences by ranks and test for the zero difference

1. Obtain a data vector $d_1, d_2, ..., d_n$ of differences with $0s$ deleted.
2. Order the absolute differences from least to greatest allocating i^{th} rank to the i^{th} absolute differences.
3. Introduce the sign difference and compute the statistic sum of positive ranks.
4. Obtain SP and state the conclusion.

▼ Detailed explanation

1. Obtain d_i which is all the differences $d_1, d_2, ..., d_n$ between x and y
2. Take the absolute values of all d_i
3. Rank the values from smallest to largest

i.e. Given $d_1 = 3, d_2 = -1, d_3 = 4, d_4 = -2$

Rank in ascending order is d_2, d_4, d_1, d_3

4. Sum up the ranks of the positive differences (W_+)

Positive differences here are d_1, d_3

With respective ranks of 3, 4

Hence, $W_+ = 7$

5. Determine the mean of W_+ with sample size n

$$E(W_+) = \frac{4(4+1)}{4} = 5$$

6. Is W_+ far from $E(W_+)$?

▼ Conclusion

Wilcoxon signed rank test rejects the null hypothesis that there are no systematic differences within pairs

WHEN the rank sum W_+ is far from its mean

Use p-value, critical regions etc.

Under the null hypothesis of zero median difference

For sample size of n (excluding any zero differences)

▼ Random variable W_+

Observed value: Wilcoxon test statistic ω_+

▼ Mean (Expectation)

$$E(W_+) = \frac{n(n+1)}{4}$$

▼ Variance

$$V(W_+) = \frac{n(n+1)(2n+1)}{24}$$

Distribution of

$$Z = \frac{W_+ - E(W_+)}{SD(W_+)}$$

is approximately standard normal

**The normal approximation is generally adequate as long as the sample size n is at least 16

▼ Example 6.1

Example 6.1

A study of early childhood education asked kindergarten students to retell two fairy tales that had been read to them earlier in the week. Each child told two stories. The first had been read to them, and the second had been read but also illustrated with pictures. An expert listened to a recording of the children and assigned a score for certain uses of language. Here are the data for five "low-progress" readers in a pilot study:[9]

Child	1	2	3	4	5
Story 2	0.77	0.49	0.66	0.28	0.38
Story 1	0.40	0.72	0.00	0.36	0.55
Difference	0.37	−0.23	0.66	−0.08	−0.17

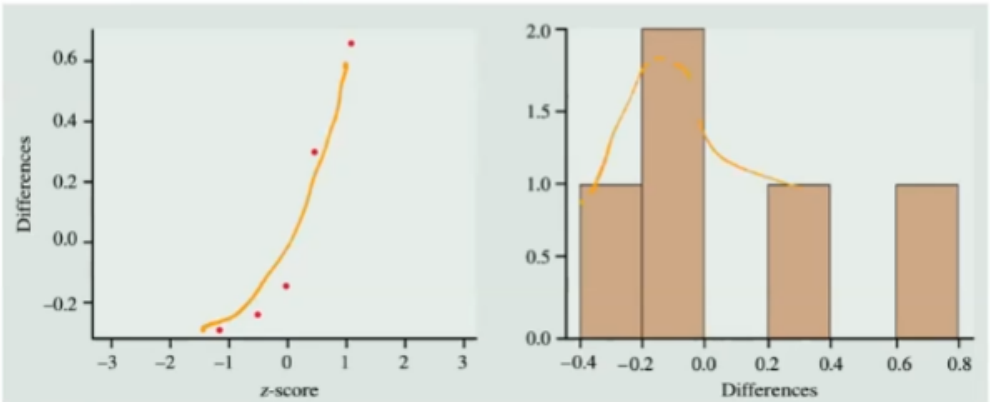
We wonder if illustrations improve how the children retell a story. We would like to test the hypotheses

H_0: scores have the same distribution for both stories
H_a: scores are systematically higher for Story 2

Dr Larry Gui

Since we cannot be sure that the scores are normally distributed, we use non-parametric test instead

Because this is a matched pairs design, we base our inference on the differences. The matched pairs t test gives $t = 0.635$ with one-sided P-value $P = 0.280$. Displays of the data suggest some lack of normality. We would therefore like to use a rank test.



Rank: $-d_4, -d_5, -d_2, +d_1, +d_3$

Sum of positive ranks (SP)(W_+) = 4 + 5 = 9

$$\mu_{W^+} = \frac{5(5+1)}{4} = 7.5$$

$$\sigma_{W^+} = \sqrt{\frac{5(5+1)(2(5)+1)}{24}} = \sqrt{\frac{330}{24}} = 3.708$$

$$W^+ \approx W_N^+ \sim N(7.5, 3.708^2) \text{ approx.}$$

Since W_+ is discrete, we need to use continuity corrections to approximate its normal distribution

$P(W_+ \geq 9) \approx P(W_+ \geq 8.5)$

We treat $W_+ \geq 9$ as occupying the interval from 8.5 to 9.5

Lastly, standardise the equation:

$$P(W_+ \geq 8.5) = P\left(\frac{W_+ - 7.5}{3.708} \geq \frac{8.5 - 7.5}{3.708}\right)$$

$$P(Z \geq 0.27) = 0.39358 \approx 0.3$$

Since probability is not small,

$W_+ = 9$ is not extremely large compared to $\mu_{W^+} = 7.5$

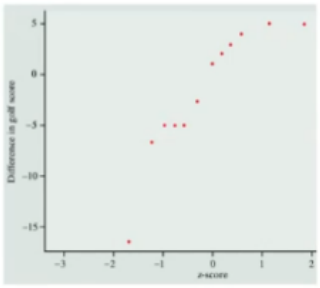
Hence, we do not reject H_0 and conclude that there is insufficient evidence to illustrate an improvement in the child's ability to retell the story.

▼ Example 6.2 (practice)

Example 6.2

Here are the golf scores of 12 members of a college women's golf team in two rounds of tournament play. (A golf score is the number of strokes required to complete the course, so that low scores are better.)

Player	1	2	3	4	5	6	7	8	9	10	11	12
Round 2	94	85	89	89	81	76	107	89	87	91	88	80
Round 1	89	90	87	95	86	81	102	105	83	88	91	79
Difference	5	−5	2	−6	−5	−5	5	−16	4	3	−3	1



Negative differences indicate better (lower) scores on the second round. We see that 6 of the 12 golfers improved their scores. We would like to test the hypotheses that in a large population of collegiate woman golfers

H_0: scores have the same distribution in rounds 1 and 2
H_a: scores are systematically lower or higher in round 2

A normal quantile plot of the differences on the right shows some irregularity and a low outlier. We will use the Wilcoxon signed rank test.

▼ Analyse results of Mann-Whitney test

This test is an alternative to testing two independent groups of data when the two-sample t test may not be applicable because of lack of normality.

This test can be used to test the null hypothesis that the distributions of the populations from which two independent samples (A and B) were drawn are identical.

Test statistic is U_A , the sum of the ranks for sample A

$$U_A \approx N\left(\frac{n_A(n_A+n_B+1)}{2}, \frac{n_A n_B(n_A+n_B+1)}{12}\right)$$

Where n_A and n_B are the respective sample sizes

Approximation is adequate if both sample sizes are at least 8 - 10

▼ Steps

1. Pool the two samples and then sort the combined data in ascending order.
2. Allocate a rank to each data value, the smallest being given rank 1. As usual, if two or more data values are equal, allocate the average of ranks to each.

   e.g. rank 10 and 11 = 10.5

	0.6	1.6	1.9	2.1	2.2	2.5	3.1	3.3	3.7	4.0
Sample	B	B	B	A	B	B	B	A	A	A
Rank	1	2	3	4	5	6	7	8	9	10.5

	4.0	4.1	4.8	5.4	5.4	6.1	6.2	6.3
Sample	B	B	A	A	B	A	B	A
Rank	10.5	12	13	14.5	14.5	16	17	18

3. Add up the ranks for each sample. Let:

   U_A = sum of ranks for sample A

   U_B = sum of ranks for sample B

   $$U_A + U_B = \frac{1}{2}(n_A+n_B)(n_A+n_B+1)$$

4. Very small or very large U_A imply the rejection of the null hypothesis

   They suggest that the A-values are "too frequently" smaller than or larger than B-values

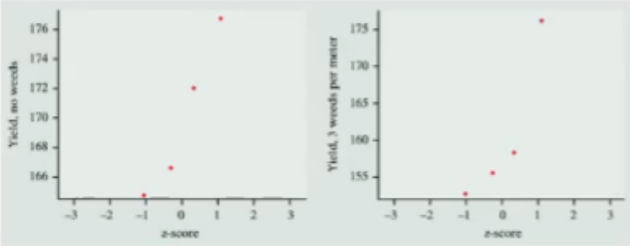
   Observed value of U_A may be compared with the null distribution of U_A to get the significance probability (SP) for the test

▼ Example 6.3

Example 6.3

Does the presence of small numbers of weeds reduce the yield of corn? Lamb's-quarter is a common weed in corn fields. A researcher planted corn at the same rate in 8 small plots of ground, then weeded the corn rows by hand to allow no weeds in 4 randomly selected plots and exactly 3 lamb's-quarter plants per meter of row in the other 4 plots. Here are the yields of corn (bushels per acre) in each of the plots.[1]

Weeds per meter	Yield (bu/acre)			
0	166.7	172.2	165.0	176.9
3	158.6	176.4	153.1	156.0



Normal quantile plots suggest departure from normality. Samples too small to assess adequacy or robustness of 2-sample t-test. We use a non-parametric test.

Dr Larry

$n_1 = 4, n_2 = 4, N = 8$

1. Rank all 8 observations

Rank all 8 observations:

Yield	153.1	156.0	158.6	165.0	166.7	172.2	176.4	176.9
Rank	1	2	3	4	5	6	7	8

If the presence of weeds reduces corn yields, we expect the ranks of the yields from plots with weeds to be smaller as a group than the ranks from plots without weeds. We might compare the *sums* of the ranks from the two treatments:

Treatment	Sum of ranks
No weeds	23
Weeds	13

Example 6.3/..

Treatment	Sum of ranks
No weeds	23
Weeds	13

We want to test H_0: no difference in distribution of yields

against the one-sided alternative

H_a: yields are systematically higher in weed-free plots

The sum of ranks for the weed-free plots:

$$\mu_W = \frac{n_1(N+1)}{2} = \frac{(4)(9)}{2} = 18 \qquad \sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}} = \sqrt{\frac{(4)(4)(9)}{12}} = \sqrt{12} = 3.464$$

The continuity correction (page 379) acts as if the whole number 23 occupies the entire interval from 22.5 to 23.5. We calculate the P-value $P(W \geq 23)$ as $P(W \geq 22.5)$ because the value 23 is included in the range whose probability we want. Here is the calculation:

$$\begin{aligned} P(W \geq 22.5) &= P\left(\frac{W - \mu_W}{\sigma_W} \geq \frac{22.5 - 18}{3.464}\right) \\ &= P(Z \geq 1.30) \\ &= 0.0968 \end{aligned}$$

The continuity correction gives a result closer to the exact value $P = 0.10$.

Since p-value = 0.0968, we reject H_0 under 10% significance level

And conclude that we have sufficient evidence to suggest weeds do reduce the yield of corn.

▼ R-code

```
X <- c(8.2, 9.4, 9.6, 9.7, 10.0, 14.5, 15.2, 16.1, 17.6, 19.4)
Y <- c(4.7, 4.9, 5.8, 6.4, 7.0, 7.3, 10.1, 11.2, 11.3, 13.2)
wilcox.test(X, Y)
```

▼ Explain the chi-squared distribution

Continuous random variable W is the sum of r independent squared observations on the standard normal random variable Z

$$W = Z_1^2 + Z_2^2 + ... + Z_r^2$$

W has a chi-squared χ^2 distribution with r degrees of freedom

$$W \sim \chi^2(r)$$

Mean:

$$\mu_W = r$$

Variance:

$$\sigma_W^2 = 2r$$

The chi-squared goodness-of-fit test is applied to situations in which we want to determine whether

A set of data may be looked upon as a random sample from a population having a given distribution

In a random sample of size n observations,

Each observation can be classified into one of k distinct classes

O_i = number of observations falling into class i

θ_i = probability that an observation falls into class i

$n\theta_i = E_i$ = expected number of observations falling into class i

For large values of n, the distribution of the quantity

$$\chi^2 = \sum_{1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k-1)$$

is chi-squared distribution with $k-1$ degrees of freedom

Essentially, sum up the squared differences between observed and expected frequency divided by expected frequency

▼ Example

Using the following data set, determine whether the number of errors a compositor makes in setting a galley of a type is random and follows a Poisson distribution.

Errors:	0	1	2	3	4	5	6	7	8	9
Frequency:	18	53	103	107	82	46	18	10	2	1

Solution:

We shall estimate the population parameter mean of the Poisson distribution

$$\mu = \frac{\sum_{1}^{n} f_i x_i}{\sum_{1}^{n} f_i} = \frac{1341}{440} = 3.05 \cong 3$$

H_0: The data come from the Poisson distribution with $\mu = 3$

H_1: The data do not come from the Poisson distribution with $\mu = 3$

No. of Errors	Observed freq.	P (X=x)	Expected freq.
0	18	0.0498	21.9
1	53	0.1494	65.7
2	103	0.2240	98.6
3	107	0.2240	98.6
4	82	0.1680	73.9
5	46	0.1008	44.4
6	18	0.0504	22.2
7	10	0.0216	9.51
8	2	0.0081	3.56
9	1	0.0027	1.19

$$\chi^2=\frac{(18-21.9)^2}{21.9}+\frac{(53-65.7)^2}{65.7}+...+\frac{(13-14.26)^2}{14.26}=5.91$$

Using probability, multiply by 440 (the total number of observations) to get expected frequency

Chi-squared approximation is adequate if the expected frequency of each class under the null hypothesis is 5.

If not so, the class division must be redefined by combining adjacent classes so that the expected frequency for each class becomes 5.

If need to combine classes (separate example):

7	10	0.0216	3.6
8	2	0.0081	1.7
9	1	0.0027	1.18

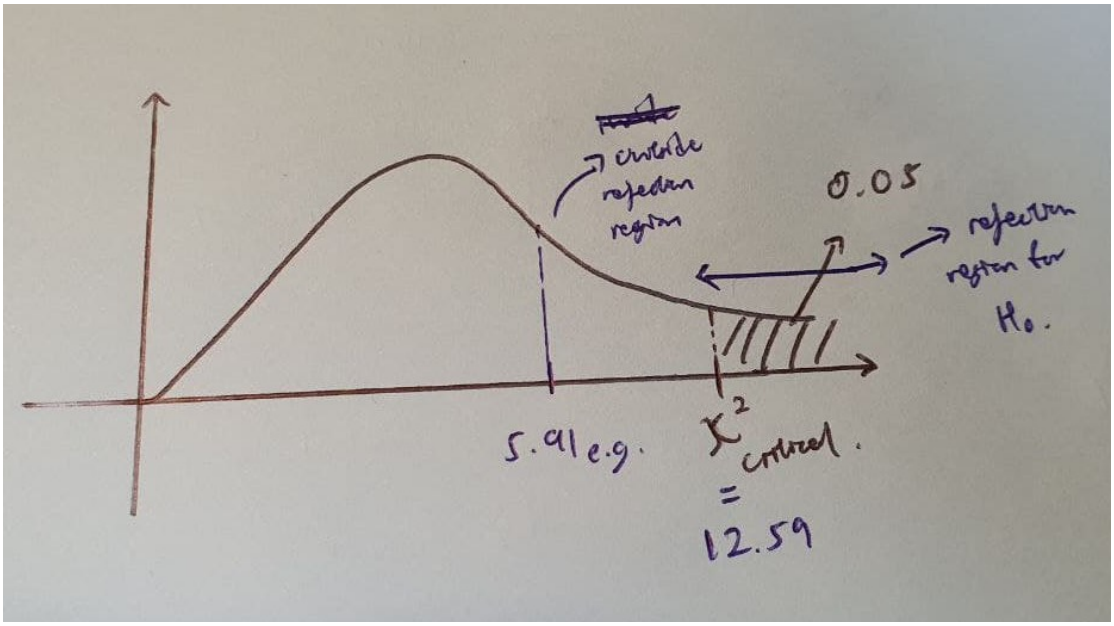
$$\chi^2=\frac{(18-21.9)^2}{21.9}+\frac{(53-65.7)^2}{65.7}+...+\frac{(13-6.5)^2}{6.5}=6.83$$

Conclusion:

For $\alpha=0.05$, $\chi^2_{0.05,6}=12.59>5.91$

Hence, the null hypothesis cannot be rejected.

$$\chi^2_{\text{test}}=5.91<\chi^2_{\text{critical}}=12.59$$



Hence, the data does come from a Poisson distribution with $\mu = 3$

If model involves p parameters from the data set,

then the test statistic follows a distribution with $df = k - p - 1$

▼ Explanation

Example

A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 100 times, with the following observed counts:

Number of Sixes	Number of Rolls
0	48
1	35
2	15
3	3

The casino becomes suspicious of the gambler and wishes to determine whether the dice are fair. What do they conclude?

If a die is fair, we would expect the probability of rolling a 6 on any given toss to be 1/6. Assuming the 3 dice are independent (the roll of one die should not affect the roll of the others), we might assume that the number of sixes in three rolls is distributed Binomial(3,1/6). To determine whether the gambler's dice are fair, we may compare his results with the results expected under this distribution. The expected values for 0, 1, 2, and 3 sixes under the Binomial(3,1/6) distribution are the following:

Often, the null hypothesis involves fitting a model with parameters estimated from the observed data.

In the above gambling example, for instance, we might wish to fit a binomial model to evaluate the probability of rolling a six with the gambler's loaded dice.

We know that this probability is not equal to 1/6, so we might estimate this value by calculating the probability from the data.

By estimating a parameter, we lose a degree of freedom in the chi-square test statistic. In general, if we estimate *d* parameters under the null hypothesis with *k* possible counts the degrees of freedom for the associated chi-square distribution will be *k - 1 - d*.

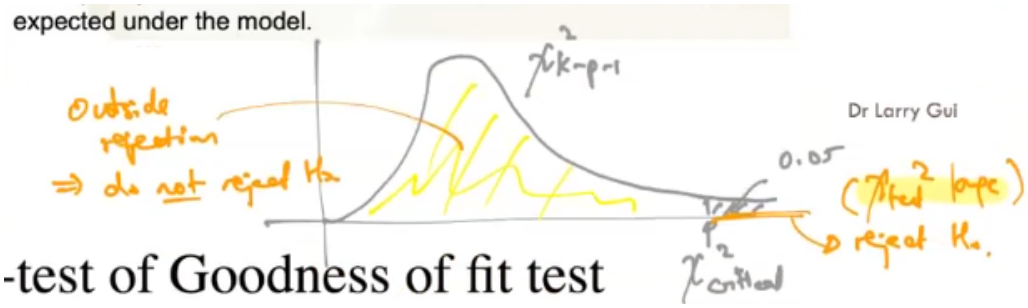
The test statistic is:

$$\chi^2_{\text{test}} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- The categories must be chosen in such a way that the expected frequency for each category is at least 5. Under the null hypothesis that the data arise from the hypothesized model, the distribution of the test statistic x^2 is approximately chi-squared with k-p-1 degrees of freedom, where p is the number of parameters whose values were estimated from the data.
- The significance probability is given by the upper tail probability of x^2 (k-p-1) for values exceeding the observed test statistic. The assessment of goodness of fit is based on quantifying the discrepancy between the data observed and the values that are expected under the model.

Chi-squared approximation is adequate if the expected frequency of each class under the null hypothesis is ≥ 5

If not so, the class division must be redefined by combining adjacent classes so that the expected frequency for each class becomes ≥ 5



χ^2 Distribution

Degree of Freedom	Probability of Exceeding the Critical Value								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38
	Not Significant							***Significant***	

▼ Use chi-squared statistics to perform goodness-of-fit test

▼ Example 6.5

Example 6.5

A department store, A, has four competitors: B,C,D, and E. Store A hires a consultant to determine if the percentage of shoppers who prefer each of the five stores is the same. A survey of 1100 randomly selected shoppers is conducted, and the results about which one of the stores shoppers prefer are below. Is there enough evidence using a significance level $\alpha = 0.05$ to conclude that the proportions are really the same?

Store	A	B	C	D	E
Number of Shoppers	262	234	204	190	210

uniform distn.??

H_0: uniform (discrete) distribution fits data

H_1: fit is not good.

$\longrightarrow$ χ^2-test for goodness-of-fit.

Dr Larry Gui

(iii) $\alpha = 0.05$.

(iv) The *degrees of freedom*: $k - 1 = 5 - 1 = 4$. k-p-1

(v) The *test statistic* can be calculated using a table:

Preference	**% of Shoppers**	E	O	$O - E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
A	20%	$0.2 \times 1100 = 220$	262	42	1764	8.018
B	20%	$0.2 \times 1100 = 220$	234	14	196	0.891
C	20%	$0.2 \times 1100 = 220$	204	-16	256	1.163
D	20%	$0.2 \times 1100 = 220$	190	-30	900	4.091
E	20%	$0.2 \times 1100 = 220$	210	-10	100	0.455

(iii) $\alpha = 0.05$.

(iv) The *degrees of freedom*: $k - 1 = 5 - 1 = 4$. k-p-1 = 5-0-1 = 4

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum \frac{(O-E)^2}{E} = 14.618.$$

(vi) From $\alpha = 0.05$ and $k - 1 = 4$, the *critical value* is 9.488.

(vii) Is there enough evidence to reject H_0? Since $\chi^2 \approx 14.618 > 9.488$, there is enough statistical evidence to reject the null hypothesis and to believe that customers do not prefer each of the five stores equally.

▼ Explain the use of chi-squared statistics to test for association between variables

Chi-squared test for data in contingency tables:

Suppose that we are interested in testing two attributes represented by rxc contingency table for association between them.

A1 / A2	1	2...	j ...	... c	Sum
1	f_{11}	f_{12}	f_{1j}	f_{1c}	$f_{1.}$
2					
i.	f_{i1}	f_{i2}	f_{ij}	f_{ic}	$f_{i.}$
.					
r.	f_{r1}	f_{r2}	f_{rj}	f_{rc}	$f_{r.}$
Sum	$f_{.1}$	$f_{.2}$	$f_{.j}$	$f_{.c}$	.N

The expected frequency for cell(i, j) is obtained by

Multiplying the total of the row to which the cell belongs by the total of the column to which the cell belongs

And then dividing by the grand total.

Expected frequency for cell (i, j)

$$e_{ij} = \hat{p}_{ij} N = \hat{p}_{i.} \hat{p}_{.j} N$$
$$= \frac{f_{i.}}{N} \frac{f_{.j}}{N} N = \frac{f_{i.} f_{.j}}{N}$$

Essentially $e_{ij} = \frac{f_i f_j}{N}$

Sum of row x sum of column divide by total

The test statistic is:

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

With $df = (r-1)(c-1)$

▼ Example

		"Ability in Maths"		
		Low	Average	High
"Interest in Statistics"	Low	63	42	15
	Average	58	61	31
	High	14	47	29

Solution:

H_0: ability in mathematics and statistics are independent

H_1: ability in mathematics and statistics are not independent

The expected frequencies for: row 1 = 45, 50, 25

row 2 = 56.25, 62.5, 31.25

row 3 = 33.75, 37.5, 18.75

$$\chi^2 = \frac{(63-45)^2}{45} + \frac{(42-50)^2}{50} + \ldots\ldots + \frac{(29-18.75)^2}{18.75}$$
$$= 32.14 > \chi^2_{0.01} = 13.277$$

R1C1:

$$\frac{(63+58+14)(63+42+15)}{360} = 45$$

df :

$$(3-1)(3-1) = 4$$