TECHNISCHE UNIVERSITÄT WIEN

Introduction to Statistics
Means, Proportions and Variances
Inference based on one and two samples

LV Nr. 105.692

Summer Semester 2021

Hypothesis Testing

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1$$

1. The parameter spaces Θ_0 and Θ_1 satisfy

$$\Theta_0 \cap \Theta_1 = \emptyset, \Theta_0 \cup \Theta_1 = \Theta$$

   where Θ is the parameter space.

2. When Θ_0 is a singleton set (contains exactly one point), the null hypothesis is said to be *simple*.

   - For example, $\mu = 0$

Type I and II errors

In a test of hypotheses, the sample space is partitioned into two disjoint regions: the rejection and the acceptance region

- Ω_1: the rejection region are the values of the test statistic T for which we reject the null at level α
- Ω_0: the acceptance region are the values of the test statistic for which we *cannot reject* the null at level α
- Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ is a vector of n iid random variables and $T(\mathbf{X})$ is a test statistic. The decision rule is

$$\text{if } \mathbf{X} \in \Omega_1 \quad \text{then reject } H_0$$

Type I and II errors

- We define two types of error with associated probabilities:

$$
\begin{aligned}
\alpha &= \mathbb{P}_{\theta \in \Theta_0}\left(\mathbf{X} \in \text{rejection region}\right) \\
&= \mathbb{P}\left(\mathbf{X} \in \Omega_1 \mid H_0 \text{ is true}\right) \\
&= \mathbb{P}(\text{Type I error}) \\
\beta &= \mathbb{P}_{\theta \in \Theta_1}\left(\mathbf{X} \in \text{acceptance region}\right) \\
&= \mathbb{P}\left(\mathbf{X} \in \Omega_0 \mid H_0 \text{ is false}\right) \\
&= \mathbb{P}(\text{Type II error})
\end{aligned}
$$

- The power is

$$
\begin{aligned}
\textit{power} = 1 - \beta &= \mathbb{P}_{\theta \in \Theta_1}\left(\mathbf{X} \in \Omega_1\right) \\
&= \mathbb{P}\left(\mathbf{X} \in \Omega_1 \mid H_0 \text{ is false}\right)
\end{aligned}
$$

	Truth	
Decision	H_0	H_1
Accept H_0	correct decision $1-\alpha$	Type II error β
Reject H_0	Type I error α	correct decision $1-\beta$

How to take a decision using α

A small level of significance α is chosen

- If p-value $< \alpha$, then the test decides reject H_0.
- If p-value $\geqslant \alpha$, then the test decides accept H_0.

Type I error and p-value

E.g. for testing $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$, when n is large,

1. Fix α and reject the null if

$$T(\mathbf{X}) = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} > z_\alpha \iff \bar{x}_n > \mu_0 + z_\alpha \frac{z_n}{\sqrt{n}}$$

2. or, equivalently, compute

$$p - value = \mathbb{P}\left(Z > \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}\right)$$

and reject if p-value $\leqslant \alpha$.

Combining Decision Theory and evidence based tests of hypotheses

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1$$

with

$$\Theta_0 \cap \Theta_1 = \emptyset, \; \Theta_0 \cup \Theta_1 = \Theta$$

1. Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ is a vector of n iid random variables and $T(\mathbf{X})$ is a test statistic.
2. Let Ω_0 denote the acceptance and Ω_1 the rejection region of H_0. That is,

$$\text{if } \mathbf{X} \in \Omega_1 \quad \text{reject } H_0$$

3. The general definition of power of this test is the probability of rejecting H_0 as a function of θ:

$$\pi(\theta) = \mathbb{P}_\theta(\text{reject } H_0) = \mathbb{P}_\theta(\mathbf{X} \in \Omega_1) \tag{1}$$

Power

If α denotes the significance level, i.e., the probability value such that

$$p - value < \alpha \iff \text{reject } H_0$$

then

$$\alpha = \mathbb{P}_{\theta \in \Theta_0}(\mathbf{X} \in \Omega_1) = \pi(\theta),\ \theta \in \Theta_0$$

and

$$\text{power} = 1 - \beta = \mathbb{P}_{\theta \in \Theta_1}(\mathbf{X} \in \Omega_1) = \pi(\theta),\ \theta \in \Theta_1$$

Two types of Tests

- Wald Tests: based on the asymptotic distribution of a pivotal statistic that is used as test statistic
  - For X_i iid from $f_\theta(x)$, $i = 1, \ldots, n$, with n large, with $\mathbb{E}(X_i) = \mu < \infty$, $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2 < \infty$, to test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$, we use the test statistic

    $$z = \frac{\bar{x}_n - \mu_0}{\sigma / \sqrt{n}} \approx_{H_0} \mathcal{N}(0, 1)$$

    with rejection region

    $$|z| \geqslant z_{\alpha/2} \iff \bar{x}_n \geqslant \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \text{ or } \bar{x}_n \geqslant \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Likelihood Ratio or Neyman-Pearson Tests: based on the finite sample distribution of a test statistic and they have maximal power (usually)
  - They control the probability of Type I Error at a certain fixed level (α, significance) and minimize the Type II error or maximize power.

Likelihood Ratio Test (Neyman-Pearson)

Assume $\mathbf{X} \sim f_\theta(\mathbf{x})$ and consider two simple hypotheses:

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta = \theta_1$$

- Denote the **likelihood ratio** as

$$\lambda(\mathbf{x}) = \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} = \frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}$$

- Then, the likelihood ratio test with rejection region

$$\Omega_1 = \left\{ \mathbf{x} : \lambda(\mathbf{x}) = \frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} \geqslant C \right\} \tag{2}$$

is the most powerful (MP) test among all tests at significance levels not larger than α, where

$$\alpha = \mathbb{P}_{\theta_0}(\lambda(\mathbf{X}) \geqslant C).$$

MP test for the normal mean

1. $X_1, \ldots, X_n$ iid $\mathcal{N}(\mu, \sigma^2)$, σ is known.
2. We want to test $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1$, $\mu_1 > \mu_0$.
3. The LR is

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{L(\mu_1; \mathbf{x})}{L(\mu_0; \mathbf{x})} \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i - \mu_1)^2 - \sum_{i=1}^{n}(x_i - \mu_0)^2\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mu_0 - \mu_1)(2x_i - \mu_0 - \mu_1)\right) \\
&= \exp\left(\frac{1}{\sigma^2}(\mu_1 - \mu_0)n\left(\bar{x}_n - \frac{\mu_1 + \mu_0}{2}\right)\right)
\end{aligned}$$

MP test for the normal mean

4. By NP Lemma, the MP test at level α rejects H_0 if

$$\lambda(\mathbf{X}) \geqslant C$$

   with C satisfying

$$\mathbb{P}_{\mu_0}(\lambda(\mathbf{X}) \geqslant C) = \alpha$$

5. To solve this and find C we have to find the distribution of the LR under the null: typically difficult (or at least tedious).

MP test for the normal mean

6. Alternatively, we observe $\lambda(\mathbf{x})$ is an increasing function of $\bar{x}_n$ for $\mu_1 > \mu_0$ and that
$$\lambda(\mathbf{x}) \geqslant C \quad \Longleftrightarrow \quad \bar{x}_n \geqslant C^\star$$
where
$$C = \exp\left(-\frac{1}{\sigma^2}(\mu_1 - \mu_0)n\left(C^\star - \frac{\mu_1 + \mu_0}{2}\right)\right)$$
7. The MP test at level α can then be re-written in terms of $\bar{X}_n$: reject H_0 if
$$\bar{X}_n \geqslant C^\star$$
with C satisfying
$$\mathbb{P}_{\mu_0}(\bar{X}_n \geqslant C^\star) = \mathbb{P}_{\mu_0}(\lambda(\mathbf{X}) \geqslant C) = \alpha$$

Generalized Likelihood Ratio Tests

- Let $\mathbf{X} = (X_1, \ldots, X_n) \sim f_\theta(\mathbf{x})$, $\theta \in \Theta$ and test

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta \in \Theta_1$$

  with $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 = \Theta$.

- The generalized likelihood ratio (GLR) for composite hypotheses is

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta} L(\theta, \mathbf{x})}{\sup_{\theta \in \Theta_0} L(\theta, \mathbf{x})} = \frac{\sup_{\theta \in \Theta} f_\theta(\mathbf{x})}{\sup_{\theta \in \Theta_0} f_\theta(\mathbf{x})}$$

Generalized Likelihood Ratio Tests

- The corresponding generalized likelihood ratio test (GLRT) at level α rejects the null if
$$\lambda(\mathbf{x}) \geqslant C$$
where C satisfies
$$\sup_{\theta \in \Theta_0} \mathbb{P}(\lambda(\mathbf{x}) \geqslant C) = \alpha$$

Calculating the GLRT

To calculate $\lambda(\mathbf{x})$ and the GLRT:

1. Find the MLE $\hat{\theta}$ of θ to calculate the numerator $\sup_{\theta\in\Theta} L(\theta,\mathbf{x}) = L(\hat{\theta},\mathbf{x})$
2. Find the MLE $\hat{\theta}_0$ of θ_0 under the restriction $\theta \in \Theta_0$ to calculate the denominator $\sup_{\theta\in\Theta_0} L(\theta,\mathbf{x}) = L(\hat{\theta}_0,\mathbf{x})$
3. Form the generalized likelihood ratio

$$\lambda(\mathbf{x}) = \frac{L(\hat{\theta},\mathbf{x})}{L(\hat{\theta}_0,\mathbf{x})}$$

and find an equivalent simpler test statistic $T(\mathbf{x})$ if possible such that $\lambda(\mathbf{x})$ is its increasing function

4. Find the corresponding critical value for $T(\mathbf{x})$ solving

$$\alpha = sup_{\theta\in\Theta_0}\mathbb{P}(T(\mathbf{X}) \geqslant C)$$

Well known GLRTs

GLR Test are most powerful level α tests

1. one and two-sample t-tests for normal means
2. χ^2 test for the normal variance
3. F-test for normal variances
4. F-test for comparing nested models in regression
5. Pearson's χ^2-test for goodness of fit

Guide to Selecting a One-Sample Hypothesis Test

Data Type: Qualitative (2 outcomes: S, F)

1. Target parameter $p = \mathbb{P}(S)$
2. Sample size n:
   1. Large (both $np_0 \geqslant 15$ and $n(1 - p_0) \geqslant 15$)
   2. Test Statistic for testing $H_0 : p = p_0$

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim_{H_0} \mathcal{N}(0, 1)$$

   1. Small ($np_0 < 15$ or $n(1 - p_0) < 15$)
   2. Another course

Guide to Selecting a One-Sample Hypothesis Test

Data Type: Quantitative

Target parameter: Mean μ

- Test Statistic for testing $H_0 : \mu = \mu_0$ depends on n:
  1. Large sample size n (both $n \geqslant 30$)
     1. σ known:
        $$z = \frac{\hat{\mu} - \mu_0}{\sigma / \sqrt{n}} \approx_{H_0} \mathcal{N}(0, 1)$$
     2. σ unknown:
        $$z = \frac{\hat{\mu} - \mu_0}{s_n / \sqrt{n}} \approx_{H_0} \mathcal{N}(0, 1)$$
  2. Small ($n < 30$)
     1. If the population is $\mathcal{N}(\mu, \sigma^2)$, then
        $$t = \frac{\hat{\mu} - \mu_0}{s_n / \sqrt{n}} \approx_{H_0} t(n-1)$$

Guide to Selecting a One-Sample Hypothesis Test

Data Type: Quantitative

Target parameter: Variance σ^2

- Test Statistic for testing $H_0 : \sigma = \sigma_0$
  1. All sample sizes n
     1. If the population is $\mathcal{N}(\mu, \sigma^2)$, then

$$X^2 = (n-1)\frac{s_n^2}{\sigma_0^2} \sim_{H_0} \chi^2(n-1)$$

Rejection Regions for One-Sample Hypothesis Tests

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \begin{cases} \theta > \theta_0 \\ \theta < \theta_0 \\ \theta \neq \theta_0 \end{cases}$$

Test Statistic Distn	H_1	RR at α level
Z	$\theta > \theta_0$	$z > z_\alpha$
	$\theta < \theta_0$	$z < -z_\alpha$
	$\theta \neq \theta_0$	$\lvert z\rvert > z_{\alpha/2}$
t	$\theta > \theta_0$	$t > t_\alpha(n-1)$
	$\theta < \theta_0$	$t < -t_\alpha(n-1)$
	$\theta \neq \theta_0$	$\lvert t\rvert > t_{\alpha/2}(n-1)$
X^2	$\theta > \theta_0$	$X^2 > \chi^2_\alpha(n-1)$
	$\theta < \theta_0$	$X^2 < \chi^2_{1-\alpha}(n-1)$
	$\theta \neq \theta_0$	$X^2 > \chi^2_{\alpha/2}(n-1)$ or $X^2 < \chi^2_{1-\alpha/2}(n-1)$

Guide to Forming a One-sample Confidence Interval

Data Type: Qualitative (2 outcomes: S, F)

1. Target parameter $p = \mathbb{P}(S)$
    1. Sample size n: Large (both $np \geqslant 15$ and $n(1-p) \geqslant 15$)
        1. A $100(1-\alpha)\%$ CI for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} = \frac{\sum_i x_i}{n}$$

    2. Sample size n: Small ($np < 15$ or $n(1-p) < 15$)
        1. A Wilson adjusted $100(1-\alpha)\%$ CI for p is

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}, \quad \tilde{p} = \frac{\sum_i x_i + 2}{n+4}$$

Guide to Forming a One-sample Confidence Interval

Data Type: Quantitative

Target parameter: Mean μ

- A $100(1-\alpha)\%$ CI for μ is
  1. Large sample size n (both $n \geqslant 30$)
     1. σ known:
        $$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$
     2. σ unknown:
        $$\bar{x} \pm z_{\alpha/2} \frac{s_n}{\sqrt{n}}$$
  2. Small ($n < 30$)
     1. If the population is $\mathcal{N}(\mu, \sigma^2)$, then
        $$\bar{x} \pm t_{\alpha/2}(n-1) \frac{s_n}{\sqrt{n}}$$

Guide to Forming a One-sample Confidence Interval

Data Type: Quantitative

Target parameter: Variance σ^2

- A $100(1-\alpha)\%$ CI for σ^2 is
  1. Sample size n large or small and the population is $\mathcal{N}(\mu, \sigma^2)$

$$\frac{(n-1)s_n^2}{\chi^2_{\alpha/2}} \leqslant \sigma^2 \leqslant \frac{(n-1)s_n^2}{\chi^2_{1-\alpha/2}}$$

Example

Latex allergy in health care workers. Health care workers who use latex gloves with glove powder on a daily basis are particularly susceptible to developing a latex allergy. Symptoms of a latex allergy include conjunctivitis, hand eczema, nasal congestion, skin rash, and shortness of breath. Each in a sample of 46 hospital employees who were diagnosed with latex allergy based on a skin prick test reported on their exposure to latex gloves (Current Allergy and Clinical Immunology, Mar. 2004). Summary statistics for the number of latex gloves used per week are $\bar{x} = 19.3$, $s = 11.9$.

1. Give a point estimate for the average number of latex gloves used per week by all health care workers with a latex allergy.
2. Form a 95% confidence interval for the average number of latex gloves used per week by all health care workers with a latex allergy.
3. Give a practical interpretation of the interval in part 2.
4. Give the conditions required for the interval to be valid.
5. Test the hypothesis that the average number of latex gloves used per week by all health care workers with a latex allergy exceeds 20.

Answer

1. Give a point estimate for the average number of latex gloves used per week by all health care workers with a latex allergy.

   Answer: If we let X_i, $i = 1, \ldots, n = 46$ denote the measurements of the exposure to latex gloves of the 46 hospital employees in the survey,

   $$\hat{\mu} = \bar{x} = 19.3$$

2. Form a 95% confidence interval for the average number of latex gloves used per week by all health care workers with a latex allergy.

   Answer: The sample size is large ($n = 46 > 30$) so we can use the large sample size confidence interval for μ:

   $$\bar{x} \pm z_{\alpha/2} \frac{s_n}{\sqrt{n}}$$

   $$19.3 \pm 1.96 \frac{11.9}{\sqrt{46}} = (15.86106, 22.73894)$$

Answer

3. Give a practical interpretation of the interval in part 2.

   Answer: The probability (15.86106, 22.73894) contains the true average number of latex gloves used per week by all health care workers with a latex allergy is 95%.

4. Give the conditions required for the interval to be valid.

   Answer: Random sample from a population with finite mean and variance.

Answer

5. Test the hypothesis that the average number of latex gloves used per week by all health care workers with a latex allergy exceeds 20.

Answer:

1. State the hypotheses:

$$H_0 : \mu = 20 \quad vs \quad H_1 : \mu > 20$$

2. Test Statistic: Testing hypotheses on a single population mean with a large sample

$$z = \frac{\bar{x} - \mu_0}{s_n/\sqrt{n}} = \frac{19.3 - 20}{11.9/\sqrt{46}} = -0.3989606$$

3. Rejection Region at $\alpha = 0.05$:

$$z > z_\alpha = z_{0.05} = 1.644854$$

Since $z - 0.3989606$ is smaller than 1.644854, we cannot reject the null at leval 0.05.

4. p-value:

$$p - value = \mathbb{P}(Z > z) = 0.6550389$$

In R: `1-pnorm(-0.3989606)`

Guide to Selecting a Two-sample Hypothesis Test and Confidence Interval

Data Type: Qualitative (2 outcomes: S, F)

1. Target parameter $p_1 - p_2 = \mathbb{P}(S \text{ in popn } 1) - \mathbb{P}(S \text{ in popn } 2)$
2. Assumption: Independent samples and large sample sizes: ($n_1 p_1 \geqslant 15$ and $n_1(1-p_1) \geqslant 15$) and ($n_2 p_2 \geqslant 15$ and $n_2(1-p_2) \geqslant 15$)
   1. A $100(1-\alpha)\%$ CI for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

   2. Test Statistic for testing $H_0 : p_1 - p_2 = 0$

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

Guide to Selecting a Two-sample Hypothesis Test and Confidence Interval

Data Type: Quantitative

- Target parameter $\mu_1 - \mu_2 = \mu_d$ for paired samples: mean of the two paired popn differences
  1. Large sample size n (both $n \geqslant 30$)
     1. Test statistic:
        $$z = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} \approx_{H_0} \mathcal{N}(0, 1)$$
     2. $100(1-\alpha)\%$ CI for μ_d:
        $$\bar{d} \pm z_{\alpha/2} \frac{s_d}{\sqrt{n}}$$
  2. Small ($n < 30$)
     1. If the population of differences is $\mathcal{N}(\mu_d, \sigma_d^2)$, then
        $$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} \approx_{H_0} t(n-1)$$
     2. $100(1-\alpha)\%$ CI for μ_d:
        $$\bar{d} \pm t_{\alpha/2}(n-1) \frac{s_d}{\sqrt{n}}$$

Guide to Selecting a Two-sample Hypothesis Test and Confidence Interval

Data Type: Quantitative

- Target parameter $\mu_1 - \mu_2$ for independent samples
- μ_j: mean of popn $j, j = 1, 2$.
  1. Large sample sizes (both $n_1 \geqslant 30$ and $n_2 \geqslant 30$)
     1. Test statistic for $H_0 : \mu_1 - \mu_2 = \mu_0$

$$z = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx_{H_0} \mathcal{N}(0, 1)$$

     2. $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Guide to Selecting a Two-sample Hypothesis Test and Confidence Interval

Data Type: Quantitative

- Target parameter $\mu_1 - \mu_2$ for independent samples
- μ_j: mean of popn $j, j = 1, 2$.
- Small ($n_1 < 30$ or $n_2 < 30$)
  1. If both populations are normally distributed with equal variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then
     1. the test statistic for $H_0 : \mu_1 - \mu_2 = \mu_0$ is

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{\sqrt{s_p^2 \frac{1}{n_1} + \frac{1}{n_2}}} \approx_{H_0} t(n_1 + n_2 - 2)$$

where

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

     2. $100(1 - \alpha)\%$ CI for μ_d:

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}(n_1 + n_2 - 2) \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Guide to Selecting a Two-sample Hypothesis Test and Confidence Interval

Data Type: Quantitative

- Target parameter $\mu_1 - \mu_2$ for independent samples
- μ_j: mean of popn $j, j = 1, 2$.
- Small ($n_1 < 30$ or $n_2 < 30$)
  1. If both populations are normally distributed with unequal variances $\sigma_1^2 \neq \sigma_2^2$, then
     1. the test statistic for $H_0 : \mu_1 - \mu_2 = \mu_0$ is Welch's t
        $$t = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx_{H_0} t(\nu)$$
        where
        $$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1} + \frac{\left(s_2^2/n_2\right)^2}{n_2-1}} \tag{3}$$
     2. $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$:
        $$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}(\nu)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Guide to Selecting a Two-sample Hypothesis Test

Data Type: Quantitative

- Target parameter σ_1^2/σ_2^2 for independent samples
- σ_j^2: variance of popn j, $j = 1, 2$.
- All sample sizes and both populations are normally distributed
- The test statistic for $H_0 : \sigma_1^2 = \sigma_2^2$, or, equivalently, $H_0 : \sigma_1^2/\sigma_2^2 = 1$ is

$$F = \frac{\text{larger sample variance}}{\text{smaller sample variance}} \begin{cases} H_1 : \sigma_1^2 > \sigma_2^2 & \text{RR: } F > F_\alpha \\ H_1 : \sigma_1^2 < \sigma_2^2 & \text{RR: } F > F_\alpha \\ H_1 : \sigma_1^2 \neq \sigma_2^2 & \text{RR: } F > F_{\alpha/2} \end{cases}$$

$$F \sim_{H_0} F(\text{num df, den df})$$

1. The F-test is much less robust to departures from normality than the t-test for comparing population means. If normality is in doubt, use a nonparametric test, such as *Levene's* test.

Example: Comparing two variances

A manufacturer of paper products wants to compare the variation in daily production levels at two paper mills. Independent random samples of days are selected from each mill, and the production levels (in units) are recorded. The data are shown in the table. Do these data provide sufficient evidence to indicate a difference in the variability of production levels at the two paper mills? Use $\alpha = 0.10$.

Table: Production levels at two paper mills.

Mill 1	34	18	28	21	40	23	29
	25	10	38	32	22	22	
Mill 2	31	13	27	19	22	18	23
	22	21	18	15	24	13	19
	18	19	23	13			

Example: Comparing two variances

Let

$$\sigma_1^2 = \text{popn variance of production levels at mill 1}$$
$$\sigma_2^2 = \text{popn variance of production levels at mill 2}$$

The hypotheses of interest are

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad vs \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

The test statistic requires the computation of the two sample variances. Let x_{ij}, $i = 1, \dots, n_j$, $j = 1, 2$ denote the observations from the two mills, respectively, with $n_1 = 13$, $n_2 = 18$.

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = \begin{cases} 8.36^2 \\ 4.85^2 \end{cases}$$
$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} = \begin{cases} 26.31 \\ 19.89 \end{cases}$$

Example: Comparing two variances

Since $s_1^2 = 69.8896 > s_2^2 = 23.5225$, the test statistic is

$$\begin{aligned} F &= \frac{\text{larger sample variance}}{\text{smaller sample variance}} \\ &= \frac{s_1^2}{s_2^2} = \frac{69.8896}{23.5225} = 2.97 \\ &\sim_{H_0} F(n_1 - 1, n_2 - 1) = F(12, 17) \end{aligned}$$

The Rejection Region for $\alpha = 0.10$ is

$$F > F_{\alpha/2}(12, 17) = F_{0.05}(12, 17) = 2.38$$

Since $F = 2.97 > 2.38$, we reject the null in favor of the alternative at $\alpha = 0.10$. The p-value is

$$p - value = 2 \times \mathbb{P}(F(12, 17) > F) = 2 \times 0.02 = 0.04 < \alpha = 0.10$$

Example: Comparing two variances

Let

$$\sigma_1^2 = \text{popn variance of production levels at mill 1}$$
$$\sigma_2^2 = \text{popn variance of production levels at mill 2}$$

The hypotheses of interest are

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad vs \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

The test statistic requires the computation of the two sample variances. Let x_{ij}, $i = 1, \ldots, n_j$, $j = 1, 2$ denote the observations from the two mills, respectively, with $n_1 = 13$, $n_2 = 18$.

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = \begin{cases} 8.36^2 \\ 4.85^2 \end{cases}$$

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} = \begin{cases} 26.31 \\ 19.89 \end{cases}$$

Example: Hull failures of oil tankers

Owing to several major ocean oil spills by tank vessels, improvements in the structural design of a tank vessel have been proposed. To aid in this development, *Marine Technology* (1995) reported on the spillage amount (in thousands of metric tons) and cause of puncture for 42 major oil spills from tankers and carriers. Cause of puncture is classified as either Collision (C), fire/explosion (FE), hull failure (HF), or grounding (G).

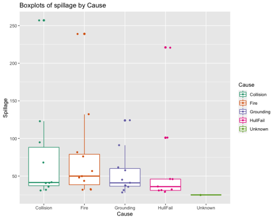
1. Use a graphical method to describe the cause of oil spillage for the 42 tankers. Does the graph suggest that any cause is more likely to occur than any other?
2. Find and interpret descriptive statistics for the 42 spillage amounts. Use this information to form an interval that can be used to predict the spillage amount of the next major oil spill.

Example: Hull failures of oil tankers

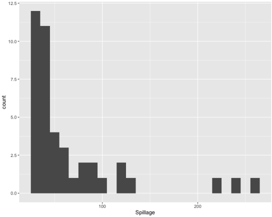
3. Construct a 90% confidence interval for the difference between the mean spillage amount of accidents caused by collision and the mean spillage amount of accidents caused by fire/explosion. Interpret the result.
4. Conduct a test of hypothesis to compare the mean spillage amount of accidents caused by grounding and the mean spillage amount of accidents caused by hull failure.
5. Refer to 3 and 4. State any assumptions required for the inference derived from the analyses to be valid. Are these assumptions reasonably satisfied?
6. Conduct a test of hypothesis to compare the variation in spillage amounts for accidents caused by collision and accidents caused by grounding.

Answers

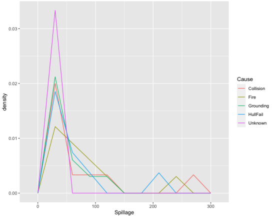
1. The box plots, histogram and frequency plots for the four causes follow. The R code is



(a) Box-plots



(b) Histogram of all data



(c) Frequency plots by Cause

Figure: Graphical displays of Spillage by Cause

Code for 1.

```
df <- read.table("HULL.csv",
                 header = TRUE,
                 sep = ";")

n=nrow(df)
attach(df)

# visualize the data
library(ggplot2)

ggplot(data = df, aes(x=Cause,y=Spillage, color=Cause)) +
  geom_boxplot()+
  scale_color_brewer(palette="Dark2") +
  geom_jitter(shape=16, position=position_jitter(0.2))+
  labs(title = 'Boxplots␣of␣spillage␣by␣Cause',
       y='Spillage',x='Cause')

ggplot(data = df) +
  geom_histogram(mapping = aes(x = Spillage), binwidth = 10)

ggplot(data = df, mapping = aes(x = Spillage, y = ..density..)) +
  geom_freqpoly(mapping = aes(colour = Cause), binwidth = 30)
```

Code for 2.

2. We'll use the following to answer the question:

```
> summary(Spillage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  25.00   32.25   43.00   66.19   74.00  257.00

> sd(Spillage)
[1] 56.0493
```

Therefore, if μ = the mean spillage for tankers and carries, then an approximate 95% CI for μ is based on the asymptotic distribution of $\bar{X} = \sum_{i=1}^{n} X_i/n$, where X_i is the ith amount of spillage and $n = 42$:

$$\bar{x} \pm 2 \times se(\bar{x}) = \bar{x} \pm 2 \times \frac{s_n}{\sqrt{n}} = 66.19 \pm 2 \times \frac{56.0493}{42}$$

3. Let μ_{CO} denote the mean spillage amount of accidents caused by collision and μ_{FE} the mean spillage amount of accidents caused by fire/explosion. We have to construct a 90% CI for

$$\mu_{CO} - \mu_{FE}$$

The formula depends on whether the data are normal and the group variances are equal.

Inspecting the box-plots and frequency plots in Figure 1, the normality assumption does not seem to hold.

Let's also look at summary statistics by group:

Code for 3 and 4.

```
# Summary Statistics by group with tidyverse
library(tidyverse)

df %>%
  group_by(Cause) %>%
  summarise(count = n(),
            mean = mean(Spillage),
            stdev=sd(Spillage),
            min = min(Spillage),
            max = max(Spillage),
            med=quantile(Spillage, 0.5))

# A tibble: 5 x 7
  Cause     count  mean stdev   min   max   med
* <chr>     <int> <dbl> <dbl> <int> <int> <dbl>
1 Collision    10  76.6  70.4    31   257  41.5
2 Fire         11  75    61.9    32   239  50
3 Grounding    11  53.7  29.4    28   124  41
4 HullFail      9  63.7  63.1    29   221  36
5 Unknown       1  25    NA      25    25  25
```

3. We will compute the CI under the assumption of normality but not equal variances! A 90% CI is

$$\hat{\mu}_{CO} - \hat{\mu}_{FE} \pm t_{0.10/2}(\nu) \times \sqrt{\frac{s_{CO}^2}{n_{CO}} + \frac{s_{FE}^2}{n_{FE}}}$$

where

$$\nu = \frac{\left(\frac{s_{CO}^2}{n_{CO}} + \frac{s_{FE}^2}{n_{FE}}\right)^2}{\frac{\left(s_{CO}^2/n_{CO}\right)^2}{n_{CO}-1} + \frac{\left(s_{FE}^2/n_{FE}\right)^2}{n_{FE}-1}}$$

```
CO=(1:n)[Cause=="Collision"]
FE=(1:n)[Cause=="Fire"]
HF=(1:n)[Cause=="HullFail"]
G=(1:n)[Cause=="Grounding"]

t.test(Spillage[CO],Spillage[FE],conf.level = 0.90)

        Welch Two Sample t-test

data:  Spillage[CO] and Spillage[FE]
t = 0.055099, df = 18.067, p-value = 0.9567
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -48.74458  51.94458
sample estimates:
mean of x mean of y
     76.6      75.0
```

4. We will carry out the test

$$H_0 : \mu_G - \mu_{HF} = 0 \quad vs \quad H_1 : \mu_G - \mu_{HF} \neq 0$$

using Welch's t-test since again the variances are very far apart. The test statistic is

$$t = \frac{\bar{x}_G - \bar{x}_{HF}}{\sqrt{\frac{s_G^2}{n_G} + \frac{s_{HF}^2}{n_{HF}}}} \approx_{H_0} t(\nu)$$

where ν is defined according to (3).
The R code is

```
t.test(Spillage[G],Spillage[HF])

        Welch Two Sample t-test

data:  Spillage[G] and Spillage[HF]
t = -0.43522, df = 10.828, p-value = 0.672
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -60.30229  40.42350
sample estimates:
mean of x mean of y
 53.72727  63.66667
```

so we fail to reject the null since p-value is 0.672, i.e. there is no evidence in the data to support the two means are not equal.

5. Both Welch's CIs and tests of hypotheses require the populations be independent and normal but allow for unequal variances.

   The group spillages do not seem to be normally distributed, nevertheless the sample size is very small in order to really form an opinion on this.

   Independence is hard to check so we simply assume the popns defined by different causes are independent.

   The sample variances point to their population counterparts being unequal.

6. We are asked to test

$$\sigma^2_{CO} = \sigma^2_G \quad vs \quad H_1 : \sigma^2_{CO} \neq \sigma^2_G$$

From the computer output, $s^2_{CO} > s^2_G$, hence the test statistic is

$$F = \frac{s^2_{CO}}{s^2_G} = 5.733907$$

with

$$p - value = 2 \times \mathbb{P}\left(F(n_{CO} - 1, n_G - 1) > 5.733907\right) = 0.0116$$

which is rather small, so we reject the null at any α greater than the p-value.

The R code is

```
 2*(1-pf(5.733907,9,10))
[1] 0.01161879
```

Practice problem

A manufacturer of automobile shock absorbers was interested in comparing the durability of its shocks with that of the shocks produced by its biggest competitor. To make the comparison, one of the manufacturer's and one of the competitor's shocks were randomly selected and installed on the rear wheel of each of six cars. After the cars had been driven 20000 miles, the strength of each test shock was measured, coded, and recorded. Results are shown in the table.

Car number	**Manufacturer's shock**	**Competitor's shock**
1	8.8	8.4
2	10.5	10.1
3	12.5	12.0
4	9.7	9.3
5	9.6	9.0
6	13.2	13.0

Do these data present sufficient evidence to conclude there is a difference in the mean strength of the two types of shocks after 20000 miles of use?