AMS 572 Review

Solha Park

Contents

1. What is Inferential Statistic?

Finals

- Chapters 7, 8, 9, 10, 12, 14

Inferential Statistics

Key concepts:

1. Population and Sample
2. Hypothesis Testing
3. Confidence Intervals
4. Regression Analysis
5. ANOVA
6. ...

Table of Contents Titles

CH7: Inferences for Single Samples

CH8: Inferences for Two Samples

CH9: Inferences for Proportions and Count Data

CONTENTS

7.1 Inferences on Mean (Large Samples)

7.1 Inferences on Mean (Large Samples)

- To estimate by a confidence interval (CI) or to test a hypothesis on the **unknown mean $\boldsymbol{\mu}$** of a population using a **random sample** $X_1, \ldots, X_n$ from that population

- For a large sample size n, the **CLT** tells us that $\bar{X}$ is approximately $N(\mu, \sigma^2/n)$ distributed, even if the population n is not normal.

- As long as the sample size is large enough (say ≥ 30) the following methods can be applied *even if* the sample comes from a nonnormal population with unknown variance.

- Use ***z*-test**

7.1 Inferences on Mean (Large Samples)

7.1.1 Large Sample Confidence Intervals on Mean

Pivotal random variable

(a function of the sample and parameter of interest whose probability distribution does not depend on the unknown parameters)

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Two-sided 100(1-α)% CI for μ:

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

(The following probability statement leads to the CI for μ)

$$P\left[-z_{\frac{\alpha}{2}} \leq Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

7.1 Inferences on Mean (Large Samples)

7.1.1 Large Sample Confidence Intervals on Mean

Sample Size Determination for a *z*-Interval

Margin of error

$$E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

and solve for ***n***,

$$n = \left[\frac{z_{\frac{\alpha}{2}} \sigma}{E}\right]^2$$

7.1 Inferences on Mean (Large Samples)

7.1.2 Hypothesis Tests on Mean (Large Sample)

$$H_0: \mu = \mu_0 \;\; vs \; H_1: \; \mu \neq \mu_0$$

When H_0 is true,

$$E(\bar{X}) = \mu_0, Var(\bar{X}) = \frac{\sigma^2}{n}$$

The test statistic

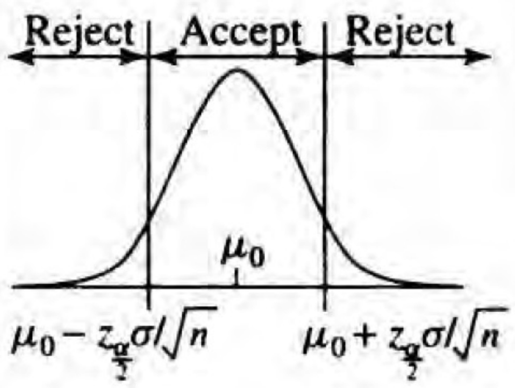
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Reject H_0 if

$$|z| > z_{\alpha/2}$$

equivalently,

$$|\bar{x} - \mu_0| > z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

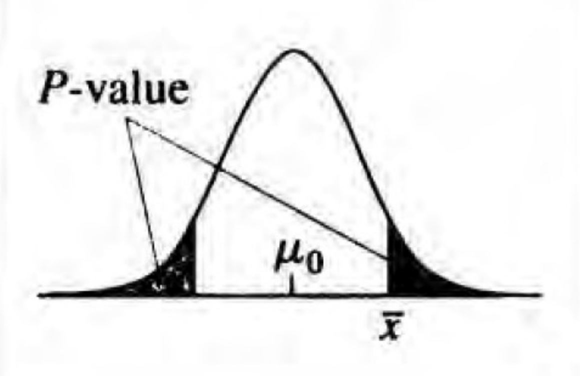


7.1 Inferences on Mean (Large Samples)

7.1.2 Hypothesis Tests on Mean (Large Sample)

P-value : a probability of observing more extreme or equally extreme test statistic values than observed test statistic values under the null hypothesis

$$P(|Z| \geq |z|\,|H_0) = 2\big(1 - \Phi(z)\big)$$



7.1 Inferences on Mean (Large Samples)

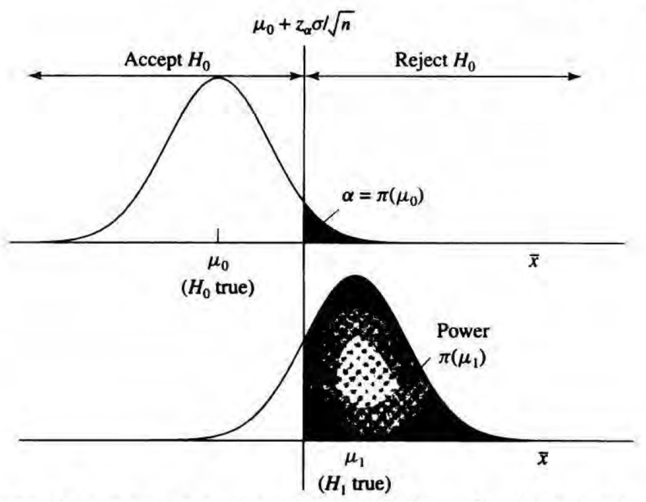
7.1.2 Hypothesis Tests on Mean (Large Sample)

Power Calculation for two-sided z-tests

$$\pi(\mu) = P(\text{Test rejects } H_0 \mid \mu)$$

Consider the problem of testing

$$H_0: \mu = \mu_0 \;\; vs \; H_1: \mu \neq \mu_0$$



$$\begin{aligned}
\pi(\mu) &= P\left(\bar{X} < \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \,\middle|\, \mu\right) + P\left(\bar{X} > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \,\middle|\, \mu\right) \\
&= P\left(Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < -z_{\alpha/2} + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right) + P\left(Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \; z_{\alpha/2} + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right). \\
&= \Phi\left[-z_{\alpha/2} + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right] + 1 - \Phi\left[z_{\frac{\alpha}{2}} + \frac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}}\right] \\
&= \Phi\left[-z_{\alpha/2} + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right] + \Phi\left[z_{\alpha/2} + \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right]
\end{aligned}$$

7.1 Inferences on Mean (Large Samples)

7.1.2 Hypothesis Tests on Mean (Large Sample)

Sample Size Determination for a two-Sided *z*-test

The treatment effect

$$\delta = \mu - \mu_0$$
$$\Leftrightarrow \mu = \mu_0 + \delta$$

Consider power function

$$\pi(\mu) = \Phi\left[-z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right] + \Phi\left[z_{\alpha/2} + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right]$$

Put $\mu_0 + \delta$ or $\mu_0 - \delta$ instead of μ

$$\pi(\mu_0 + \delta) = \pi(\mu_0 - \delta)$$
$$= \Phi\left[-z_{\alpha/2} - \frac{\delta}{\sigma/\sqrt{n}}\right] + \Phi\left[-z_{\alpha/2} + \frac{\delta}{\sigma/\sqrt{n}}\right] = 1 - \beta$$

A simple approximation can be obtained because for $\delta > 0$, $\Phi\left[-z_{\alpha/2} - \frac{\delta}{\sigma/\sqrt{n}}\right]$ is negligible

Using the fact that $\Phi[z_\beta] \cong 1 - \beta$,

$$-z_\alpha + \frac{\delta}{\sigma/\sqrt{n}} \cong z_\beta$$

Solve for n,

$$n = \left[\frac{(z_{\alpha/2} + z_\beta)\sigma}{\delta}\right]^2$$

7.2 Inferences on Mean (Small Samples)

7.2.1 Confidence Intervals on Mean (Small Sample)

Pivotal random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Two-sided 100(1-α)% CI for μ:

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

(The following probability statement leads to the CI for μ)

$$P\left[-t_{n-1,\alpha/2} \le T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \le t_{n-1,\alpha/2}\right] = 1 - \alpha$$

7.2 Inferences on Mean (Small Samples)

7.2.2 Hypothesis Tests on Mean (Small Sample)

$$H_0: \mu = \mu_0 \;\; vs \; H_1: \; \mu \neq \mu_0$$

When H_0 is true,

$$E(\bar{X}) = \mu_0, Var(\bar{X}) = \frac{s^2}{n}$$

The test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Reject $\boldsymbol{H_0}$ if

$$|t| > t_{n-1,\alpha/2}$$

equivalently,

$$|\bar{x} - \mu_0| > t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

7.2 Inferences on Mean (Small Samples)

7.2.2 Hypothesis Tests on Mean (Small Sample)

P-value

$$P(|T_{n-1}| \geq |t| \mid H_0) \text{ or } 2P(T_{n-1} \geq |t| \mid H_0)$$

<u>Power Calculation for two-sided *z*-tests</u>

$$\begin{aligned} \pi(\mu) &= P(\text{Test rejects } H_0 \mid \mu) \\ &= P\left(\left| \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \right| > |\, t_{n-1,\alpha/2} \right| \mu) \end{aligned}$$

where μ is the true mean

7.3 Inferences on Variance

7.3.1 Confidence Intervals on Variance

Pivotal random variable

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Two-sided 100(1-α)% CI for σ^2:

$$\frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}$$

(The following probability statement leads to the CI for σ^2)

$$P\left[\chi^2_{n-1,1-\frac{\alpha}{2}} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{n-1,\frac{\alpha}{2}}\right] = 1-\alpha$$

7.3 Inferences on Variance

7.3.2 Hypothesis Tests on Variance

$$H_0: \sigma^2 = \sigma_0^2 \;\; vs \; H_1: \; \sigma^2 \neq \sigma_0^2$$

When H_0 is true,
The test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Reject $\boldsymbol{H_0}$ if

$$\chi^2 > {\chi^2}_{n-1,\alpha/2} \quad \text{or} \quad \chi^2 < {\chi^2}_{n-1,1-\alpha/2}$$

P-value

$$P(\chi_{n-1}^2 \geq \chi^2 \,|H_0)$$

HW Problem

In order to test the accuracy of speedometers purchased from a subcontractor, the purchasing department of an automaker orders a test of a sample of speedometers at a controlled speed of 55 mph. At this speed, it is estimated that the readings will range ±2 mph around the mean.

(a) Set up the hypotheses to detect if the speedometers have any bias.

(b) How many speedometers need to be tested to have a 95% power to detect a bias of 0.5 mph or greater using a 0.01-level test? Use the rough estimate of σ obtained from the range.

(c) A sample of the size determined in (b) has a mean of $\bar{x} = 55.2$ and $s = 0.8$. Can you conclude that the speedometers have a bias?

(d) Calculate the power of the test if 50 speedometers are tested and the actual bias is 0.5 mph. Assume $\sigma = 0.8$.

Contents

8.1 Independent Samples and Matched Pairs Designs

8.1 Independent Samples and Matched Pairs Designs

- Independent sample design

Sample 1:	$x_1, x_2, \ldots, x_{n_1}$
Sample 2:	$y_1, y_2, \ldots, y_{n_2}$.

- Matched pairs design

Pair:	1	2	...	n
Sample 1:	x_1	x_2	...	x_n
Sample 2:	y_1	y_2	...	y_n

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(1) Inferences for Large Samples

- Suppose that the observations $x_1, x_2, \dots, x_{n_1}$ and $y_1, y_2, \dots, y_{n_2}$ are random samples from two populations wi th means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively.

- The goal is to compare μ_1 and μ_2 in terms of their differences $\boldsymbol{\mu_1 - \mu_2}$.

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$$

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(1) Inferences for Large Samples

The standardized random variable

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

By the central limit theorem (CLT).

For large samples, σ_1^2 and σ_2^2 can be replaced by s_1^2 and s_2^2

Two-sided 100(1-α)% CI for $\mu_1 - \mu_2$:

$$\bar{x} - \bar{y} - z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x} - \bar{y} + z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(1) Inferences for Large Samples

Testing the hypothesis

$$H_0: \mu_1 - \mu_2 = \delta_0 \;\; vs \; H_1: \; \mu_1 - \mu_2 \neq \delta_0$$

where $\delta_0 = \mu_1 - \mu_2$ under H_0.

Typically $\delta_0 = 0$ is used, which corresponds to testing

$$H_0: \mu_1 = \mu_2 \;\; vs \; H_1: \; \mu_1 \neq \mu_2$$

The test statistic

$$z = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(1) Inferences for Large Samples

Reject H_0 if

$$|z| > z_{\alpha}$$

or equivalently if

$$|\bar{x} - \bar{y} - \delta_0| > z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

P-value

$$P\big(|Z| \geq |z|\big|H_0\big) = 2P(Z \geq |z|)$$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(i) Case 1: $\sigma_1^2 = \sigma_2^2$

(ii) Case 2: $\sigma_1^2 \neq \sigma_2^2$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(i) Case 1: $\sigma_1^2 = \sigma_2^2$

- Denote the common value of σ_1^2 and σ_2^2 by σ^2, which is unknown.
- An unbiased estimator of this parameter is the sample mean difference $\bar{X} - \bar{Y}$.
- The **sample variances** from the two samples,

$$S_1^2 = \frac{\sum(X_i - \bar{X})^2}{n_1 - 1} \quad \text{and} \quad S_2^2 = \frac{\sum(Y_i - \bar{Y})^2}{n_2 - 1}$$

are both unbiased estimators of σ^2

- The **pooled estimator** is given by

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

which has $n_1 + n_2 - 2$ d.f.

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(i) Case 1: $\sigma_1^2 = \sigma_2^2$

The **pivotal random variable** for $\mu_1 - \mu_2$ is

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which has $n_1 + n_2 - 2$ d.f.

Two-sided 100(1-α)% CI for $\boldsymbol{\mu_1 - \mu_2}$:

$$\bar{x} - \bar{y} - t_{n_1+n_2-\frac{\alpha}{2}}\, s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{x} - \bar{y} + t_{n_1+n_2-\frac{\alpha}{2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(i) Case 1: $\sigma_1^2 = \sigma_2^2$

Testing the hypothesis

$$H_0: \mu_1 - \mu_2 = \delta_0 \;\; vs \; H_1: \; \mu_1 - \mu_2 \neq \delta_0$$

where $\delta_0 = \mu_1 - \mu_2$ under H_0.

Typically $\delta_0 = 0$ is used, which corresponds to testing

$$H_0: \mu_1 = \mu_2 \;\; vs \; H_1: \; \mu_1 \neq \mu_2$$

The test statistic

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(i) Case 1: $\sigma_1^2 = \sigma_2^2$

Reject H_0 if

$$|t| > t_{n_1+n_2-2,\alpha/2}$$

or equivalently if

$$|\bar{x} - \bar{y} - \delta_0| > t_{n_1+n_2-2,\alpha/2}\, s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

P-value

$$P\left(\left|T_{n_1+n_2-2}\right| \geq |t|\right) = 2P\left(\left|T_{n_1+n_2-2}\right| \geq |t|\right)$$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(i) Case 1: $\sigma_1^2 = \sigma_2^2$

- Denote the common value of σ_1^2 and σ_2^2 by σ^2, which is unknown.
- An unbiased estimator of this parameter is the sample mean difference $\bar{X} - \bar{Y}$.
- The **sample variances** from the two samples,

$$S_1^2 = \frac{\sum(X_i - \bar{X})^2}{n_1 - 1} \quad \text{and} \quad S_2^2 = \frac{\sum(Y_i - \bar{Y})^2}{n_2 - 1}$$

are both unbiased estimators of σ^2

- The **pooled estimator** is given by

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

which has $n_1 + n_2 - 2$ d.f.

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(i) Case 1: $\sigma_1^2 = \sigma_2^2$

The **pivotal random variable** for $\mu_1 - \mu_2$ is

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which has $n_1 + n_2 - 2$ d.f.

Two-sided 100(1-α)% CI for $\boldsymbol{\mu_1 - \mu_2}$:

$$\bar{x} - \bar{y} - t_{n_1+n_2-\frac{\alpha}{2}}\, s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{x} - \bar{y} + t_{n_1+n_2-\frac{\alpha}{2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(ii) Case 2: $\sigma_1^2 \neq \sigma_2^2$

The **pivotal random variable** for $\mu_1 - \mu_2$ is

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_1^2}{n_2}}}$$

This T does not have a Student t-distribution. But the distribution of T can be *approximated* by Student's t wit h d.f. v. computed as follows.

Denote the standard errors of the means by $SEM_1 = SEM(\bar{x}) = s_1/\sqrt{n_1}$ and $SEM_2 = SEM(\bar{y}) = s_2/\sqrt{n_2}$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(ii) Case 2: $\sigma_1^2 \neq \sigma_2^2$

Let

$$w_1 = SEM_1^2 = \frac{s_1^2}{n_1} \quad \text{and} \quad w_2 = SEM_2^2 = \frac{s_2^2}{n_2}$$

Then the **degrees of freedom** are given by

$$v = \frac{(w_1 + w_2)^2}{w_1^2(n_1 - 1) + w_2^2(n_2 - 1)}$$

****** The d.f. are estimated from data and are not a function of the sample sizes alone.

** The d.f. are generally fractional. For convenience, we will truncate them town to the nearest integer.

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(ii) Case 2: $\sigma_1^2 \neq \sigma_2^2$

Approximate Two-sided 100(1-α)% CI for $\mu_1 - \mu_2$:

$$\bar{x} - \bar{y} - t_{v,\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + t_{n_1+n_2-\frac{\alpha}{2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(ii) Case 2: $\sigma_1^2 \neq \sigma_2^2$

Testing the hypothesis

$$H_0: \mu_1 - \mu_2 = \delta_0 \;\; vs \; H_1: \; \mu_1 - \mu_2 \neq \delta_0$$

where $\delta_0 = \mu_1 - \mu_2$ under H_0.
Typically $\delta_0 = 0$ is used, which corresponds to testing

$$H_0: \mu_1 = \mu_2 \;\; vs \; H_1: \; \mu_1 \neq \mu_2$$

The test statistic

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

8.3 Comparing Means of Two Populations

8.3.1 Independent Samples Design

(2) Inferences for Small Samples

(ii) Case 2: $\sigma_1^2 \neq \sigma_2^2$

Reject H_0 if

$$|t| > t_{v,\alpha/2}$$

or equivalently if

$$|\bar{x} - \bar{y} - \delta_0| > t_{v,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

P-value

$$P(|T_v| \geq |t|) = 2P(|T_v| \geq |t|)$$

This method of obtaining approximate CI's and hypothesis tests based on the approximate t-distribution of r.v. T is known as the **Welch-Satterthwaite method**.

8.4 Comparing Variances of Two Populations

- Check the assumption of equal variances used for the pooled variances (Case 1) methods in Section 8.3.1.
- The methods below are applicable only under the assumption of normality of the data.
- We consider only the independent samples design.

Sample 1: $x_1, x_2, \ldots, x_{n_1}$ is a random sample from an $N(\mu_1, \sigma_1^2)$ distribution
Sample 2: $y_1, y_2, \ldots, y_{n_2}$ is a random sample from an $N(\mu_2, \sigma_2^2)$ distribution

- To compare the two population variances, we use the ratio $\frac{\sigma_1^2}{\sigma_2^2}$.
- The ratio is estimated by s_1^2/s_2^2

8.4 Comparing Variances of Two Populations

Testing the hypothesis

$$H_0\colon\ \frac{\sigma_1^2}{\sigma_2^2} = 1 \ \text{ vs } H_1\colon\ \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

The pivotal r.v.

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

which follows F-distribution with $n_1 - 1$ and $n_2 - 1$ d.f.

Reject $\boldsymbol{H_0}$ if

$$F < f_{n_1-1,n_2-1,1-\alpha/2}$$

or

$$F > f_{n_1-1,n_2-1,\alpha/2}$$

8.4 Comparing Variances of Two Populations

Two-sided 100(1-α)% CI for σ_1^2/σ_2^2:

$$\frac{1}{f_{n_1-1,n_2-1,\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{1}{f_{n_1-1,n_2-1,1-\alpha/2}}$$

** Note that

$$\frac{1}{f_{n_1-1,n_2-1,1-\alpha/2}} = f_{n_2-1,n_1-1,\alpha/2}$$

HW Problem

Two brands of water filters are to be compared in terms of the mean reduction in impurities measured in parts per million (ppm). Twenty-one water samples were tested with each filter and reduction in the impurity level was measured, resulting in the following data:

Filter 1: $n_1 = 21 \quad \bar{x} = 8.0 \quad s_1^2 = 4.5$

Filter 2: $n_2 = 21 \quad \bar{y} = 6.5 \quad s_2^2 = 2.0$

(a) Calculate a 95% confidence interval for the mean difference $\mu_1 - \mu_2$ between the two filters, assuming $\sigma_1^2 = \sigma_2^2$. Is there a statistically significant difference at $\alpha = .05$ between the two filters?

(b) Repeat (a) without assuming $\sigma_1^2 = \sigma_2^2$. Compare the results.

CONTENTS

9.1 Inferences on Proportion

9.1.1 Large Sample Confidence Interval for Proportion

Sample proportion

$$\hat{p} = \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

is an unbiased estimator of p.

By applying the **CLT**, $\hat{p}$ is approximately $N\left(p, \frac{pq}{n}\right)$ distributed for large n.

The **guideline for treating n as large** is

$$n\hat{p} \geq 10 \; and \; n(1-\hat{p}) \geq 10.$$

Approximate $(1-\alpha)$-level CI for p is

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

9.1 Inferences on Proportion

9.1.1 Large Sample Confidence Interval for Proportion

Sample Size Determination for a Confidence Interval on Proportion

$$E = z_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Solving this equation for n gives

$$n = \left(\frac{z_\alpha}{E}\right)^2 \hat{p}\hat{q}$$

9.1 Inferences on Proportion

9.1.2 Large Sample Hypothesis Tests on Proportion

$$H_0\colon p = p_0 \;\; vs \; H_1\colon p \neq p_0$$

When H_0 is true,

$$\hat{p} \approx N\left(p_0, \frac{p_0 q_0}{n}\right), \qquad Y = n\hat{p} \approx N(np_0, np_0 q_0)$$

The standardized statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{y - np_0}{\sqrt{n\hat{p}\hat{q}}}$$

Then the α-level two-sided z-test of $H_0\colon p = p_0$ is equivalent to rejecting H_0 when p_0 falls outside the $(1-\alpha)$-l evel CI.

9.1 Inferences on Proportion

9.1.2 Large Sample Hypothesis Tests on Proportion

Power Calculation and Sample Size Determination for Large Sample Tests on Proportion

$$E(Z) = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \text{ and } Var(Z) = \frac{pq}{p_0 q_0}$$

$$\pi(p) = P\{Z > Z_\alpha | p\} = \Phi\left[\frac{p - p_0\sqrt{n} - z_\alpha\sqrt{p_0 q_0}}{\sqrt{pq}}\right]$$

$$n = \left[\frac{z_\alpha\sqrt{p_0 q_0} + z_\beta\sqrt{p_1 q_1}}{\delta}\right]^2$$

9.1 Inferences on Proportion

9.1.3 Small Sample Hypothesis Tests on Proportion

- Large sample hypothesis tests on p are based on the asymptotic normal distribution of the sample proportion $\hat{p}$ or equivalently of $n\hat{p} = Y$, which is the sample sum.

Exact binomial distribution

$$H_0: p \leq p_0 \;\; vs \; H_1: p > p_0$$

$$P - value = P(Y \geq y \mid p = p_0) = \sum_{i=y}^{n} \binom{n}{i} p_0^i (1 - p_0)^{n-i}$$

9.2 Inferences for Comparing Two Proportions

9.2.1 Independent Sample Design

$$\text{relative risk} = \frac{p_1}{p_2}, \qquad \text{odds ratio} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

(1) Inferences for Large Samples

The guideline for large samples:

$$n_1\hat{p}_1, n_1(1-\hat{p}_1) \geq 10 \quad \textit{and} \quad n_2\hat{p}_2, n_2(1-\hat{p}_2) \geq 10$$

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}$$

For large n_1, n_2,

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} \approx N(0,1)$$

9.2 Inferences for Comparing Two Proportions

9.2.1 Independent Sample Design

The null hypothesis to be tested is $H_0\colon\ p_1 = p_2\ \ (i.e.\delta_0 = 0)$.

A **pooled estimate of *p*** is

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x + y}{n_1 + n_2}$$

An alternative test statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

9.2 Inferences for Comparing Two Proportions

9.2.1 Independent Sample Design

(2) Inferences for Small Samples
Fisher's exact test

Table 9.3 A 2 × 2 Table for Data from Two Independent Bernoulli Samples

	Outcome		Row
	Success	Failure	Total
Sample 1	x	$n_1 - x$	n_1
Sample 2	y	$n_2 - y$	n_2
Column Total	m	$n - m$	n

The test is derived by regarding the total number of successes m as *fixed*, i.e. by conditioning on $X + Y = m$.

$$P(X = i \mid X + Y = m\} = \frac{\binom{n_1}{i}\binom{2}{m-i}}{\binom{n}{m}}$$

9.2 Inferences for Comparing Two Proportions

9.2.2 Matched Pair Design

$A+B+C+D=n$ and the probabilities of the four possible outcomes on a single trial: p_A, p_B, p_C, p_D, where $p_A+p_B+p_C+\ p_D=1$.
Then A,B,C,D have a **multinomial distribution** with sample size $=n$ and the given outcome probabilities.

Table 9.5 A 2×2 Table for Data from Two Matched Pairs Bernoulli Samples

		Condition 2 Response	
		Yes	No
Condition 1	Yes	a	b
Response	No	c	d

The response (success) probability under condition 1 is $p_1=p_A+p_B$ and under condition 2 is $p_2=p_C+p_D$.
Note that $p_1-p_2=p_B-p_C$. (Testing the difference between p_1 and p_2)
$B \sim Bin\left(m, p=\frac{p_B}{p_B+p_C}\right)$

9.2 Inferences for Comparing Two Proportions

9.2.2 Matched Pair Design

$H_0\colon\, p_B = p_C$ becomes $H_0\colon\, p = \frac{1}{2}$, which can be tested by using binomial distribution. (**McNemar's test**). ** (2* 2 contingency table)

$$H_0\colon\, p = \frac{1}{2} \;\text{ vs }\; H_1\colon p > \frac{1}{2}$$

The P-value corresponding to the observed test statistic b is

$$P - value = P(B \geq b | B + C = m) = \sum_{i=b}^{m} \binom{m}{i} \left(\frac{1}{2}\right)^{i} \left(\frac{1}{2}\right)^{m-i} = \left(\frac{1}{2}\right)^{m} \sum_{i=b}^{m} \binom{m}{i}$$

9.2 Inferences for Comparing Two Proportions

9.2.2 Matched Pair Design

If m is large, then the large sample z-statistic with a **continuity correction** can be applied by calculating

$$z = \frac{b - mp_0 - \frac{1}{2}}{\sqrt{mp_0(1 - p_0)}} = \frac{b - \frac{m}{2} - \frac{1}{2}}{\sqrt{\frac{m}{4}}} = \frac{b - c - 1}{\sqrt{b + c}}$$

9.3 Inferences for One-Way Count Data

9.3.1 A test for the Multinomial Distribution

- Denote the cell probabilities by $p_1, p_2, \dots, p_c$, the observed cell counts by $n_1, n_2, \dots, n_c$, and the correspondin g random variables by $N_1, N_2, \dots, N_c$ with $\sum_{i=1}^{c} p_i = 1$ and $\sum_{i=1}^{c} n_i = \sum_{i=1}^{c} N_i = n$.

- The joint distribution of the N_i is the **multinomial distribution** given by
$$P\{N_1 = n_1, N_2 = n_2, \dots, N_c = n_c\} = \frac{n!}{n_1!\, n_2! \dots n_c!} p_1^{n_1} p_2^{n_2} \dots p_c^{n_c}$$
- We consider the problem of testing
$$H_0\colon\ p_1 = p_{10}, p_2 = p_{20}, \dots, p_c = p_{c0} \ \text{ vs } H_1\colon At\ least\ one\ p_i \neq p_{i0}$$

- Assuming that H_0 is true, the expected cell counts e_i is
$$e_i = np_{i0} \quad (i = 1,2, \dots, c)$$

9.3 Inferences for One-Way Count Data

9.3.1 A test for the Multinomial Distribution

The measure of discrepancy is using **Pearson chi-square statistic**

$$\chi^2 = \sum_{i=1}^{c} \frac{(n_i - e_i)^2}{e_i} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

When H_0 is ture, the large sample distribution of the r.v. can be rejected at level α if

$$\chi^2 > \chi^2_{c-1,\alpha}$$

where $\chi^2_{c-1,\alpha}$ is the upper α critical point of the χ^2-distribution with c-1 d.f.

9.3 Inferences for One-Way Count Data

9.3.2 Chi-Squared Goodness of Fit Test

- To determine whether a specified distribution fits a set of data
- If any parameters of the distribution are estimated from the data, then one d.f. is deducted for each indepen dent estimated parameter from the total d.f. c-1.

Thank you!