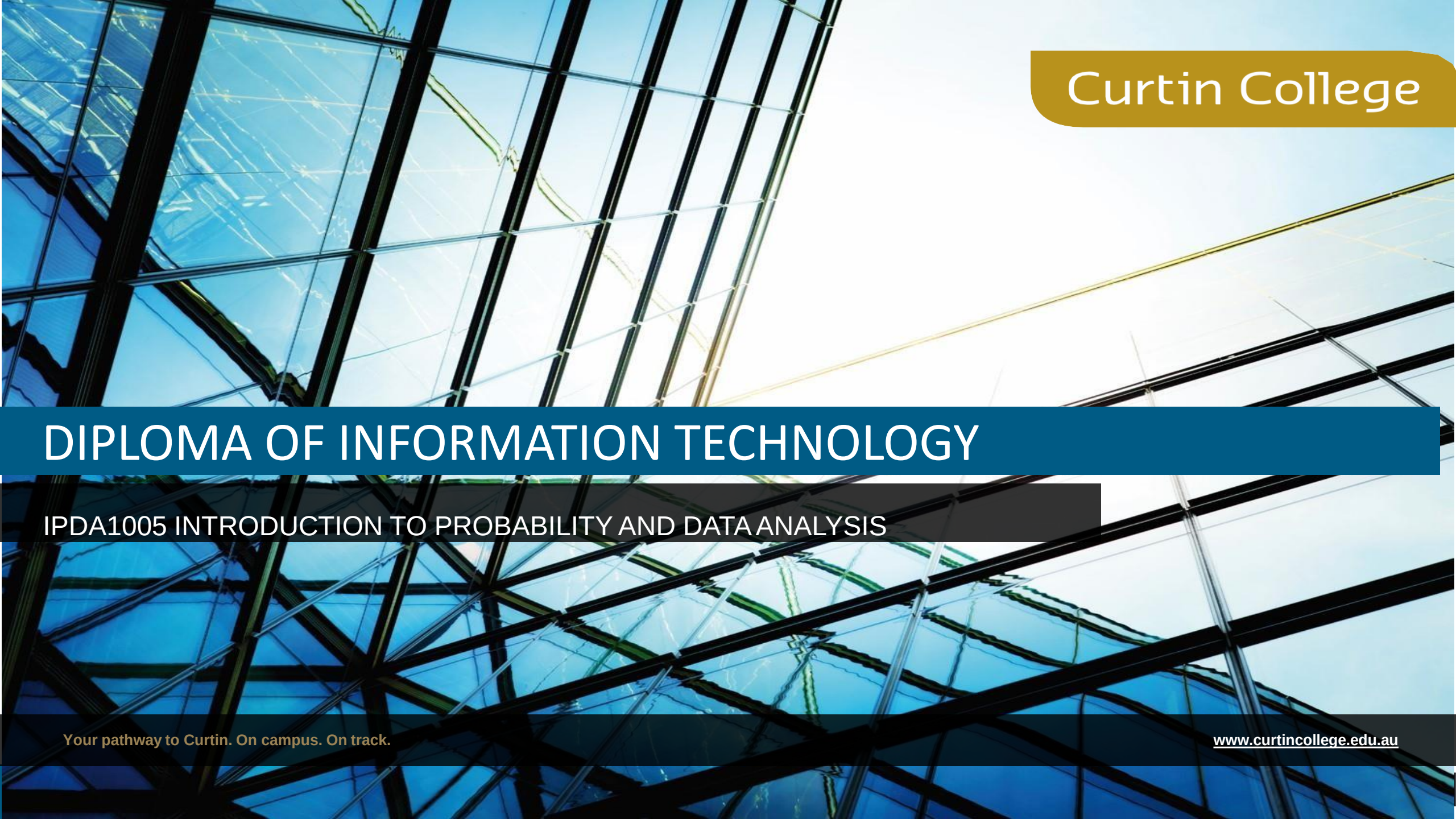

DIPLOMA OF INFORMATION TECHNOLOGY

IPDA1005 INTRODUCTION TO PROBABILITY AND DATA ANALYSIS

Acknowledgement

We respectfully acknowledge the Elders and custodians of the Whadjuk Nyungar nation, past and present, their descendants and kin. Curtin College Bentley Campus enjoys the privilege of being located in Whadjuk / Nyungar Boodjar (country) on the site where the Derbal Yerrigan (Swan River) and the Djarlgarra (Canning River) meet. The area is of great cultural significance and sustains the life and well being of the traditional custodians past and present.

Outline

1. Hypothesis Test for Population Mean

2. Summary of one-sample inference
3. Inference for variance and standard deviation
   a. Inference for variance - Confidence intervals
   b. Hypothesis Testing

4. Inference for Poisson Mean

5. Two-sample Inference
   a. Comparing Two Population Means
   b. Comparing Two Population Proportions

Hypothesis test on a population mean

The general principles of hypothesis testing apply to ALL hypothesis tests without exception. In this lecture I focus on the distinctive features of each situation. Firstly: a hypothesis test on a population mean.

> **Example:** A manufacturer of small appliances employs a market research firm to investigate retail sales of its products by gathering information from a sample of retail stores. Concerning handheld mixers, an SRS of 75 stores found average monthly sales of 24, with $s = 11$. What evidence is there that hand-mixer sales have changed from the previous monthy average of 22?

- $H_0 : \mu = 22 \qquad H_0 : \mu \neq 22$
  where μ is the mean monthly sales of hand-mixers in retail stores.
- We assume $\alpha = .05$

Hypothesis test on a population mean

- As we saw last week, the sampling distribution of $\dfrac{\overline{x} - \mu}{s/\sqrt{n}}$ is t_{n-1}.
  Strictly, this applies only if X is normally distributed.
  In practice (based on simulation studies), we are OK if -
  - the distribution of X is not strongly skewed, and
  - there aren't any outliers

  - and these conditions gradually relax as sample size increases.
- We have a large sample ($n = 75$) so use of the t distribution is pretty safe.
- Also, we are told that the sampling was random, so probability calculations will be valid.
- Test statistic $t = \dfrac{24 - 22}{11/\sqrt{75}} = 1.575.$

Hypothesis test on a population mean

- p-value$= 2P(t_{74} > 1.575) = 2 \times 0.05976 = .1195$
- The p-value exceeds $\alpha = .05$ so we don't reject H_0 at a 5% significance level.
- The data are consistent with monthly mean sales of hand-mixers still being 22.

We are now in a position to summarise inference on 1 population.

One-sample inference summary

	Proportion	Mean
Data	n categorical observations including x successes.	n numerical observations
Hypotheses	$H_0 : p = p_0$ $H_1 : p \neq p_0$ $H_1 : p < p_0$ $H_1 : p > p_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$ $H_1 : \mu < \mu_0$ $H_1 : \mu > \mu_0$
Test stat	$\dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}; \quad \hat{p} = \dfrac{x}{n}$	$\dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$
Sampling distribution	Standard Normal	t_{n-1}

One-sample inference summary

	Proportion	Mean
Assumptions	Large random sample. np_0 and $n(1-p_0)$ both > 10	Random sampling. Population measurements roughly symmetric. No outliers. Can relax as n increases.
Confidence interval	$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\overline{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$
Sample size	$n \geq \left(\frac{z_{\alpha/2}}{e}\right)^2 p(1-p)$	$n \geq \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$

Hypothesis Testing for Paired Data

- Last week we looked at confidence intervals for matched-pair data. The main message is that inference on paired data *is* one-sample inference. Take differences between the pairs and treat the differences as a single sample.
- **Example:** These are average weekly work-hours lost due to accidents in 10 industrial plants before and after a safety program was implemented.

Before	45	73	46	124	33	57	83	34	26	17
After	36	60	44	119	35	51	77	29	24	11

Decide whether the program is effective.

Hypothesis Testing for Paired Data

- **Solution:**
- Calculate the pairwise differences.

Before (B)	45	73	46	124	33	57	83	34	26	17
After (A)	36	60	44	119	35	51	77	29	24	11
B−A	9	13	2	5	-2	6	6	5	2	6

- The difference data set has $n = 10$, $\bar{x} = 5.2$ and $s = 4.08$.
- We are testing $H_0 : \mu_d = 0$ vs. $H_A : \mu_d > 0$, where μ_d is the population mean reduction in lost work hours after implementing the safety program.
- The test requires that the differences are random observations of a Normally-distributed random variable. The sample size is too small for the data to give much indication of the population distribution. We have no information as to how the industrial plants were selected.
- The test statistic is $t = \dfrac{\bar{x}}{s/\sqrt{n}} = \dfrac{5.2}{4.08/\sqrt{10}} = 4.03.$

Hypothesis Testing for Paired Data

- p-value $= P(T_9 > 4.03) = 0.00149$ using *R*'s `pt(4.03,9)`.
- We would reject H_0 at a 1% significance level.
- Subject to evaluation of the assumptions, there is strong evidence that the mean effect of implementing the safety programme was to reduce weekly work hours lost due to accidents. The mean reduction was 5.2 hours per week. It appears that the safety program was effective.

- *R*'s `t.test()` command will do the entire calculation from raw data.

Inference for variance - sampling distribution

- If we have an random sample $\{X_1, X_2 \dots X_n\}$ from a $N(\mu, \sigma^2)$ population, then $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- It is also true that $S^2 \sim \frac{\sigma^2}{(n-1)}\chi^2_{n-1}$
- χ^2_{n-1} is a *chi-squared* distribution with $n-1$ degrees of freedom.
- The sampling distribution for S^2 is sensitive to the population distribution assumption. It should not be used where sample data appear non-normal. In such cases, more sophisticated approaches are available, which we do not cover in this unit.

Confidence interval for variance and standard deviation

- The sampling distribution leads to the following confidence interval for population variance.
- Given a random sample of size n from a normally-distributed population, a $100(1-\alpha)\%$ confidence interval for σ^2 is:

$$\left(\frac{(n-1)\,s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\,,\ \frac{(n-1)\,s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)$$

- Confidence intervals for σ can be found by taking square-roots.
  - χ^2 distributions are not symmetric, so this CI is NOT in the form
    estimate $\pm$ margin of error,
    and is not symmetric around the point estimator s^2.
- Look up each chi-squared value in tables or use `R`'s `qchisq` command.

Confidence Interval for standard deviation

- A food inspector examined 12 jars of a certain brand of peanut butter and determined the percentages of impurities, yielding these results: $2.3, 1.9, 2.1, 2.8, 2.3, 3.6, 1.4, 1.8, 2.1, 3.2, 2.0, 1.9$
- Construct 90% confidence intervals for the variance and the standard deviation.
- **Solution:** From the data, $s^2 = 0.3906$.
- From tables or R, χ^2_{11} values for 0.95 and 0.05 are 19.675 and 4.575.
- So the 90% confidence interval for the variance is given by

$$\left(\frac{11 \times 0.3906}{19.675} \, , \, \frac{11 \times 0.3906}{4.575} \right) = (0.2184, 0.9392) \, .$$

- The 90% confidence interval for the standard deviation is $(.467, .969)$.
- A boxplot shows the data are positively skewed, so we should not take the confidence intervals too seriously.

Hypothesis Testing for Variance and Standard Deviation

- Testing the hypothesis $H_0 : \sigma^2 = \sigma_0^2$ (or $H_0 : \sigma = \sigma_0$) is as follows.
- Test statistic is $\chi^2 = \dfrac{(n-1)\, s^2}{\sigma_0^2}$.
- Compare to χ^2_{n-1} distribution.
- Assumptions:
  - random sampling
  - normally distributed population

Hypothesis Testing for Variance and Standard Deviation

An experiment was conducted to measure the specific heat of iron and a random sample of size 9 resulted in a standard deviation of 0.0086. Assuming normality of the underlying population, test the hypothesis that $\sigma = 0.01$ against the alternative hypothesis that $\sigma < 0.01$. Use the 0.05 level of significance.

- **Solution:** $H_0 : \sigma = 0.01$ vs. $H_A : \sigma < 0.01$,
  where σ is the population standard deviation for specific heat measurements on iron.
- We are given that sampling was random and instructed to assume normality of population measurements.
- With $n = 9, s = 0.0086$ our test statistic is
  $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{8(.0086)^2}{.01^2} = 5.92$
- p-value $= P(\chi_8^2 < 5.92) = 0.3438$, using R's `pchisq(5.92,8)`.
- Since p-value $> .05$ we do not reject H_0 at a 5% significance level.
- The data provide little evidence that the population standard deviation is less than 0.01.

Hypothesis Testing for Variance and Standard Deviation

Based on a sample of size 10 that gave $s = 1.5$, test $H_0 : \sigma = 1$ against $H_A : \sigma \neq 1$ at the 0.05 level of significance. Past experience indicates that population measurements are approximately normally distributed.

- **Solution:** $H_0 : \sigma = 1$ vs. $H_A : \sigma \neq 1.$,
  where σ is the population standard deviation.
- We require random sampling from a normally distributed population. We are given that the population is approximately normal, but there is no information about how the sample was selected.
- $n = 10, s = 1.5$, so test stat $\chi^2 = \dfrac{(n-1)\,s^2}{\sigma_0^2} = \dfrac{9(1.5)^2}{1^2} = 20.25.$
- p-value$= 2P(\chi_9^2 > 20.25) = 0.03286.$
- Since p-value $< .05$ we reject H_0 at the 5% level.
- If sampling was random, there is significant evidence that the population standard deviation is not 1. The estimated value is 1.5.

Inference for Poisson Mean

- For inference on the parameter λ of a Poisson distribution, we restrict ourselves to cases where we can approximate the Poisson distribution by a normal distribution.
- For confidence intervals, the approximation is valid if $n\hat{\lambda} = \sum x_i \geq 20$.
- For hypothesis tests, the approximation is valid if $n\lambda_0 \geq 20$.
- For Poisson(λ), $\mu = \sigma^2 = \lambda$, so we approximate Poisson(λ) by $N(\lambda, \lambda)$.
- Then $E(\overline{X}) = \lambda$, $Var(\overline{X}) = \dfrac{\lambda}{n}$ and $\overline{X} \sim N\left(\lambda, \dfrac{\lambda}{n}\right)$

Confidence Interval for Poisson Mean

An approximate $100(1-\alpha)\%$ confidence interval for Poisson mean λ is given by

$$\overline{x} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{x}}{n}}.$$

- **Example:** A random sample of size 25 from a Poisson(λ) population resulted in a sample mean of 7. Construct a 95% confidence interval for λ.
- **Solution:** $\sum x_i = n\overline{x} = 175 \geq 20$, so the condition needed for normal approximation is met. The 95% confidence interval is given by

$$7 \pm 1.96\sqrt{\frac{7}{25}} = 7 \pm 1.037 = (5.963, 8.037)$$

Confidence Intervals for Poisson Mean

- **Example:** The daily number of accidents on a section of a highway is assumed to have a Poisson distribution. A random sample of 100 days showed a total of 35 accidents. Construct a 99% confidence interval for the average number of accidents per day.
- **Solution:** $\sum x_i = 35 \geq 20$, so the condition needed for normal approximation is met. Here $\overline{x} = \frac{35}{100} = 0.35$. The 99% confidence interval is given by

$$0.35 \pm 2.576\sqrt{\frac{0.35}{100}} = 0.35 \pm 0.152 = (0.198, 0.502).$$

- Hypothesis testing for Poisson mean uses the same sampling distribution, with assumptions
  - random sampling
  - $n\lambda_0 \geq 20$
- The test statistic is $z = \dfrac{\overline{x} - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}}$.

Hypothesis Testing for Poisson Mean

Example: For the highway accident example, test the hypothesis that the average daily number of accidents is less than 0.4. Use a 5% significance level.

- **Solution:** We test $H_0 : \lambda = 0.4$ against $H_A : \lambda < 0.4$, where λ is the population mean daily number of accidents.
- The test requires random sampling and $n\lambda_0 \geq 20$. We were told that the sampled days were randomly selected, and $n\lambda_0 = 100 \times 0.4 = 40$, well over 20, so the size is sufficient.
- The test statistic is $z = \dfrac{\overline{x} - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}} = \dfrac{0.35 - 0.4}{\sqrt{\frac{0.4}{100}}} = -0.7906$
- p-value $= P(Z < -0.7906) = 0.2146$
- Since p-value exceeds 5% we do not reject H_0 at the 5% significance level.
- There is insufficient evidence to conclude that the average number of accidents per day is less than 0.4.

Two-sample Inference

- This section covers confidence intervals and hypothesis testing involving parameters from two populations.
- All these methods require *independent random samples*. That is, the samples are selected -
  - randomly from each population, and
  - independently of each other.
- This is in contrast with paired data, where there is pair-wise dependence between two sets of data.
- We shall briefly cover:
  - Difference of Normal means
  - Difference of proportions

Difference of Normal Means

Comparing two populations or two treatments is a common statistical investigation. We select random samples from two populations, or randomly allocate individuals between two treatments.

Examples:

- Patients with lower back pain are randomly assigned to two groups. 142 received a single examination and advice from a therapist; another 144 received regular physical therapy for up to 5 weeks. After a year, the change in their level of pain was assessed by a doctor who did not know which treatment the patient received.
- A psychologist develops a test that measures social insight. He investigates gender variation in social insight by giving the test to a sample of female students and a sample of male students.
- A bank wants to know which of two incentive plans will increase the use of its credit cards. It offers each incentive to a random sample of credit card customers and compares the amounts charged over the next six months.

Comparing Two Population Means - sampling distribution

- Inference concerning $\mu_1 - \mu_2$ depends on the sampling distribution of $$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
- This sampling distribution is not known exactly, but the approximate sampling distribution is a t-distribution with degrees of freedom $$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$
- If working from raw data, R calculates df for you.
- If calculating manually,

  let $A = s_1^2/n_1, \quad B = s_2^2/n_2, \quad$ then $df = \dfrac{(A+B)^2}{\frac{A^2}{n_1 - 1} + \frac{B^2}{n_2 - 1}}$

Comparing Two Population Means - assumptions

Assumptions

- Formally, the sampling distribution requires normally distributed populations, but in practice we can be a bit more relaxed.
- Moderate skewness is OK.
- If the populations are more strongly skewed but *similarly* skewed, this is also well tolerated.
- As in the one-sample test, outliers are undesirable.
- As in the one-sample test, the conditions relax as sample size increases.

Confidence Interval for Difference of Means

A $100(1-\alpha)\%$ confidence interval for the difference between the means of two Normally distributed populations is given by

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

where $\nu =$ degrees of freedom as given earlier.

Example: Two independent random samples of size 40 and 50 from normally distributed populations yielded sample means 13.5 and 9.3. The corresponding sample variances were 14.4 and 112.5. Find an approximate 99% confidence interval for the difference between the means.

Confidence Interval for Difference of Means

Solution:

- We first calculate the degrees of freedom.
  $A = 14.4/40 = 0.36, \quad B = 112.5/50 = 2.25$
  $$df = \frac{(A+B)^2}{\frac{A^2}{n_1-1} + \frac{B^2}{n_2-1}} = \frac{(0.36+2.25)^2}{\frac{0.36^2}{39} + \frac{2.25^2}{49}} = 63.88$$

- The relevant t-score is qt(.005, 63.88) $= -2.655$

- Then the CI is
  $$\begin{aligned} 13.5 - 9.3 \pm 2.655\sqrt{\frac{14.4}{40} + \frac{112.5}{50}} &= 4.2 \pm 2.655 \times 1.6155 \\ &= 4.2 \pm 4.289 \\ &= (-0.089, 8.489) \end{aligned}$$

Hypothesis Testing for Difference of Means

Example: Consider the situation given in the previous example. Test at the .01 level of significance the hypothesis that the means of these populations are same against the alternative hypothesis that the first population has a larger mean.

- **Solution:** The null and the alternative hypotheses are $H_0 : \mu_1 - \mu_2 = 0$ against $H_A : \mu_1 - \mu_2 > 0$.
- We require independent random samples, ideally from normally distributed populations. The given information is that that is exactly what we have.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{13.5 - 9.3}{\sqrt{\frac{14.4}{40} + \frac{112.5}{50}}} = \frac{4.2}{1.6155} = 2.600$$

- Using the df calculated earlier, p-value$= P(T_{63.88} > 2.6) = 0.0058$.
- Since p-value $< .01$ we reject H_0 at a 1% significance level. There is significant evidence that the first population has a larger mean. Estimates of the two means are $\mu_1 \approx 13.5,\ \mu_2 \approx 9.3$.

Confidence Interval for Difference of Proportions

- Suppose we have two population proportions p_1 and p_2 which we estimate using samples of size n_1 and n_2.
- $\hat{p}_1$ and $\hat{p}_2$ denote the sample proportions that we use to estimate p_1 and p_2 respectively.

An approximate $100(1-\alpha)\%$ confidence interval for the difference of proportions $p_1 - p_2$ is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\hat{V}}$$

where $\hat{V} = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$ is the estimated variance of $\hat{P}_1 - \hat{P}_2$.

- Like the 1-sample CI for a proportion, the interval is approximate because we are using sample proportions to estimate the variance.

Example

Example: In a one-year mortality investigation, 25 of the 100 ninety-year-old males and 30 of the 150 ninety-year-old females died before the end of the year. Construct a 95% confidence interval for the difference in mortality rates between males and females.

Solution:

- The sample proportions are $\hat{p}_m = \frac{25}{100} = 0.25, \quad \hat{p}_f = \frac{30}{150} = 0.20$
- The variance estimate is $\hat{V} = \frac{0.25(0.75)}{100} + \frac{0.2(0.8)}{150} = 0.002942$
- The 95% confidence interval for $p_1 - p_2$ is

$$0.25 - 0.20 \pm (1.96)\sqrt{0.002942} = 0.05 \pm 0.106 = (-0.056, 0.156)$$

Hypothesis Testing for Difference of Proportions

- Here we test $H_0 : p_1 = p_2$ (or $H_0 : p_1 - p_2 = 0$) against a one-sided or two-sided alternative.
- The test requires large independent random samples.
- In the CI the estimated variance $\hat{V} = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$
- Under H_0 the two population proportions are equal, and the best estimator of that unknown common proportion is the *pooled sample proportion* given by

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

- Then $\hat{V}$ becomes $\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$
- and the test statistic is $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

Hypothesis Testing for Difference of Proportions

Example: Does the mortality rate data indicate that males have a higher mortality rate than females?

Solution:

- We test $H_0 : p_m = p_f$ against $H_A : p_m > p_f$, where p_m, p_f are respectively the male and female mortality rates.
- We require large independent random samples. We have no information regarding random sampling among the males and females. Since the samples contain different people they are probably independent. The sample sizes of 100 and 150 are reasonably large.
- Pooled proportion $\hat{p} = \dfrac{25 + 30}{100 + 150} = 0.22.$
- Test statistic $z = \dfrac{0.25 - 0.2}{\sqrt{0.22 \times 0.78 \left(\frac{1}{100} + \frac{1}{150}\right)}} = \dfrac{0.05}{0.05348} = 0.935$
- p-value $= P(Z > 0.935) = 0.1749$, so we do not reject H_0 even at a high significance level like 10%.
- The population mortality rates may be the same.