Course summary

Stephanie Ranft S2459825

March 22, 2020

Statistics 1B
PSBE1-09

1 Overview

§ 5 Statistical inference: estimation

§ 5.1 Point and interval estimation

- *Point estimate*: estimates a parameter with a single number, e.g. $\bar{x}$ is a point estimate of μ.
- *Interval estimate*: estimates a parameter with a range, e.g. confidence interval $\bar{x} \pm z^* \sigma / \sqrt{n}$ is an interval estimate of μ.
- *Confidence interval* = point estimate ± margin of error. The confidence level comes from: the probability that the estimate is within $z^*_{1-\alpha/2}$ standard errors from the parameter is equal to $(1-\alpha)\%$, e.g. the probability that $\bar{x}$ is in the range $(\mu - 1.96 \times \sigma/\sqrt{n}, \mu + 1.96 \times \sigma/\sqrt{n})$ is 95%.
  - In the long run, $(1-\alpha)\%$ of such intervals would contain the parameter.
  - With repeated sampling, there is a $(1-\alpha)\%$ chance that a random confidence interval will cover the true parameter value.
  - $(1-\alpha)\%$ percent of the time (in repeated sampling from the same population) the parameter will lie in the range of the confidence interval.
- *Margin of error* = critical statistic (z^* or t^*, for e.g.) × standard error of the point estimate. As the sample size increases, the standard error of the estimate decreases and thus so does the margin of error.
- *Unbiased estimate*: the expected value of the point estimate is equal to the parameter - another way to write it is $\text{Bias}(\bar{x}) = \mu - \mathbb{E}(\bar{x})$. An estimator is unbiased if the bias equals zero. Recall: the expected value is basically the mean.
- *Efficient estimator*: an unbiased estimator which has a small standard error (relatively speaking) is also efficient. You can Google the Cramer-Rao inequality if you have interest, but it is not pertinent to this course. For example, $\bar{x}$ is efficient but Q_2 is not.
- *Notation*: $\hat{}$ over a parameter denotes an estimator of the parameter, e.g. $\hat{\mu}$ estimates μ.
- *Central Limit Theorem (CLT)*: due to the *Law of Large Numbers (LLN)*, as the sample size increases the estimate tends towards the parameter which is seen by the decreasing standard error.

§ 5.2 Confidence interval for a proportion

- *CI for proportion π*:

$$\hat{\pi} \pm z^*_{1-\alpha/2} \times se_{\hat{\pi}}, \quad \text{where } se_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}.$$

§ 5.3 Confidence interval for a mean:

- *CI for mean μ*:

$$\bar{x} \pm z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \quad \text{where } \sigma \text{ is known.}$$

$$\bar{x} \pm t^*_{1-\alpha/2,\ \text{df}} \times \frac{s}{\sqrt{n}}, \quad \text{where } \sigma \text{ is unknown.}$$

$t^*_{1-\alpha/2,\ \text{df}}$ is larger than $z^*_{1-\alpha/2}$ to account for the additional error introduced by using s to estimate σ for the standard error. They are equal for "infinite sample sizes", or near equal for very large sizes.

- *Robustness*: a method is robust if it performs well despite violations of certain assumptions, e.g. normality. This is important from the book: *...the confidence interval for a mean using the t distribution is robust against violations of the normal population assumption. Even if the population is not normal, confidence intervals based on the t distribution still work quite well, especially when n exceeds about 15. As the sample size gets larger, the normal population assumption becomes less important, because of the Central Limit Theorem. The sampling distribution of the sample mean is then bell shaped even when the population distribution is not. The actual probability that the 95% confidence interval method contains μ is close to 0.95 and gets closer as n increases.*

  An important case when the method does not work well is when the data are extremely skewed or contain extreme outliers. Partly this is because of the effect on the method, but also because the mean itself may not then be a representative summary of the centre. This last part refers to using the median or the mode to summarise the centre, instead of the mean, because of skew.

§ 5.4 Choice of sample size

- *Limiting the margin of error for a certain confidence level*: suppose you want to ensure that the margin of error is less than (or equal to) some value, say A, with probability $(1-\alpha)\,\%$ then you need to alter the sample size, e.g.

$$\underbrace{\overbrace{z^*_{1-\alpha/2}}^{\text{critical value}} \times \overbrace{\frac{\sigma}{\sqrt{n}}}^{\text{standard error}}}_{\text{margin of error}} \le A \qquad \iff \qquad n \ge \left(z^*_{1-\alpha/2} \times \frac{\sigma}{A}\right)^2 .$$

  This works with any standard error formula (e.g. using σ or s) and any critical value (e.g. $z^*_{1-\alpha/2}$ or $t^*_{1-\alpha/2,\ \mathrm{df}}$); just rearrange to solve for n.

§ 5.5 Estimation Methods: Maximum Likelihood and the Bootstrap

- From the book: *For point estimation, Fisher proposed the maximum likelihood estimate. This estimate is the value of the parameter that is most consistent with the observed data, in the following sense: If the parameter equalled that number (i.e., the value of the estimate), the observed data would have had greater chance of occurring than if the parameter equalled any other number.*
- Also: *In fact, with random sampling, the maximum likelihood estimate of a population proportion is necessarily the sample proportion.*
- *Maximum likelihood estimator (MLE)*: is the most efficient for large enough samples, consistent (bias diminishes to zero as n increases), and have approximately normal sampling distribution. This normality of the sampleing distribution allows us to compute confidence intervals.
- $\bar{x}$ is the MLE for μ, as the standard error for Q_2 is $1.25\sigma/\sqrt{n}$, which is more than the standard error for $\bar{x}$.
- *Bootstrap method*: treat the sampling distribution as if it is the population distribution, e.g. $\bar{X} \sim \mathcal{N}\left(\mu, \sigma/\sqrt{n}\right)$ is the sampling distribution of the mean. Then take n samples (with replacement) from the sampling distribution, and compute the mean. Do this about 1000 times, and the means of the means will be near the parameter value for the population. (See the simulation `http://onlinestatbook.com/stat_sim/sampling_dist/index.html`.)

§ 6 Statistical Inference: Significance Tests

(a) The five parts of a significance test

- *Hypothesis*: a predictive statement about a parameter that can be tested. The null hypothesis always corresponds to "no difference", "no effect" or "no association" to protect society from any dramatic changes (Type I errors). The alternative is the converse of the null hypothesis, or the research hypothesis - if "no effect" then what we expect. We do not "accept" either, we only choose to "reject" or "not reject" the null hypothesis.
- *Significance test*: uses sample data to summarise evidence about a hypothesis.
- The five parts are: assumptions, hypotheses, test statistic, p-value, and conclusion.
- *Test statistic*: summary statistic used in significance test, e.g. z-value or t-value.
- *p-value*: the probability of achieving a test statistic equal to or more extreme than that which we observed, e.g. $\mathbb{P}\left(|Z| > z\right) = p$ (two-tailed H_a), $\mathbb{P}\left(Z > z\right) = p$ (right-tailed H_a) or $\mathbb{P}\left(Z < z\right) = p$ (left-tailed H_a). The smaller the p-value, the greater the evidence against H_0 and thus for rejecting it. We decide to reject for $p < \alpha$, where α is our accepted probability of a Type I error.

- *α-level*: if you decrease α, the critical value z^* or t^* increases. The converse is also true. α is the probability of a Type I error.

(b) Significance test for a mean

- *Assumptions:* random sampling and parameter is normally distributed.
- *Hypotheses:* H_0: $\mu = \mu_0$ and H_a: $\mu \neq \mu_0$ for two-sided, $\mu > \mu_0$ for right-tailed or $\mu < \mu_0$ for left-tailed.
- *Test statistic:* z-value for σ in the population known, or t-value for σ in the population unknown.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(\text{df} = n - 1).$$

- *p-value*: probability of observing a test statistic at least or more extreme (in the direction of H_a).
- *Conclusion*: reject H_0 if $p < \alpha$ and assumptions are met.

(c) Significance test for a proportion

- *Assumptions:* random sampling and sampling distribution of estimate is approximately normal. The latter implies that $n\hat{\pi}$ and $n(1 - \hat{\pi})$ must both be at least 10.
- *Hypotheses:* H_0: $\pi = \pi_0$ and H_a: $\pi \neq \pi_0$ for two-sided, $\pi > \pi_0$ for right-tailed or $\pi < \pi_0$ for left-tailed.
- *Test statistic:*

$$z = \frac{\hat{\pi} - \pi_0}{se_0} \sim \mathcal{N}(0,1); \qquad se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}.$$

- *p-value*: probability of observing a test statistic at least or more extreme (in the direction of H_a).
- *Conclusion*: reject H_0 if $p < \alpha$ and assumptions are met.

(d) Decisions and types of errors in tests

- *Type I error*: probability of incorrectly rejecting the null hypothesis, α. (H_0 is true in reality.)
- *Type II error*: probability of incorrectly not rejecting the null hypothesis, β. (H_a is true in reality.)
- *Power*: probability of correctly rejecting the null hypothesis, $1 - \beta$. (H_a is true in reality.)
- *Sensitivity*: probability of correctly not rejecting the null hypothesis, $1-\alpha$. (H_0 is true in reality.)
- *alpha* and β are inversely proportional, meaning that a decrease in α results in an increase in β, and vice-versa. Recall that power is $1 - \beta$, so decreasing α decreases the power of your test.
- If the test statistic is in the rejection region, then $p < \alpha$ and the hypothesised parameter value, e.g. μ_0, is not included in the confidence interval of the same α level.

(e) Limitations of significance tests

- *Statistical vs. practical significance*: a small p-value only signifies that the probability, under the null hypothesis, of observing a test statistic as or more extreme is very small. This does not give any indication of the true value of the parameter. Confidence intervals are "better" in the sense that they do give an indication of the true value of the parameter.
- *Effect size*: used to summarise practical significance, e.g. indicates how likely a researcher is observe an effect. Cohen's d gives the z-value, which should be at least 0.8 to have a large effect; medium is around 0.5 and small is ≤ 0.2.
- Read page 173.

(f) *Finding $\mathbb{P}$(Type II error)*

- Calculate the cut-off value, i.e. the sample statistic for which you would reject H_0, e.g. $\bar{x} > \mu_0 + z^* \times \sigma/\sqrt{n}$ is the cut-off value for a right-tailed hypothesis test.
- Use the cut-off value to calculate the z-value using $\mu = \mu_a$, where μ_a is the true parameter value (not H_0), e.g.

$$z = \frac{\left[\mu_0 + z^* \times \sigma/\sqrt{n}\right] - \mu_a}{\sigma/\sqrt{n}} = \frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} + z^*.$$

- Compute the p-value, e.g. $p = \mathbb{P}(Z > z)$ where z was calculated above. This p-value is the probability of a Type II error.
- $1 - p$-value gives you the power.

(g) Small-sample test for a proportion— the Binomial distribution

- $X \sim \mathcal{B}in(n,p)$ is the probability distribution of "success" with probability p and n trials, and $\mathbb{P}(X = k)$ is the probability of k "successes".

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

$\hat{p}$ estimates p and can be written at x/n, where x is the number of successes in a sample of size n.

- The mean is np and the standard deviation is $\sqrt{np(1-p)}$.
- If the expected frequencies for both success and failure, np and $n(1-p)$, are both greater than 10, then you can approximate using the *Normal approximation to the Binomial distribution*, $X \sim \mathcal{N}\left(\mu = np, \sigma = \sqrt{np(1-p)}\right)$.

§ 7 Comparison of two groups

§ 7.1 Preliminaries for comparing groups

- *Response variable*: the dependent variable, usually the y values is (x,y) bivariate data sets.
- *Explanatory variable*: the independent variable, usually the x values is (x,y) bivariate data sets.
- *Longitudinal study*: same subjects but testing at 2 or more time periods. Usually you compare the difference. Also called matched pairs or dependent sample.
- Read for yourself from page 193 about *cross-sectional studies*, *experimental studies*, *experimental* vs. *control group*, etc.
- If two estimates from independent samples have standard errors se_1 and se_2, then the sampling distribution of their difference has the estimated standard error $\sqrt{se_1^2 + se_2^2} = \sqrt{s_1^2/n_1 + s_2^2/n_2}$.
- *Ratios*: if the ratio of two values equals 1, then they are equal, e.g. if the ratio of sample means is $11.9/8.3 = 1.43$ then one of the sample means is 1.43 times the other. Can be used for effect size.

§ 7.2 Categorical data: comparing two proportions

- If you have two populations with proportions π_1 and π_2 and you calculate the sample proportions $\hat{\pi}_1$ and $\hat{\pi}_2$, the sampling distribution of their difference is Normal with mean $\pi_2 - \pi_1$ and variance $\hat{\pi}_1(1-\hat{\pi}_1)/n_1 + \hat{\pi}_2(1-\hat{\pi}_2)/n_2$ (take the square root of the variance for the standard error). We estimate $\pi_2 - \pi_1$ with $\hat{\pi}_1 - \hat{\pi}_2$.
- *Confidence interval for difference of proportions*:

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z^*_{1-\alpha/2} \times se; \qquad se = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

If the confidence interval contains zero, then you cannot reject the claim that the proportions are the same. This inference depends on the standard error, of course, as a large standard error would indicate that there is a lot of error associated with the estimate. If zero is in the CI, then the lower bound must be *negative* and the upper bound must be *positive.*
If the CI contains only negative values then we infer that $\pi_2 < \pi_1$, and conversely if the CI contains only positive values then we infer that $\pi_2 > \pi_1$.

- *Hypothesis testing*: we assume that there is not difference; H_0: $\pi_1 = \pi_2$, $\pi = \pi_2 - \pi_1 = 0 = \pi_0$.

$$\hat{\pi} = \frac{\hat{\pi}_1 n_1 + \hat{\pi}_2 n_2}{n_1 + n_2} = \frac{\sum x_1 + \sum x_2}{n_1 + n_2},$$

where x_1 is the sum of successes in sample 1, and similarly for x_2 - it is the sum of successes in sample 2.

$$\Longrightarrow z = \frac{(\hat{\pi}_2 - \hat{\pi}_1) - \pi_0}{se_0} = \frac{\hat{\pi}_2 - \hat{\pi}_1}{se_0};$$

$$se_0 = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

§ **7.3** Quantitative data: comparing two means

- *Confidence interval for the difference in means*: for independent random samples from two normally distributed populations, the CI for $\mu_2 - \mu_1$ is

$$(\bar{y}_2 - \bar{y}_2) \pm t^*_{1-\alpha/2,\ \mathrm{df}=n_1+n_2-2} \times se; \qquad se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

  This method is robust to violations of the normal assumption provided that $n \geq 30$ because of the CLT. Beaware of extreme outliers or extreme skew which may nullify the robustness for $n \approx 30$. Similar to the inference about proportions, you want that zero is in the interval and if not, then at least you can infer that one group has a higher mean than the other.
  In this case, we don't assume equal variance in the populations (homoskedasticity).
- *Significance tests are differences in means*: we claim that there is no difference; H_0: $\mu_1 = \mu_2$ or $\mu = \mu_2 - \mu_1 = 0 = \mu_0$.

$$t = \frac{(\mu_2 - \mu_1) - \mu_0}{se} = \frac{\mu_2 - \mu_1}{se}; \qquad se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; \qquad \mathrm{df} = n_1 + n_2 - 2.$$

§ **7.4** Comparing means with dependent samples

- Can be called matched pairs, longitudinal study (with the same subjects), before and after, repeated measures, crossover study, etc.
- The sampling distribution of the difference between two groups is Normally distributed with mean $\mu_d = \mu_2 - \mu_1$ and standard error s_d, which is the standard deviation of the difference between scores (pairwise).
- *Paired-difference t test*: we test H_0: $\mu_d = \mu_2 - \mu_1 = 0 = \mu_{d,0}$, where

$$(y_1, y_2) \quad \Longrightarrow y_d = y_2 - y_1$$
$$\Longrightarrow t = \frac{\bar{y}_d - \mu_{d,0}}{se_d} = \frac{\bar{y}_d}{s_d/\sqrt{n}}; \qquad se_d = \frac{s_d}{\sqrt{n}}.$$

  Also, the confidence interval is given by

$$\bar{y}_d \pm t^*_{1-\alpha/2,\ \mathrm{df}=n-1} \times se_d.$$

- The standard error for the difference in scores (dependent sample) is lower than the standard error for comparing two independent samples.

§ **7.5** Other methods for comparing means

- *Pooled standard deviation*: if we assume that variances of two independent samples are equal in the population (homoskedasticity), then we can pool their individual variances:

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}} \qquad \Longrightarrow se_p = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

  The confidence interval for $\mu_2 - \mu_1$ in this case is

$$\Longrightarrow (\bar{y}_2 - \bar{y}_1) \pm t^*_{1-\alpha/2,\ \mathrm{df}=n_1+n_2-2} \times se_p.$$

  For testing against the claim H_0: $\mu_1 = \mu_2$ or $\mu = \mu_2 - \mu_1 = 0 = \mu_0$,

$$\begin{aligned} t &= \frac{(\bar{y}_2 - \bar{y}_1) - \mu_0}{se_p} \\ &= \frac{\bar{y}_2 - \bar{y}_1}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \end{aligned}$$

- Read page 207 about randomised block design.
- The assumption of homoskedasticity $\left(\sigma^1 \approx \sigma^2\right)$ modifies your t-test through the standard error and degrees of freedom. In general, the degrees of freedom is bounded: $\min\{n_1 - 1, n_2 - 1\} \leq \mathrm{df} \leq n_1 + n_2 - 2$. The upper bound is when homoskedasticity is assumed, and the lower bound is a rough estimate of the solution to the Welch-Satterthwaite equation - used in Welch's t-test. The df is smaller and standard error is larger if you do not assume homoskedasticity. If the sample sizes are equal, then the standard errors are also equal despite the homoskedasticity assumption.
- Read page 209-210, up to § 7.6.

§ 8 Analyzing association between categorical variables

§ 8.1 Contingency tables

- You assume under H_0: the variables are independent or $\mathbb{P}(A|B) = \mathbb{P}(A)$, which is equivalent to $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$ using *Bayes' Rule for Independent Events*. For a 2×2 contingency table, e.g. Table 1, you calculate the expected frequencies as what you expect under H_0, so the expected frequency for the cell $A \cap C$ is $\mathbb{E}(A \cap C) = n \times \mathbb{P}(A) \times \mathbb{P}(C)$.

	A	B	Marginal
C	$\mathbb{P}(A \cap C)$	$\mathbb{P}(B \cap C)$	$\mathbb{P}(C)$
D	$\mathbb{P}(A \cap D)$	$\mathbb{P}(B \cap D)$	$\mathbb{P}(D)$
Marginal	$\mathbb{P}(A)$	$\mathbb{P}(B)$	$\mathbb{P}(\Omega) = 1$

Table 1

If you have the observed frequencies, e.g. Table 2, you can calculate the expected frequencies as $\mathbb{E}(A \cap C) = O_A \times O_C / n$.

	A	B	Marginal
C	$O_{A \cap C}$	$O_{B \cap C}$	O_C
D	$O_{A \cap D}$	$O_{B \cap D}$	O_D
Marginal	O_A	O_B	$O_\Omega = n$

Table 2

- *Joint distribution*: $\mathbb{P}(A \cap B)$.
- *Marginal distribution*: $\mathbb{P}(A)$.
- *Conditional distribution*: $\mathbb{P}(A|B)$.
- Two categorical variables are statistically independent if, in the population, the $\mathbb{P}(A|C) = \mathbb{P}(A|D) = \mathbb{P}(A)$ and $\mathbb{P}(B|C) = \mathbb{P}(B|D) = \mathbb{P}(B)$, with reference to Table 1. The variables are statistically dependent if they are not equal. Refer to table 8.3 on page 229.

§ 8.2 Chi-squared test of independence

- We claim H_0: the variables are statistically independent. If the expected cell frequencies f_E are at least 5 (in *every* cell), then we can use the Pearson χ^2 statistic:

$$\chi^2 = \sum \frac{(f_O - f_E)^2}{f_E} \sim \chi^2 (\mathrm{df} = (r-1)(c-1)).$$

Read the sub-section "THE CHI-SQUARED PROBABILITY DISTRIBUTION" from page 231 onward.

- If there is one cell which has expected cell frequency less than 5, then you need to use Fisher's Exact test.
- For 2×2 contingency tables, you can model this test of independence as a difference of proportions z-test (see table 8.6 on page 234).

$$z^2 = \frac{(\hat{\pi}_2 - \hat{\pi}_1)^2}{se_0^2} \sim \chi^2 (\mathrm{df} = 1).$$

Interesting to note: $\mathbb{P}\left(\chi_1^2 > 1.96^2\right) = \mathbb{P}\left(\chi_1^2 > 3.84\right) = 0.05$ and $\mathbb{P}(|z| > 1.96 = 0.05)$, so for $\mathrm{df} = 1$ the χ^2 distribution has the same critical values as the z-distribution.

Week 1

Power

It can be difficult to understand statistical power, but essentially it is how well your test works. You, as a psychologist/statistician, want to ensure that you only make **necessary** changes to society and do so at the right time. The impact of the changes you make, or the affect to society, can be immense which is why you might choose a low significance level, such as 0.05. This says that, if the data you collect is representative of the population (large n) and it tells a different story to what you already know (null hypothesis), then you should change your mind (reject H_0). What you assumed cannot be true, so you must believe the data (accept H_a). You are able to conclude that your assumption has been contradicted, because the sample representing the population is clustered around another number (the mean is different to what you assumed) and the potential for that to happen (probability p-value) is extremely low, and in fact is lower than what you prescribe as an acceptable margin for deviation within the population (α).

Consider the first homework question from week 1 (Figure 10), where H_0 is that $\mu_{\bar{X}} = \mu_{\bar{Y}}$ and H_a is that $\mu_{\bar{Y}} < \mu_{\bar{X}}$ (one-sided). We already found that the p-value for this hypothesis test was 0.0041, but what is the "cut-off score", i.e. from what value do we consider rejecting H_0?

$$z = \frac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_{\bar{X}} - \mu_{\bar{Y}}\right)}{\sigma/\sqrt{n}} > z^*_{1-\alpha} \tag{1}$$

For $\alpha = 0.05$, our one-sided critical z-value is 1.645, and under our null hypothesis $\mu_{\bar{X}} - \mu_{\bar{Y}} = 0$. After inputting σ and n into the equation, we yield:

$$\frac{\bar{X} - \bar{Y}}{9/\sqrt{23}} > 1.645 \tag{2}$$

$$\implies \bar{X} - \bar{Y} > 1.645 \times \frac{9}{\sqrt{23}} \approx 3.087. \tag{3}$$

This says that, if the difference between $\bar{X}$ and $\bar{Y}$ is more than 3, then we reject H_0. In our sample, the difference we found was 7, which is more than 3 so we rejected H_0. Suppose that a very large research institute was able to collate samples of the same test from all over the world, and concluded that the true mean of the population is in fact 5 (the experimental group performs worse, on average, by 5 points than the control group). How powerful was your test? If the test you conduct is "powerful" it means that you found a statistic which was relevant to the population, i.e. the sample mean will be calculated correctly (rejecting H_0) in a high percentage of times.

$$\begin{aligned}
\text{Power} &= 1 - \beta = \mathbb{P}\left(\text{reject } H_0 \,|H_0 \text{ is false}\right) \qquad (4)\\
&= \mathbb{P}\left(\bar{X} - \bar{Y} > 3.087 \,|\mu_{\bar{X}-\bar{Y}} = 5\right) \qquad (5)\\
&= \mathbb{P}\left(\frac{\left(\bar{X} - \bar{Y}\right) - 5}{9/\sqrt{23}} > \frac{3.087 - 5}{9/\sqrt{23}}\right) \qquad (6)\\
&= \mathbb{P}\left(Z > -1.02\right) \qquad (7)\\
&= 0.8459. \qquad (8)
\end{aligned}$$

So, given that we know the standard deviation in the population to be 9 and prescribed $\alpha = 0.05$, in samples of size 23 we will correctly reject the null hypothesis 84.6% of the time. This can be seen graphically in Figure 1: the graph of the distribution of $\bar{X} - \bar{Y}$ is translated from centering around the hypothesised mean of zero, to centering around the true mean. We have decided that if the sample statistic $\bar{x} - \bar{y}$ is outside of the 95% of cases, then it is significantly different to the hypothesised case - significant because it changes our mind.

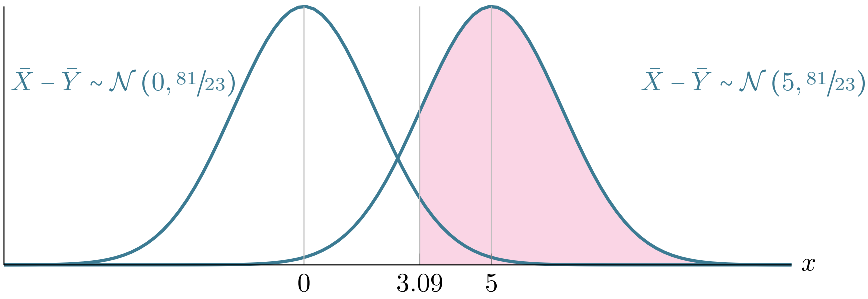


Figure 1: This graph shows the distribution of the difference in means from the experiment laid out in question 1 of homework 1. In a sample of size 23, we correctly assume that the experimental group has lower test scores than the control group if the difference in sample means is more than 3 at a 5% significance level. If the true population has a mean of 5 points, then this correct assumption will be reached in 84.6% of tests.

Notice that this value depends on three significant factors: the sample size n, the hypothesised μ_0 and actual mean μ, and the significance level α. What happens if we change the sample size? For a sample of size 23, the standard error of the sampling distribution of the mean is $9/\sqrt{23} \approx 1.8766$ which decreases as the sample size increases. For a sample of size 50, the standard error is $9/\sqrt{50} \approx 1.2728$ meaning that we reject H_0 when $\bar{X} - \bar{Y} > 1.645 \times 1.2728 \approx 2.094$ which is smaller than the cut-off value of 3.087 when our sample size was 23. So, as n increases, the cut-off value also decreases (due to the decrease in the SE) which can be seen in Figure 2.

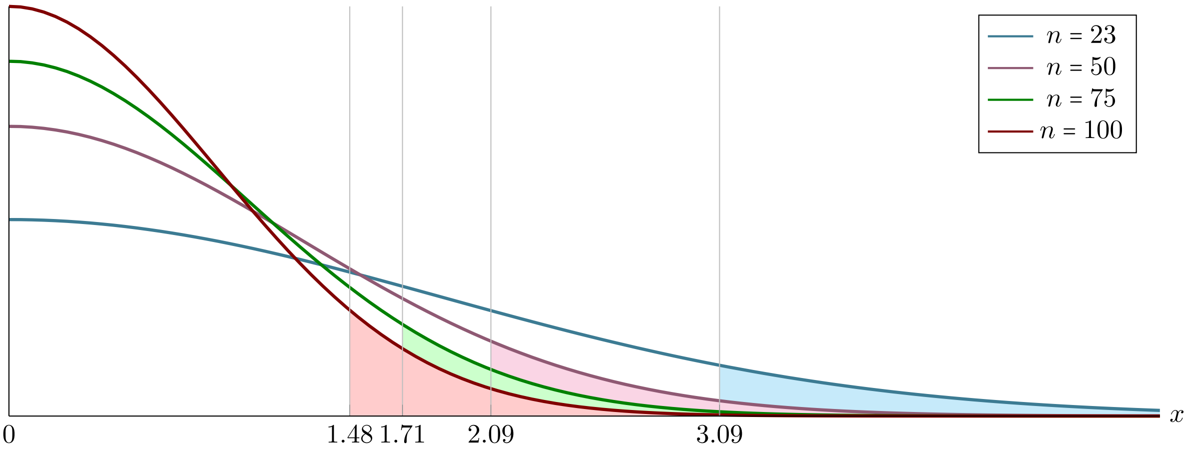


Figure 2: This graph shows the cut-off values of 3.09 ($n = 23$), 2.09 ($n = 50$), 1.71 ($n = 75$), and 1.48 ($n = 100$) for a fixed significance level ($z^*_{1-\alpha} = 1.645$; $\alpha = 0.05$) for a normally distributed random variable ($\bar{X} - \bar{Y}$) with mean zero and standard deviation $9/\sqrt{n}$. The shaded area under the graph represents the top 5% of each of the distributions (so the area is the same), however the value for which the top 5% is taken from is changing.

But how does increasing sample size impact our power? We have seen in Figure 2 that the cut-off value is decreasing which means that,

$$\text{Power} = 1 - \beta \tag{9}$$

$$= \mathbb{P}\left(\text{reject } H_0 \,|\, H_0 \text{ is false}\right) \tag{10}$$

Recall from above that the true population mean is 5, and not zero.

$$= \mathbb{P}\left(\bar{X} - \bar{Y} > \underbrace{1.645 \times \frac{9}{\sqrt{n}}}_{=14.805/\sqrt{n}} \,\middle|\, \mu_{\bar{X}-\bar{Y}} = 5\right) \tag{11}$$

$$= \mathbb{P}\left(\frac{(\bar{X} - \bar{Y}) - 5}{9/\sqrt{n}} > \frac{14.805/\sqrt{n} - 5}{9/\sqrt{n}}\right) \tag{12}$$

$$= \mathbb{P}\left(Z > 1.645 - \frac{5}{9\sqrt{n}}\right) \tag{13}$$

$$= \begin{cases} \mathbb{P}\left(Z > -2.28\right) \approx 0.9887, & n = 50; \\ \mathbb{P}\left(Z > -3.17\right) \approx 0.9992, & n = 75; \\ \mathbb{P}\left(Z > -3.91\right) \approx 1, & n = 100. \end{cases} \tag{14}$$

So, as n increases so does the power of our test, which is shown in Figure 3.

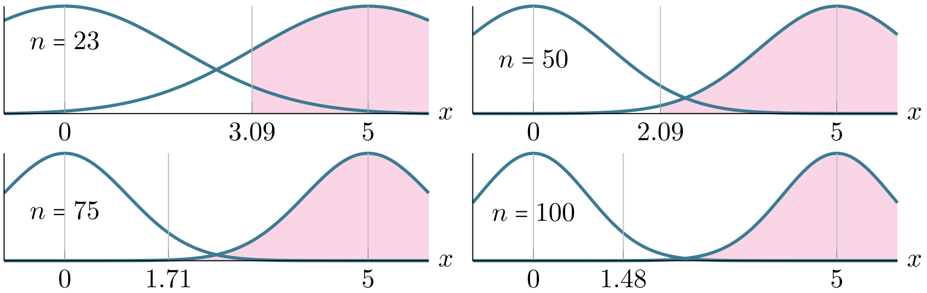


Figure 3: The shaded area represents the power of the test for increasing values of n, where the random variable $\bar{X} - \bar{Y}$ is distributed normally with standard error $9/\sqrt{n}$ ($\alpha = 0.05$; $H_0 : \mu = 0$; $H_a : \mu > 0$; true mean $\mu = 5$).

Confidence intervals

A confidence interval is the proposed range of a population parameter to a certain percentage of accuracy given a sample and associated statistic/s.

$$\text{confidence interval} = \text{estimate} \pm \text{margin of error}.$$

We construct confidence intervals to approximate the location of the true population parameter, and we use the parameter estimate from a sample to do so. In the case of a CI for the population mean μ, and we have the population standard deviation σ:

$$100 \times (1 - \alpha)\,\%\text{ CI for } \mu\text{:} \quad \bar{x} \pm z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}. \tag{15}$$

Where $\bar{x}$ is the mean of a sample taken from the population of size n. We suppose that, if a large number of samples are taken from the same population, with mean μ and standard deviation σ, we can deduce that the mean of every sample is distributed normally with the same mean but with a scaled standard deviation. We call this the **sampling distribution of the mean**, and it is written formally as $\bar{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$. So, what this confidence interval says is that, in $100 \times (1 - \alpha)\,\%$ of cases the population mean μ is equal to the estimate $\bar{x}$ with a margin of error equal to the standard normal deviation $z^*_{1-\alpha/2}$ for $100 \times (1 - \alpha)\,\%$ level multiplied by the standard error of the estimate $\sigma/\sqrt{n}$. Multiplying the critical z-value by the standard error transforms the margin of error from the standard normal to our distribution, so that the result is a margin of error for the sampling distribution of the mean.

Where does this come from? The Central Limit Theorem states that for a random variable X with standard deviation σ, 95% of sample means collected with size n will lie within 1.96 standard deviations of the mean:

$$\mathbb{P}\left(\mu - 1.96 \times \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) \approx 0.9500. \tag{16}$$

We can rearrange the above equation to solve for μ:

$$\implies \mathbb{P}\left(\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) \approx 0.9500. \tag{17}$$

So, in 95% of cases the population mean is within 1.96 standard deviations of the sample mean. There is a general formula for $100 \times (1 - \alpha)\,\%$:

$$\implies \mathbb{P}\left(\bar{X} - z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha. \tag{18}$$

The following figure represents the confidence interval graphically (for a particular distribution):

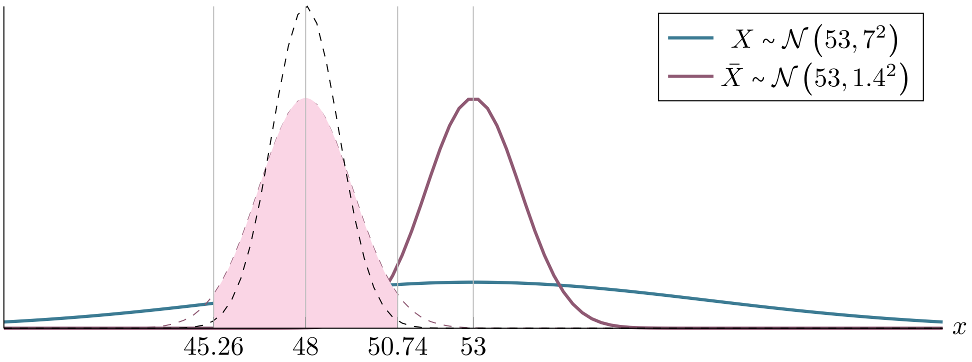


Figure 4: In question 1 of Homework 2, the population is distributed normally with mean 53 and standard deviation 7. A sample of 25 is taken, and a mean of 48 is computed from the sample which is useful in constructing a 95% confidence interval, where the margin of error is $1.96 \times 1.4 = 2.744$. The confidence interval is shaded pink, and the pink dash line is the sampling distribution of the mean centred at 48. The sampling distribution of the mean of X is distributed normally with a mean of 53 and standard deviation of $\sigma/\sqrt{n} = 7/5 = 1.4$ for a sample size of 25. The black dashed line represents the standard normal distribution centred at 48.

Essentially, you take a standard normal graph of $Z \sim \mathcal{N}(0,1)$ and centre it at your sample mean $\bar{x}$. Next, you pull up to narrow the shape of the bell curve if $\sigma/\sqrt{n} < 1$, or push down to widen the shape of the bell curve if $\sigma/\sqrt{n} > 1$. So, the values that normally represent the middle 95% of a standard normal distribution ($z = \pm 1.96$) have been moved so that the middle 95% of our sampling distribution of the mean (centred at $\bar{x}$) is covered.

In Figure 4, the hypothesised mean $\mu_0 = 53$ is not contained in the 95% CI centred at the sample mean of 48. This allows us to conclude that the true population mean is not 53, as we presume that $\bar{X}$ is an unbiased estimator of μ, and has a standard error of $\sigma/\sqrt{n}$ which reduces as the sample size increases. We hypothesised that the true mean was 53, but we allow that we can be wrong and in fact prescribe a level of significance α to define when we are wrong. So if the hypothesised mean is outside of the CI, we accept that the observed mean is a better estimate of μ than μ_0 is and deduce that μ is within this CI 95% of the time. **It is really important to remember that μ is fixed in the population, but $\bar{x}$ can change and what we are really saying here is that in 95% of samples the true mean lies within 1.96 standard deviations of $\bar{x}$.** The other 5% of samples have CI's which do not contain the true population parameter, however you have decided that 5% is not that bad and also is the highest percentage of inaccuracy that you will accept. If we ensured that our sample size n was at least as big as the variance σ^2 of the population of X, then $\bar{X}$ would estimate μ with a standard error no more than 1 ($n \geq \sigma^2 \implies \sigma/\sqrt{n} \leq 1$).

Variance and standard deviation

Consider a random variable X with mean μ. Imagine you plotted on a graph all of the values of this random variable, then you drew a square:

This square represents the squared distance from that value of X to μ. Next, you add up all of your squares and divide it by the total number of squares that you have (take the average of the squares). What results is the **population variance** σ^2 - it's the average squared distance between the values of X and the mean μ. If you don't want the squared distance and only want to know the absolute distance (shortest path from value X to μ), you can take the squared root of the variance and produce the **standard deviation** σ - the average straight-line distance between the values of X and the mean μ. It is quite impossible to do this for an entire population, as we assume that the population size is infinite and uncountable. We can estimate the population variance using a sample collected from the population. Instead of dividing by n (the total number of values of X present in the sample), we divide by $n-1$ as this ensures that it remains an unbiased estimator of σ^2.

$$\mathbb{E}\left(\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}\right) = \left(\frac{n-1}{n}\right) \times \sigma^2 \tag{19}$$

In order for an estimator to be unbiased, it's expected value must be equal to the parameter it estimates. In this case, it is not equal to σ^2.

$$\implies \mathbb{E}\left(\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}\right) \times \left(\frac{n}{n-1}\right) = \sigma^2. \tag{20}$$

We can rearrange so that

$$\mathbb{E}\left(s^2\right) = \mathbb{E}\left(\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}\right) = \sigma^2. \tag{21}$$

Beaware that the sample standard deviation s ($s = \sqrt{s^2}$) is **not** an unbiased estimator of the population standard deviation σ. This is because the square root function is **not** a linear function, and only linear functions are preserved under the expected value function:

$$\sqrt{\mathbb{E}\left(s^2\right)} \neq \mathbb{E}\left(\sqrt{s^2}\right). \tag{22}$$

Using $n - 1.5$ in place of $n - 1$ almost completely eliminates bias for estimating σ, and using n in place of $n - 1$ reduces the mean squared error. There are positive and negative effects associated with correcting for bias and we presume for the purposes of this course that s is the best estimator of σ.

It can sometimes be unclear which is the best choice, s or σ, to use when analysing data of a random variable X. Recall the formulas:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}; \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{n}}; \qquad \implies \text{SE} = \frac{\sigma}{\sqrt{n}} \tag{23}$$

$$= \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{n^2}} \tag{24}$$

$$= \frac{\sqrt{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}}{n}. \tag{25}$$

The difference between the formulas is the denominator, which corrects for bias. In a sample, the individual values might be closer to the sample mean than they are to the population mean, and $\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ might underestimate the actual sum of squares $\sum_{i=1}^{n}\left(x_i - \mu\right)^2$.

The standard error shows how close your sample mean is to the population mean. Sample standard deviation shows how much individuals within the same sample differ from the sample mean. This also means that standard error should decrease if the sample size increases, as the estimate of the population mean improves. Sample standard deviation will be unaffected by changes in the sample size[1].

Previously we spoke only about estimating σ for univariate data, and now we move on to bi- and multivariate data. If we collect two or more independent samples either from the same population or from different populations with (assumed) equal variance we can use pooled variance as an estimator for the population variance. If we cannot assume that the independent samples are taken from populations with equal or similar variances, then we use unpooled variance. Finally, if the two or more samples collected depend on each other in some manner, e.g. before and after therapy, then we are better off looking at the difference.

Pooled variance

Consider a test between I independent samples. Whilst we cannot assume that they are all from the same population (and hence have the same variance), the Central Limit Theorem allows us to conclude that the pooled variance converges to the true variance. To see how this works, and in order to better understand when and where to use pooled variance:

$$s_p^2 = \frac{\sum_{i=1}^{I}\left(n_i - 1\right)s_i^2}{\sum_{i=1}^{I}\left(n_i - 1\right)}, \tag{26}$$

where I is the total number of groups.

$$= \frac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2 + \cdots + \left(n_I - 1\right)s_I^2}{\underbrace{\left(n_1 - 1\right) + \left(n_2 - 1\right) + \cdots + \left(n_I - 1\right)}_{n-I}} \tag{27}$$

If we were to assume that one of the samples was a lot bigger than the others, suppose sample k ($n_k >> n_i$ for every sample), then we can assume that the pooled variance will converge to s_k^2:

$$= \frac{\frac{n_1-1}{n_k-1}s_1^2 + \cdots + s_k^2 + \cdots + \frac{n_I-1}{n_k-1}s_I^2}{\frac{n_1-1}{n_k-1} + \cdots + \underbrace{\frac{n_k-1}{n_k-1}}_{=1} + \cdots + \frac{n_I-1}{n_k-1}} \sim s_k^2. \tag{28}$$

[1] You can also peruse the examples given here: `https://statistics.laerd.com/statistical-guides/measures-of-spread-standard-deviation.php`.

The reason for this has something to do with the "power" of having a large n; that the sample is reliably similar to the population (lower standard error). This is an important point in sample collection, because if all samples except one have a really small size (¡30) but one sample is substantially larger (¿100), then the statistician can more readily believe the results of the largest sample (due to CLT - recall formula for SE). If we have that all of the samples are nearly the same size (choose $n_k \approx n_1 \approx \cdots \approx n_I$), then there is no dominating variance and we can assume the following:

$$\begin{aligned} s_p^2 &\approx \frac{(n_k - 1)\, s_1^2 + (n_k - 1)\, s_2^2 + \cdots + (n_k - 1)\, s_I^2}{(n_k - 1) + (n_k - 1) + \cdots + (n_k - 1)} && (29) \\ &= \frac{\cancel{(n_k - 1)} \times \left(s_1^2 + s_2^2 + \cdots + s_I^2\right)}{I \times \cancel{(n_k - 1)}} && (30) \\ &= \frac{\sum_{i=1}^{I} s_i^2}{I} && (31) \\ &= \bar{s}^2, && (32) \end{aligned}$$

which is the mean of the variances. This indicates that the pooled variances is the weighted average of variances, where the highest weight is given to the largest sample size. This is to ensure that our pooled variance is the best estimator of the population variance. Another way to look at this is to look at the fact that $s_i^2 = \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 / \left(n_i - 1\right)$, where j indexes the persons in a group.

$$\begin{aligned} \Longrightarrow\ s_p^2 &= \frac{\sum_{j=1}^{n_1} \left(y_{1j} - \bar{y}_1\right)^2 + \sum_{j=1}^{n_2} \left(y_{2j} - \bar{y}_2\right)^2 + \cdots + \sum_{j=1}^{n_I} \left(y_{Ij} - \bar{y}_I\right)^2}{n - I} && (33) \\ &= \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2}{n - I} && (34) \\ &= \frac{\text{SSE}}{\text{df}_E} = MSE. && (35) \end{aligned}$$

It is important to know when exactly to use pooled v.s. unpooled variance, but if we can **assume that the variances of the populations are equal**, i.e. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2$, then we use pooled variance.
For $I = 2$ (comparing two means), we can use the t statistic to determine the results of our test:

$$\begin{aligned} t &= \frac{\bar{y}_1 - \bar{y}_2}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t\left(\underbrace{n_1 + n_2}_{n} - 2\right) && (36) \\ \Longrightarrow\ t^2 &= \frac{\left(\bar{y}_1 - \bar{y}_2\right)^2}{s_p^2 \times \underbrace{\frac{n_1 + n_2}{n_1 n_2}}_{\frac{1}{n_1} + \frac{1}{n_2}}} && (37) \\ &= \frac{\left(\bar{y}_1 - \bar{y}_2\right)^2 / \left(n_1 + n_2\right)}{s_p^2 / n_1 n_2} && (38) \end{aligned}$$

If $n_1 \approx n_2$

$$\Longrightarrow\ t^2 \approx \frac{\frac{n}{2}\left(\bar{y}_1 - \bar{y}_2\right)^2}{s_p^2} \sim F\left(1, n_1 + n_2 - 2\right). \qquad (39)$$

How do we know if the populations have equal variance if we only have sample information? A general **rule of thumb** is that the smallest sample standard deviation must no less than half of the larger sample standard deviation:

$$\Longrightarrow\ s_1 \geq \frac{s_2}{2} > 0. \qquad (40)$$

We can rearrange this to consider only the ratio of standard deviations:

$$\Longrightarrow\ 2 \geq \frac{s_2}{s_1}. \qquad (41)$$

Further, we extend this to sample variance:

$$\Longrightarrow\ 4 \geq \frac{s_2^2}{s_1^2}. \qquad (42)$$

The more formal way of testing for equal variances is ANOVA, which is introduced in Stats 2.

Unpooled variance

If we cannot see from the data that the independent samples are drawn from the same population, or that the I variances are similar, then we use unpooled variance. In general, we do not consider using this for $I > 2$ (for comparing more than two means). It is mathematically possible, but in practice it is seldom used and certainly not a part of the scope of this course. The main reason being that the calculations required to determine the degrees of freedom is quite complicated (see (44)), and most psychologists use a computer. So, the **only time you use unpooled variance is when you have observable differences between the two groups**. So, you could note that $s_1^2 >> s_2^2$ or perhaps the n_i of each group is low (¡30) and unequal; it is something you need to determine for yourself. For example, if your test was about two machines from different manufacturers and whether they can complete the same task in the same amount of time, then you would use unpooled variance. If you were exploring behavioural data two culturally different countries, you would use unpooled variances. **<u>Rule of thumb:</u> if it's only two groups and you can assume/predict that the variances are unequal, use unpooled.**

$$s_{up} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{43}$$

$$\implies t = \frac{\bar{y}_1 - \bar{y}_2}{s_{up}} \sim t(k), \qquad \text{where } k = \frac{(n_1 - 1) \times (n_2 - 1)}{(n_2 - 1) \times C^2 + (1 - C)^2 \times (n_1 - 1)} \tag{44}$$

$$\text{and } C = \frac{s_1^2/n_1}{s_{up}^2}. \tag{45}$$

N.B.: if the sample sizes are equally large enough ($n_1 \approx n_2 = n$ and $n \geq 30$), then the formulae for standard error are equivalent, regardless of the homoskedasticity assumption.

$$s_{up} = \sqrt{\frac{s_1^2 + s_2^2}{n}} = \sqrt{\frac{s_1^2 + s_2^2}{2}} \times \sqrt{\frac{2}{n}} = s_p \sqrt{\frac{1}{n} + \frac{1}{n}}. \tag{46}$$

Paired data

I don't think I need to refer much to this, however if you are testing whether there is any effect before and then after, consider this to be paired data. In that case, you don't use either pooled or unpooled variance, but look instead at the variance of differences. That is, transform your data from x and y to $d = x - y$, for instance.

$$\implies \bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{\sum_{i=1}^n (x_i - y_i)}{n} = \frac{\sum_{i=1}^n x_i}{n} - \frac{\sum_{i=1}^n y_i}{n} = \bar{x} - \bar{y}. \tag{47}$$

$$s_d = \sqrt{\frac{\sum_{i=1}^n \left(d_i - \bar{d}\right)^2}{n}} \tag{48}$$

$$= \sqrt{\frac{\sum_{i=1}^n \left[(x_i - y_i) - (\bar{x} - \bar{y})\right]^2}{n - 1}} \tag{49}$$

The next few lines are just some boring math relations:

$$= \sqrt{\frac{\sum_{i=1}^n \left[(x_i - \bar{x}) - (y_i - \bar{y})\right]^2}{n - 1}} \tag{50}$$

$$= \sqrt{\frac{\sum_{i=1}^n \left[(x_i - \bar{x})^2 - 2(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2\right]}{n - 1}} \tag{51}$$

$$= \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} - 2\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} + \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}} \tag{52}$$

This is the important part:

$$= \sqrt{s_x^2 - 2\,\mathrm{Cov}(x, y) + s_y^2} \tag{53}$$

The $-2\,\mathrm{Cov}(x, y)$ is very important!! Similar to conducting a z-test when the variables are dependent, if we want σ_{X-Y} we calculate $\sqrt{\sigma_X^2 - 2\,\mathrm{Cov}(X, Y) + \sigma_Y^2}$. Here we use s^2 to approximate σ^2, which is unknown. The standard error of the estimate $\bar{d}$ approximating μ_d is $s_d/\sqrt{n}$.

Week 3

More on prioris

Prior to commencing a study, a psychologist/statistician needs to be aware that sample data can be misleading in that it is not always representative of the population. You should be familiar with Table 3 but you may not understand what it means graphically. It is important to remember that the probabilities α and β are conditional probabilities, in that they are conditional on which reality is actually true.

		Reality	
		H_0 is true	H_A is true
Decision	Reject H_0	**Type I error** α “false alarm”	True positive **Power** $1-\beta$ Sensitivity
	Accept H_0	True negative $1-\alpha$ Specificity	**Type II error** β “miss rate”

Table 3

In hypothesis testing, you assume that H_0 is the true reality, but you accept that you might be wrong so H_A is a possible reality. Under the null hypothesis, you assume (if you are taking about means) that $\mu = \mu_0$ where μ_0 is some value, say zero ($\mu_0 = 0$). You accept that there is an $100 \times \alpha\%$ possibility that you will have data which is not representative of the population, for whatever reason (usually attributed to human error in the data collection itself). This says that there is a $100 \times (1-\alpha)\%$ possibility that the sample you collected is representative of your population. If H_0 is the true reality it means that (in terms of $\mu = \mu_0$ and $\alpha = 0.05$) the probability that the sample mean is within 1.96 standard deviations (or lies 1.645 standard deviations away, if one-tailed) from the population mean, given that the population mean is equal to μ_0, is 0.95.
In the case that you were wrong, so H_0 is not the true reality and H_A is, you accept that this is probable with $\alpha\%$, and in fact can calculate **when** you would discover this via your sample.

Two-sided H_A: $\mu \neq \mu_0$, with $\alpha = 0.05$

$$\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} \geq 1.96 \tag{54}$$

$$|\bar{x}-\mu_0| \geq 1.96 \times \frac{\sigma}{\sqrt{n}} \tag{55}$$

$$\bar{x} \geq \mu_0 + 1.96 \times \frac{\sigma}{\sqrt{n}} \tag{56}$$

$$\bar{x} \leq \mu_0 - 1.96 \times \frac{\sigma}{\sqrt{n}} \tag{57}$$

If we have that the sample mean is more than some value equal to $\mu_0 + 1.96 \times \sigma/\sqrt{n}$, or less than some value equal to $\mu_0 - 1.96 \times \sigma/\sqrt{n}$, then our p-value is less than α (refer to Figure 5a).

$$\mathbb{P}\left(\bar{X} \geq \mu_0 + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) \approx 0.025 \tag{58}$$

$$\mathbb{P}\left(\bar{X} \leq \mu_0 - 1.96 \times \frac{\sigma}{\sqrt{n}}\right) \approx 0.025 \tag{59}$$

One-sided H_A: $\mu > \mu_0$ (right-tailed), with $\alpha = 0.05$

$$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \geq 1.645 \tag{60}$$

$$\bar{x}-\mu_0 \geq 1.645 \times \frac{\sigma}{\sqrt{n}} \tag{61}$$

$$\bar{x} \geq \mu_0 + 1.645 \times \frac{\sigma}{\sqrt{n}} \tag{62}$$

If we have that the sample mean is more than some value equal to $\mu_0 + 1.645 \times \sigma/\sqrt{n}$ then our p-value is less than α (refer to Figure 5b).

$$\mathbb{P}\left(\bar{X} \geq \mu_0 + 1.645 \times \frac{\sigma}{\sqrt{n}}\right) \approx 0.05 \tag{63}$$

One-sided H_A: $\mu < \mu_0$ (left-tailed), with $\alpha = 0.05$

$$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \leq -1.645 \tag{64}$$

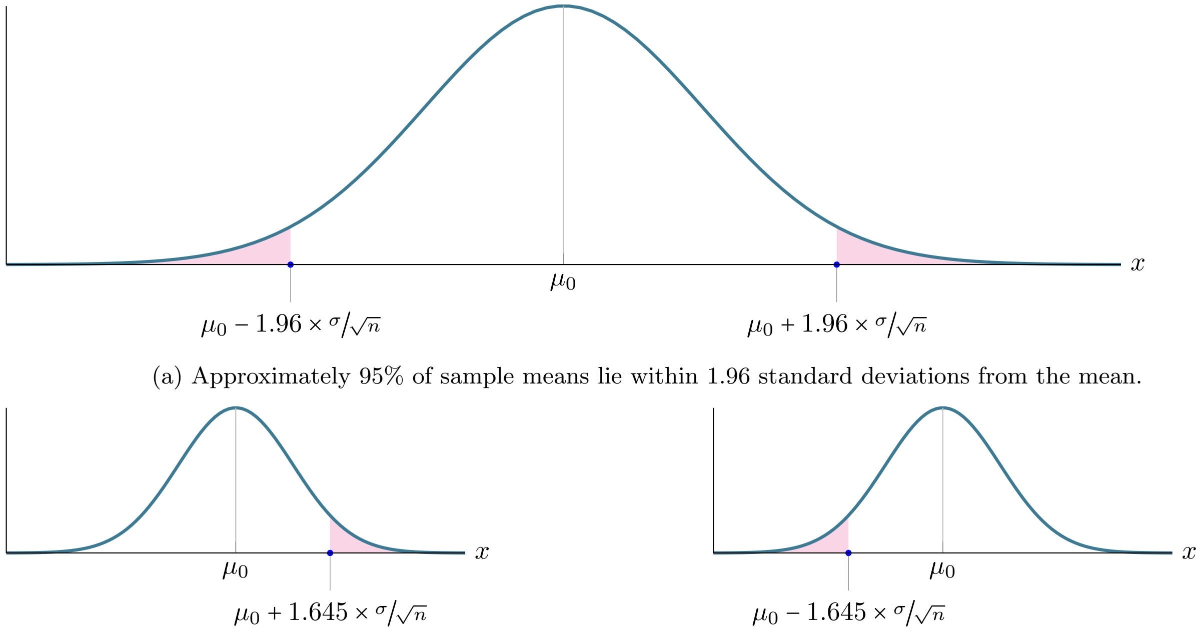
$$\bar{x} - \mu_0 \leq -1.645 \times \frac{\sigma}{\sqrt{n}} \tag{65}$$

$$\bar{x} \leq \mu_0 - 1.645 \times \frac{\sigma}{\sqrt{n}} \tag{66}$$

If we have that the sample mean is less than some value equal to $\mu_0 - 1.645 \times {}^{\sigma}/_{\sqrt{n}}$ then our p-value is less than α (refer to Figure 5c).

$$\mathbb{P}\left(\bar{X} \leq \mu_0 - 1.645 \times \frac{\sigma}{\sqrt{n}}\right) \approx 0.05 \tag{67}$$

What happens if we modify our acceptance level, α? If we *decrease* α, the shaded areas in Figure 5 will become smaller, which means that our "cut-off values" will change to values which are further from μ_0. A decrease in α will result in an increase of β in samples of unchanged size (Figure 7) - *reduction of power.* If we *increase* our α, the shaded areas in Figure 5 will become larger, which means that our "cut-off values" will change to values which are closer to μ_0. An increase in α will result in an decrease of β in samples of unchanged size (Figure 7) - *increase of power.*



(a) Approximately 95% of sample means lie within 1.96 standard deviations from the mean.

(b) Approximately 95% of sample means lie below 1.645 standard deviations above the mean.

(c) Approximately 95% of sample means lie above 1.645 standard deviations below the mean.

Figure 5: For a given random variable X which has mean μ_0 and standard deviation σ, the sampling distribution of the mean of size n is Normal with mean μ_0 and standard deviation $\sigma/\sqrt{n}$ (the Standard Error) which is denoted $\bar{X} \sim \mathcal{N}\left(\mu_0, \sigma^2/n\right)$.

Now that we have determined the "cut-off values" (when we reject H_0, based on μ_0, n and α), if we were to learn the **true value of μ in the population**, let's call it μ_A, it is possible to compute the power of our test, i.e. the probability of rejection given that H_0 **should** be rejected.

Two-sided H_A: $\mu \neq \mu_0$,
with $\alpha = 0.05$

$$\text{power} = \mathbb{P}\left(\text{reject } H_0 \,|\, H_0 \text{ is not the true reality}\right) \tag{68}$$

We define two scenarios, where the true mean μ_A is greater than the hypothesised mean μ_0 and when it is less than.

$$= \begin{cases} \mathbb{P}\left(\bar{X} \geq \mu_0 + 1.96 \times \dfrac{\sigma}{\sqrt{n}} \middle| \mu = \mu_A\right), & \mu_A > \mu_0 \\ \mathbb{P}\left(\bar{X} \leq \mu_0 - 1.96 \times \dfrac{\sigma}{\sqrt{n}} \middle| \mu = \mu_A\right), & \mu_A < \mu_0 \end{cases} \tag{69}$$

Now convert to the standard normal distribution and get your z-scores:

$$= \begin{cases} \mathbb{P}\left(\dfrac{\bar{X} - \mu_A}{\sigma/\sqrt{n}} \geq \dfrac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + 1.96\right), & \mu_A > \mu_0 \\ \mathbb{P}\left(\dfrac{\bar{X} - \mu_A}{\sigma/\sqrt{n}} \leq \dfrac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - 1.96\right), & \mu_A < \mu_0 \end{cases} \tag{70}$$

$$= \begin{cases} \mathbb{P}\left(Z \geq \dfrac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + 1.96\right), & \mu_A > \mu_0 \\ \mathbb{P}\left(Z \leq \dfrac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - 1.96\right), & \mu_A < \mu_0 \end{cases} \tag{71}$$

In two-sided hypothesis testing, we can use the "cut-off values" to calculate power using the z-test. So, if $\mu_A > \mu_0$, we transform the "cut-off value" $\bar{x} = \mu_0 + 1.96 \times \sigma/\sqrt{n}$ into a z-score using μ_A **which should be negative**. The reason it **should be negative** is that will imply that our power is more than 50%; if you achieve a positive z-score with the $\geq$ sign, then you will certainly have low statistical power. We can transform this negative z-score with the $\geq$ sign into a positive z-score with the $\leq$ sign as follows:

$$\mathbb{P}\left(Z \geq \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + 1.96\right) = \mathbb{P}\left(Z \leq \frac{\mu_A - \mu_0}{\sigma/\sqrt{n}} - 1.96\right), \qquad \mu_A > \mu_0. \tag{72}$$

We are able to do this because the standard Normal distribution is symmetric about the mean of zero. It is now possible to read this value from the "regular" z-table, which usually shows $\leq$ and positive z-values. If $\mu_A < \mu_0$, then the transformed "cut-off" value that we calculated is already positive and we already have the $\leq$ sign, so it is easy to read from the table.

One-sided H_A: $\mu > \mu_0$ (right-tailed), with $\alpha = 0.05$

$$\text{power} = \mathbb{P}\left(\text{reject } H_0 \,|H_0 \text{ is not the true reality}\right) \tag{73}$$

$$= \mathbb{P}\left(\bar{X} \geq \mu_0 + 1.645 \times \frac{\sigma}{\sqrt{n}} \,\middle|\, \mu = \mu_A\right) \tag{74}$$

$$= \mathbb{P}\left(\frac{\bar{X} - \mu_A}{\sigma/\sqrt{n}} \geq \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + 1.645\right) \tag{75}$$

$$= \mathbb{P}\left(Z \geq \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + 1.645\right) \tag{76}$$

Again, we can transform this negative z-score (**should be negative** because $\mu_0 < \mu_A$ and **power should be more than 50%**) with $\geq$ sign to a positive z-score with $\leq$ sign:

$$= \mathbb{P}\left(Z \leq \frac{\mu_A - \mu_0}{\sigma/\sqrt{n}} - 1.645\right) \tag{77}$$

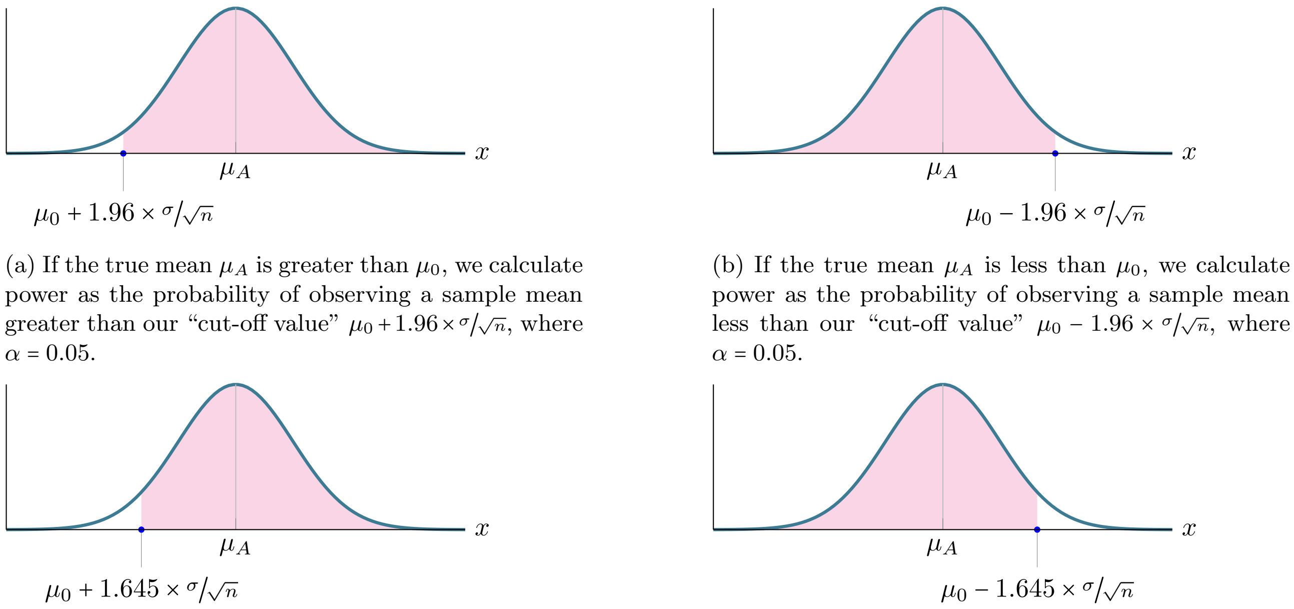
One-sided H_A: $\mu < \mu_0$ (left-tailed), with $\alpha = 0.05$

$$\text{power} = \mathbb{P}\left(\text{reject } H_0 \,|H_0 \text{ is not the true reality}\right) \tag{78}$$

$$= \mathbb{P}\left(\bar{X} \leq \mu_0 - 1.645 \times \frac{\sigma}{\sqrt{n}} \,\middle|\, \mu = \mu_A\right) \tag{79}$$

$$= \mathbb{P}\left(\frac{\bar{X} - \mu_A}{\sigma/\sqrt{n}} \leq \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - 1.645\right) \tag{80}$$

$$= \mathbb{P}\left(Z \leq \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - 1.645\right) \tag{81}$$



(a) If the true mean μ_A is greater than μ_0, we calculate power as the probability of observing a sample mean greater than our "cut-off value" $\mu_0 + 1.96 \times {}^{\sigma}\!/\!\sqrt{n}$, where $\alpha = 0.05$.

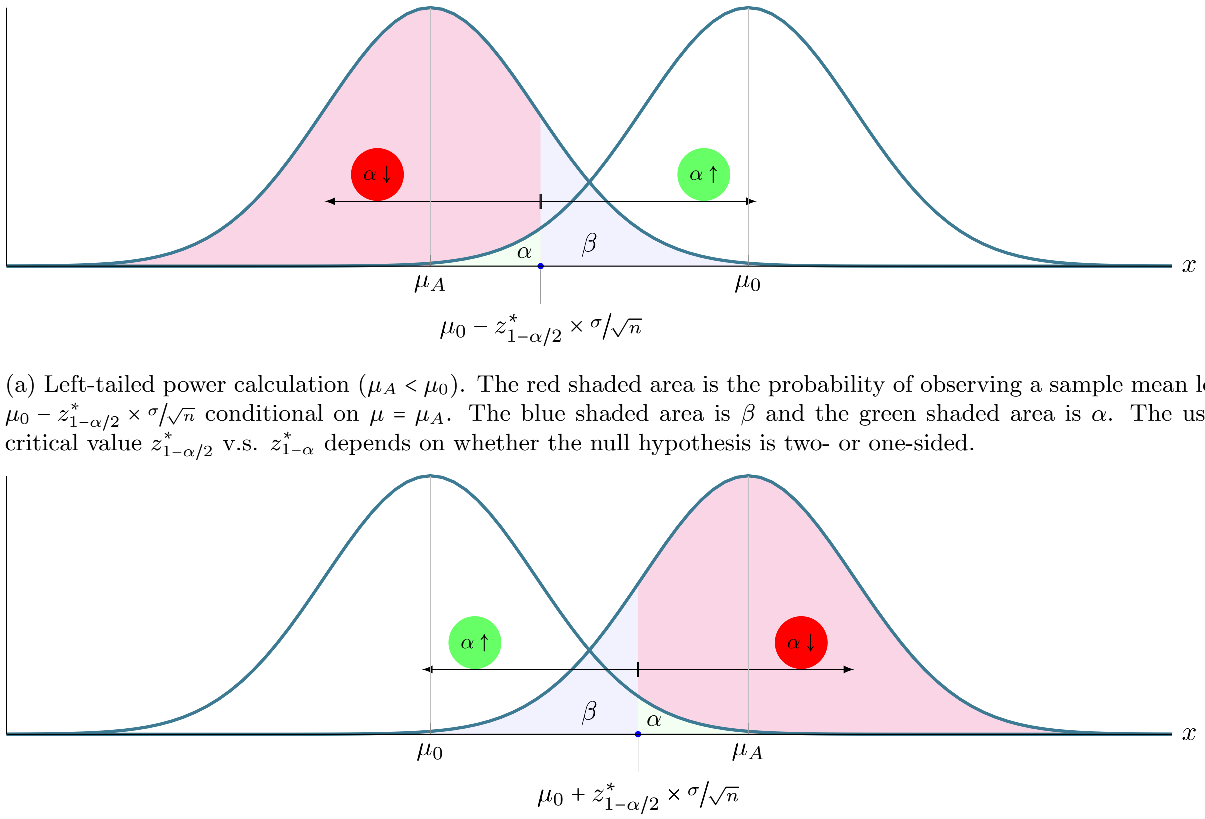
(b) If the true mean μ_A is less than μ_0, we calculate power as the probability of observing a sample mean less than our "cut-off value" $\mu_0 - 1.96 \times {}^{\sigma}\!/\!\sqrt{n}$, where $\alpha = 0.05$.

(c) For a right-tailed hypothesis test, we calculate statistical power as the probability of observing a sample mean greater than our "cut-off value" $\mu_0 + 1.645 \times {}^{\sigma}\!/\!\sqrt{n}$, where $\alpha = 0.05$.

(d) For a left-tailed hypothesis test, we calculate statistical power as the probability of observing a sample mean less than our "cut-off value" $\mu_0 - 1.645 \times {}^{\sigma}\!/\!\sqrt{n}$, where $\alpha = 0.05$.

Figure 6: When the true mean of the population is $\mu = \mu_A$, unequal to μ_0, we can calculate the statistical power based on the "cut-off values" proposed by μ_0 given in Figure 5.

So, if we are to finalise all of these ideas together (visually):



α↓
α↑
α
β
x
μ_A
μ_0
$\mu_0 - z^*_{1-\alpha/2} \times \sigma/\sqrt{n}$

(a) Left-tailed power calculation ($\mu_A < \mu_0$). The red shaded area is the probability of observing a sample mean less than $\mu_0 - z^*_{1-\alpha/2} \times \sigma/\sqrt{n}$ conditional on $\mu = \mu_A$. The blue shaded area is β and the green shaded area is α. The use of the critical value $z^*_{1-\alpha/2}$ v.s. $z^*_{1-\alpha}$ depends on whether the null hypothesis is two- or one-sided.

α↑
α↓
β
α
x
μ_0
μ_A
$\mu_0 + z^*_{1-\alpha/2} \times \sigma/\sqrt{n}$

(b) Right-tailed power calculation ($\mu_A > \mu_0$). The red shaded area is the probability of observing a sample mean more than $\mu_0 + z^*_{1-\alpha/2} \times \sigma/\sqrt{n}$ conditional on $\mu = \mu_A$. The blue shaded area is β and the green shaded area is α. The use of the critical value $z^*_{1-\alpha/2}$ v.s. $z^*_{1-\alpha}$ depends on whether the null hypothesis is two- or one-sided.

Figure 7: Power calculation for $\mu_A < \mu_0$ and $\mu_A > \mu_0$ when the population variance is known (z-test). If α is decreased, then the critical value $z^*_{1-\alpha/2}$ (or $z^*_{1-\alpha}$, if H_0 is one-sided) becomes larger which is visually shown by the red shaded area becoming smaller (smaller power). If α is increased, then the critical value $z^*_{1-\alpha/2}$ (or $z^*_{1-\alpha}$, if H_0 is one-sided) becomes smaller which is visually shown by the red shaded area becoming larger (larger power).

Week 4

t-test

Figure 7 shows the power calculation for the z-test for the mean, as we have assumed that the population variance of our random variable is known. If we do not know the population variance, we look at the t-test which requires the calculation of degrees of freedom. Notice in Figure 8 that the critical values, and thus the "cut-off" values, change with the degrees of freedom.

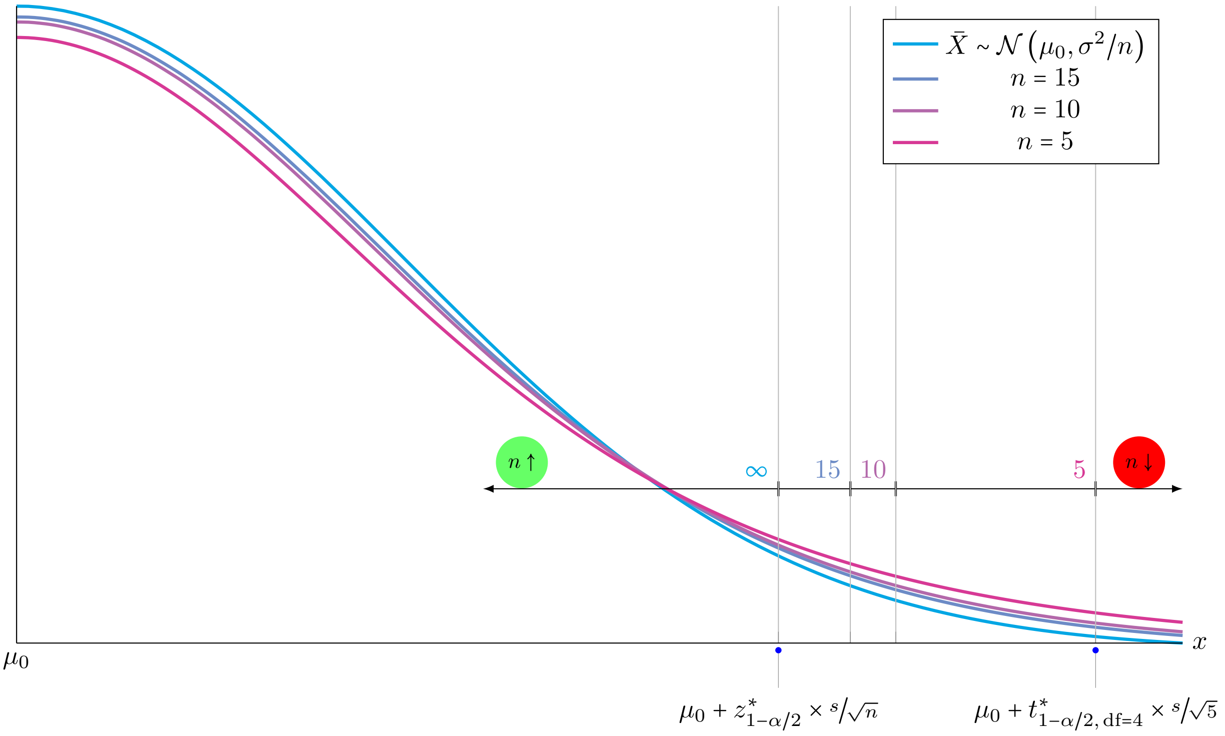


Figure 8: The right-tailed "cut-off values" for a t-test on the mean, where μ_0 is the mean under the (two-sided) null hypothesis and $s/\sqrt{n}$ is the Standard Error. The tails are "fatter" when n is small, and reshape to become "thinner" until it resembles that of the normally distributed variable with mean μ_0 and standard error $\sigma/\sqrt{n}$

There are only 4 different t-tests you need to remember:

1: one-sample or paired t-test If you are given a univariate data set, e.g. height of persons aged 21 in Groningen denoted as variable X, you can estimate the population variance using the sample variance and perform hypothesis testing using the following:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}, \qquad \text{for tests about the mean.} \tag{82}$$

Refer to (23) for the formula for s.

2: one-sample or paired t-test If you are given a (dependent) paired data set, e.g. before and after a therapy denoted variable D, you can estimate the population variance using the sample variance and perform hypothesis testing using the following:

$$t = \frac{\bar{d} - \mu_0}{s_d/\sqrt{n}} \sim t_{n-1}, \qquad \text{for tests about the mean difference.} \tag{83}$$

Refer to (47) and (48) for the formulas for $\bar{d}$ and s_d.
If you calculate $r(x, y) \approx 0$, or you are told/can assume that they are independent, then you need to decide if the homoskedasticity assumption holds.

3: two-sample t-test If you are given an independent bi-variate data set, (X, Y), you need to consider two possibilities:

a: homoskedasticity met Then you can estimate the population variance using the sample variance and perform hypothesis testing using the following:

$$t = \frac{\bar{x} - \bar{y} - \mu_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}, \qquad \text{for tests about the difference of means (homoskedasticity met).} \tag{84}$$

Refer to (26) for the formula for s_p.

b: homoskedasticity violated Then you can estimate the population variance using the sample variance and perform hypothesis testing using the following:

$$t = \frac{\bar{x} - \bar{y} - \mu_0}{s_{up}} \sim t_k, \qquad \text{for tests about the difference of means (homoskedasticity violated).} \tag{85}$$

Refer to (43) and (44) for the formulas for s_{up} and the degrees of freedom k.

$$\min\{n_1 - 1, n_2 - 1\} \le k \le n_1 + n_2 - 2 \tag{86}$$

Keeping the above relation in mind, you can also be lazy and just use the lower bound for k in place of calculating it using a computer.

Remember: if the samples are of equal size $n_1 = n_2$ then the standard error is the same whether homoskedasticity is met or violated (46), the only thing left to consider is the degrees of freedom - lower for homoskedasticity violated than met.

The main difference between the t-distribution and the normal distribution is the effect of degrees of freedom on the shape of the distribution (flat with fat tails for small df - similar to standard normal for large df).

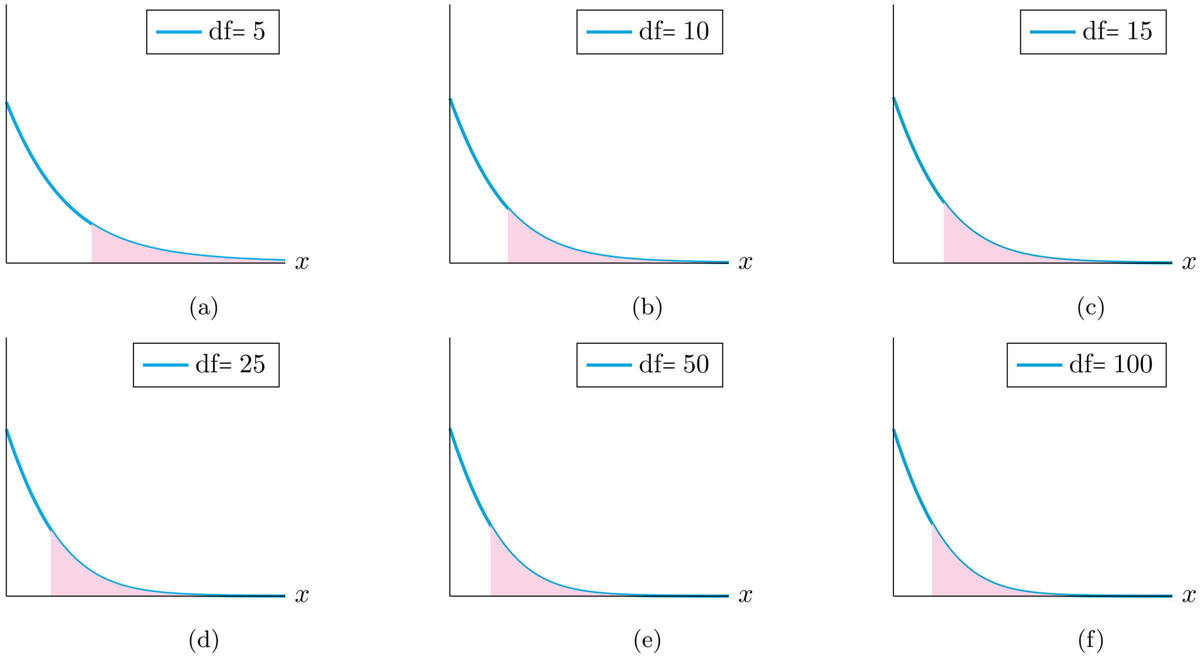


Figure 9: As the degrees of freedom increases, you can see that the shaded red area moves to the left so it appears that the shaded area is increasing. In actuality, the shaded red area is the same (representing 2.5%) and it appears to be changing due to changing graph shape. For small degrees of freedom, the shape of the bell curve is flat with long fat tails, and as df increases the bell curve reforms to look like that of the standard normal distribution.

Math expressions

Reading equations and understanding mathematical operations can be a little confusing, especially when you're not so sure about the consequences of making an incorrect inference. Firstly, fractions: what to do when you have a fraction on the denominator? When you divide by a whole number you can easily imagine this as sharing a cake with friends, e.g. divide by 5 can be pictured as taking the numerator (the cake) and cutting it into 5 pieces. When you divide by a fraction, you can't really imagine it so well as it doesn't make sense in terms of the cake analogy. If you convert the fraction in the denominator into a decimal number it can be a little more easy to understand: take the following fraction

$$\frac{1}{{}^{1}\!/_{2}} = \frac{1}{0.5} \tag{87}$$

If you take a whole cake and divide it by a half, you are actually dividing it by 0.5. What you are asking is, "how many 0.5's can you fit into 1" or "how many half cakes make up a whole cake"? With this thought in mind, we can easily conclude that 2 halves make a whole, so the answer is 2. The easiest way to do this is by the 'ole flip and multiply:

$$\frac{1}{{}^{1}\!/_{2}} = \frac{1}{1} \times \frac{2}{1} = 2. \tag{88}$$

This method is better than thinking always in cakes, as you may not always have a nice fraction with integers to work with. Consider the following fraction:

$$\frac{8}{{}^{5}\!/_{6}} = \frac{8}{0.8\bar{3}} \tag{89}$$

In terms of cakes, you are asking "how many $0.8\bar{3}$'s of a cake fit into 8 cakes"? You can easily think that there are at least 8 $0.8\bar{3}$'s of a cake in 8 cakes as $0.8\bar{3}$ is less than 1. So, now you have left over $0.1\bar{6}$ from each of the 8 cakes, which equates to $8 \times (1/6) = 8/6$. We can write this fraction as $(3+5)/6 = (3/6) + (5/6) = 0.5 + 0.8\bar{3}$, meaning that we can in total fit 9 $0.8\bar{3}$'s of a cake in 8 cakes and we will have half of a cake left over. Now comes the tricky part, what proportion of $0.8\bar{3}$ of a cake makes up half a cake. This is why we only visualise fractions in terms of cakes when the fractions are "nice". The easier way is to flip and multiply:

$$\frac{8}{{}^{5}\!/_{6}} = \frac{8}{1} \times \frac{6}{5} = \frac{48}{5} = 9 + \frac{3}{5} = 9.6. \tag{90}$$

What are you actually doing here? You don't want the fraction in the denominator, so you **multiply it out**:

$$\frac{8}{{}^{5}\!/_{6}} = x \implies ({}^{5}\!/_{6}) \times \left(\frac{8}{{}^{5}\!/_{6}}\right) = ({}^{5}\!/_{6}) \times (x) \tag{91}$$

We don't know what $\frac{8}{{}^{5}\!/_{6}}$ is yet, so we set it equal to unknown x. The next step (seen above) is to multiply both sides by the denominator ${}^{5}\!/_{6}$, which then requires that we solve for x and you will see that we get the same answer.

$$\implies 8 = \frac{5x}{6} \tag{92}$$

$$\implies 8 \times 6 = 48 = 5x \tag{93}$$

$$\implies \frac{48}{5} = 9.6 = x. \tag{94}$$

It doesn't really matter which of the three methods you choose to use: cakes, flip-a-roo, or multiplying it out. The important thing to remember is make it easiest for you; I would recommend the multiplying out method as it will make sure you don't make any errors. Thus far we have dealt with fractions in an arbitrary sense, however let's apply it to something which is more relevant. In the previous sections, we saw that we can calculate power if we know μ_A and you have heard that you can use effect size to calculate power.

$$\text{power} = \begin{cases} \mathbb{P}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha/2} \,|\mu = \mu_A\right), & H_A:\ \mu \neq \mu_0 \text{ and } \mu_A > \mu_0; \\ \mathbb{P}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha/2} \,|\mu = \mu_A\right), & H_A:\ \mu \neq \mu_0 \text{ and } \mu_0 > \mu_A; \\ \mathbb{P}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha} \,|\mu = \mu_A\right), & H_A:\ \mu > \mu_0; \\ \mathbb{P}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha} \,|\mu = \mu_A\right), & H_A:\ \mu > \mu_0. \end{cases} \tag{95}$$

$$= \begin{cases} \mathbb{P}\left(Z > \dfrac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right), & H_A:\ \mu \neq \mu_0 \text{ and } \mu_A > \mu_0; \\ \mathbb{P}\left(Z < \dfrac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right), & H_A:\ \mu \neq \mu_0 \text{ and } \mu_0 > \mu_A; \\ \mathbb{P}\left(Z > \dfrac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + z_{1-\alpha}\right), & H_A:\ \mu > \mu_0; \\ \mathbb{P}\left(Z < \dfrac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - z_{1-\alpha}\right), & H_A:\ \mu > \mu_0. \end{cases} \tag{96}$$

The above four cases are the usual ones that you will come across in this course, as you are mostly concerned about means. Suppose you want a specific power for your test, so you are looking for a particular critical z-value corresponding to

$$= 1 - \beta = \mathbb{P}\left(Z < z_{1-\beta}\right) \tag{97}$$

For instance, if you require 80% power, then $z_{1-\beta} = 0.85$ and so you need to solve $0.85 > \frac{\mu_0-\mu_A}{\sigma/\sqrt{n}} + z_{1-\alpha/2}$, for $H_A:\ \mu \neq \mu_0$ and $\mu_A > \mu_0$. You need to solve the following equations for n:

$$z_{1-\beta} > \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + z_{1-\alpha/2}, \qquad H_A:\ \mu \neq \mu_0 \text{ and } \mu_A > \mu_0; \tag{98}$$

$$z_{1-\beta} < \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - z_{1-\alpha/2}, \qquad H_A:\ \mu \neq \mu_0 \text{ and } \mu_0 > \mu_A; \tag{99}$$

$$z_{1-\beta} > \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + z_{1-\alpha}, \qquad H_A:\ \mu > \mu_0; \tag{100}$$

$$z_{1-\beta} < \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - z_{1-\alpha}, \qquad H_A:\ \mu > \mu_0. \tag{101}$$

We go about solving these equations in the following way: first subtract the $z(\alpha)$ critical values from both sides.

$$\left(z_{1-\beta}\right) - \left(z_{1-\alpha/2}\right) > \left(\frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right) - \left(z_{1-\alpha/2}\right), \qquad H_A:\ \mu \neq \mu_0 \text{ and } \mu_A > \mu_0; \tag{102}$$

$$\left(z_{1-\beta}\right) - \left(-z_{1-\alpha/2}\right) < \left(\frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right) - \left(-z_{1-\alpha/2}\right), \qquad H_A:\ \mu \neq \mu_0 \text{ and } \mu_0 > \mu_A; \tag{103}$$

$$\left(z_{1-\beta}\right) - \left(z_{1-\alpha}\right) > \left(\frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) - \left(z_{1-\alpha}\right), \qquad H_A:\ \mu > \mu_0; \tag{104}$$

$$\left(z_{1-\beta}\right) - \left(-z_{1-\alpha}\right) < \left(\frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - z_{1-\alpha}\right) - \left(-z_{1-\alpha}\right), \qquad H_A:\ \mu > \mu_0. \tag{105}$$

Next, we multiply out the fraction in the denominator.

$$\left(\frac{\sigma}{\sqrt{n}}\right) \times \left(z_{1-\beta} - z_{1-\alpha/2}\right) > \left(\frac{\sigma}{\sqrt{n}}\right) \times \left(\frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}}\right), \qquad H_A:\ \mu \neq \mu_0 \text{ and } \mu_A > \mu_0; \tag{106}$$

$$\left(\frac{\sigma}{\sqrt{n}}\right) \times \left(z_{1-\beta} + z_{1-\alpha/2}\right) < \left(\frac{\sigma}{\sqrt{n}}\right) \times \left(\frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}}\right), \qquad H_A:\ \mu \neq \mu_0 \text{ and } \mu_0 > \mu_A; \tag{107}$$

$$\left(\frac{\sigma}{\sqrt{n}}\right) \times \left(z_{1-\beta} - z_{1-\alpha}\right) > \left(\frac{\sigma}{\sqrt{n}}\right) \times \left(\frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}}\right), \qquad H_A:\ \mu > \mu_0; \tag{108}$$

$$\left(\frac{\sigma}{\sqrt{n}}\right) \times \left(z_{1-\beta} + z_{1-\alpha}\right) < \left(\frac{\sigma}{\sqrt{n}}\right) \times \left(\frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}}\right), \qquad H_A:\ \mu > \mu_0. \tag{109}$$

Now we multiply both sides by $\sqrt{n}$

$$\left(\sqrt{n}\right) \times \left(\frac{\sigma\left(z_{1-\beta} - z_{1-\alpha/2}\right)}{\sqrt{n}}\right) > \left(\sqrt{n}\right) \times \left(\mu_0 - \mu_A\right), \qquad H_A:\ \mu \neq \mu_0 \text{ and } \mu_A > \mu_0; \tag{110}$$

$$\left(\sqrt{n}\right) \times \left(\frac{\sigma\left(z_{1-\beta} + z_{1-\alpha/2}\right)}{\sqrt{n}}\right) < \left(\sqrt{n}\right) \times \left(\mu_0 - \mu_A\right), \qquad H_A:\ \mu \neq \mu_0 \text{ and } \mu_0 > \mu_A; \tag{111}$$

$$\left(\sqrt{n}\right) \times \left(\frac{\sigma\left(z_{1-\beta} - z_{1-\alpha}\right)}{\sqrt{n}}\right) > \left(\sqrt{n}\right) \times \left(\mu_0 - \mu_A\right), \qquad H_A:\ \mu > \mu_0; \tag{112}$$

$$\left(\sqrt{n}\right) \times \left(\frac{\sigma\left(z_{1-\beta} + z_{1-\alpha}\right)}{\sqrt{n}}\right) < \left(\sqrt{n}\right) \times \left(\mu_0 - \mu_A\right), \qquad H_A:\ \mu > \mu_0. \tag{113}$$

Finally, we divide by $\mu_0 - \mu_A$:

$$\begin{aligned}
\left(\frac{1}{\mu_0 - \mu_A}\right) \times \left(\sigma\left(z_{1-\beta} - z_{1-\alpha/2}\right)\right) &> \left(\frac{1}{\mu_0 - \mu_A}\right) \times \left(\sqrt{n}\left(\mu_0 - \mu_A\right)\right), && H_A:\ \mu \neq \mu_0 \text{ and } \mu_A > \mu_0; && (114)\\
\left(\frac{1}{\mu_0 - \mu_A}\right) \times \left(\sigma\left(z_{1-\beta} + z_{1-\alpha/2}\right)\right) &< \left(\frac{1}{\mu_0 - \mu_A}\right) \times \left(\sqrt{n}\left(\mu_0 - \mu_A\right)\right), && H_A:\ \mu \neq \mu_0 \text{ and } \mu_0 > \mu_A; && (115)\\
\left(\frac{1}{\mu_0 - \mu_A}\right) \times \left(\sigma\left(z_{1-\beta} - z_{1-\alpha}\right)\right) &> \left(\frac{1}{\mu_0 - \mu_A}\right) \times \left(\sqrt{n}\left(\mu_0 - \mu_A\right)\right), && H_A:\ \mu > \mu_0; && (116)\\
\left(\frac{1}{\mu_0 - \mu_A}\right) \times \left(\sigma\left(z_{1-\beta} + z_{1-\alpha}\right)\right) &< \left(\frac{1}{\mu_0 - \mu_A}\right) \times \left(\sqrt{n}\left(\mu_0 - \mu_A\right)\right), && H_A:\ \mu > \mu_0. && (117)
\end{aligned}$$

Recalling the general formula for Cohen's $d = (\mu_1 - \mu_2)/\sigma$

$$\tag{118}$$

Week 6

Contingency tables and tests for independence

The term contingency table comes from the fact that they outline all potential events in a certain situation. You usually want to know if 2 categorical variables, Cat$_1$ and Cat$_2$, are independent and test this with hypothesis testing: H_0 is always that the variables are independent (odds ratio equal to 1) and H_a is that they are dependent is some way. There are the usual alternative hypotheses:

		Cat$_1$		
		X	Y	Marginal
Cat$_2$	W	a	b	$a+b$
	Z	c	d	$c+d$
	Marginal	$a+c$	$b+d$	n

Table 4

- Left-tailed: there is negative association between the variables. That is, the observations tend to lie in lower left and upper right of the contingency table. Formally, H_A : odds ratio less than 1.
- Right-tailed: there is positive association between the variables. That is, the observations tend to lie in upper left and lower right. of the contingency table. Formally, H_A : odds ratio less than 1.
- Two-tailed: there is no prior alternative. Formally, H_A : odds ratio not equal to 1.

The odds ratio is the ratio of probabilities: the numerator is the probability that event W occurs given that event X has/will occur divided by the probability that event W occurs given that event Y has/will occur - the odds of W given X or Y. The denominator is the probability that event Z occurs given that event X has/will occur divided by the probability that event Z occurs given that event Y has/will occur - the odds of Z given X or Y. When you hear, "the odds are 3:2", this refers to the odds ratio which would then be 3/2. Similarly, if the odds are 7:1 then the odds ratio is 7/1.

$$\begin{aligned}
\frac{\mathbb{P}(W|X)/\mathbb{P}(W|Y)}{\mathbb{P}(Z|X)/\mathbb{P}(Z|Y)} &= \frac{\left(\frac{a}{a+c}\right)/\left(\frac{b}{b+d}\right)}{\left(\frac{c}{a+c}\right)/\left(\frac{d}{b+d}\right)} && (119)\\
&= \frac{\left(\frac{a}{a+c}\right)/\left(\frac{c}{a+c}\right)}{\left(\frac{b}{b+d}\right)/\left(\frac{d}{b+d}\right)} && (120)\\
&= \frac{a/c}{b/d} && (121)
\end{aligned}$$

The Cat$_1$ is independent of Cat$_2$ if this odds ratio is equal to 1 *in the population.* We do hypothesis testing using a sample to determine this. In this course, usually two-sided hypothesis testing is conducted as you are not bothered by the positive or negative association, just whether there is any association. If the odds ratio equals 1 in the population, you say that the probability that event W or Z occurs is equally likely, regardless

of X or Y occurring, i.e. the chance of W or Z is independent of X or Y.
This is attributed to Bayes' Rule for Independence, which simply states

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A) \times \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A). \tag{122}$$

The degrees of freedom is $(r-1) \times (c-1)$ where r is the number of rows and c is the number of columns of the contingency table; for a 2×2, the df is $(2-1) \times (2-1) = 1$.
It is also possible that you encounter contingency tables with more than two factors per variable, for example: suppose in a population of 1 million you take a random sample of 650 people to compare their neighbourhood with their occupation. The steps for undertaking a test for independence using contingency tables is as follows:

		Neighbourhood				
		A	B	C	D	Total
Occupation	White collar	90	60	104	95	349
	Blue collar	30	50	51	20	151
	Unemployed	30	40	45	35	150
	Total	150	150	200	150	650

Table 5: This is a 3 × 4 contingency table, for which the odds ratio is not as simply calculated as it was for the 2 × 2 contingency table.

1. **Formalise your hypotheses:**
H_0 : the variables "neighbourhood" and "occupation" are independent - $OR = 1$.
H_A : they are dependent - $OR \neq 1$.

2. **Compute cell proportions under null hypothesis:**
You begin by assuming that the variables are independent, so the probability that a person lives in neighbourhood A and has a white collar job is just the multiplication of the probabilities of either:

$$\begin{aligned}\mathbb{P}(\text{white collar job} \cap \text{n'hood A}) &= \mathbb{P}(\text{white collar job}) \times \mathbb{P}(\text{white collar job}) & (123)\\ &= \frac{349}{650} \times \frac{150}{650} & (124)\\ &\approx 0.1239. & (125)\end{aligned}$$

This says that, if the variables are independent we should have roughly 12.39% of the sample in that particular cell. Compute all and draw up a table of proportions or percentages under the null hypothesis:

		Neighbourhood				
		A	B	C	D	Total
Occupation	White collar	12.39%	12.39%	16.52%	12.39%	53.69%
	Blue collar	5.36%	5.36%	7.15%	5.36%	23.23%
	Unemployed	5.33%	5.33%	7.10%	5.33%	23.09%
	Total	23.08%	23.08%	30.77%	23.08%	100%

Table 6

3. **Check the expected cell frequencies:**
To determine the expected frequencies, you calculated what you expect the cell frequency to be (under the null hypothesis and with sample size n), i.e. $f_e = n \times \%$, where % is drawn from Table 6. The expected frequency for white collar workers living in neighbourhood A in a sample of size 650 given that the % chance of either is 53.69% and 23.08%, respectively, is

$$\mathbb{E}(\text{white collar job} \cap \text{n'hood A}) = 650 \times 12.39\% = 80.535. \tag{126}$$

Now you can do the same for the other 11 cells:

(a) If you have a 2×2 contingency table, then df=1 and you need to ensure that the expected frequency *in every cell* is more than 10. *If not*, use Fisher's Exact test.

(b) If you have an $r \times c$ contingency table, with df> 1, then you need to ensure that the expected frequency *in every cell* is more than 5. *If not*, use Fisher's Exact test.

		Neighbourhood				
		A	B	C	D	Total
Occupation	White collar	80.535	80.535	107.38	80.535	349
	Blue collar	34.84	34.84	46.475	34.84	151
	Unemployed	34.645	34.645	46.15	34.645	150
	Total	150	150	200	150	650

Table 7: Expected frequencies - it is important that the marginal frequencies are the same as in Table 5.

4. **Conduct the test:**

   (a) **Chi-squared test χ^2:**
   In Table 7, all expected frequencies are greater than 5 so we are fine conducting a *Pearson χ^2 test for independence.*

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(f_{o,i} - f_{e,i}\right)^2}{f_{e,i}} \sim \chi^2_{\text{df}}; \qquad \text{where } k = r \times c \text{ and df} = (r-1) \times (c-1). \tag{127}$$

This says that the test statistic χ^2 is the sum of $k = r \times c$ ratios of the squared difference between the observed and expected frequency, and the expected frequency, which is modelled by the chi-squared distribution with $(r-1) \times (c-1)$ degrees of freedom. For our example:

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{1} 2 \frac{\left(f_{o,i} - f_{e,i}\right)^2}{f_{e,i}} && (128)\\
&= \frac{(90-80.535)^2}{80.535} + \frac{(60-80.535)^2}{80.535} + \frac{(104-107.38)^2}{107.38} + \frac{(95-80.535)^2}{80.535} && (129)\\
&\quad + \frac{(30-34.84)^2}{34.84} + \frac{(50-34.84)^2}{34.84} + \frac{(51-46.475)^2}{46.475} + \frac{(20-34.84)^2}{34.84} && (130)\\
&\quad + \frac{(30-34.645)^2}{34.645} + \frac{(40-34.645)^2}{34.645} + \frac{(45-46.15)^2}{46.15} + \frac{(35-34.645)^2}{34.645} && (131)\\
&\approx 1.1124 + 5.2361 + 0.10639 + 2.5981 && (132)\\
&\quad + 0.67238 + 6.5966 + 0.44057 + 6.3211 && (133)\\
&\quad + 0.62278 + 0.82771 + 0.028657 + 0.0036376 && (134)\\
&= 24.566 \sim \chi^2_6. && (135)
\end{aligned}$$

A general rule of thumb for critical χ^2 values: for $\alpha = 0.05$, the critical value with df$= (r-1) \times (c-1)$ is near $r \times c$. For example, for 2 degrees of freedom (possibilities: 2×3 or 3×2 contingency tables) the critical value is 5.991 which is near $6 = 2 \times 3$. Another example, for 6 degrees of freedom (possibilities: 2×7, 3×4, 4×3, or 7×2) the critical value is 14.449 which is near $14 = 7 \times 2$.
If df=1, then you can just look at z^2 for the same critical values, i.e. $\chi^{2,*}_{1,\alpha} = \left(z^*_\alpha\right)^2$. We can use Table 4 to model z^2:

$$\begin{aligned}
\hat{\pi} &= \frac{a+c}{n} && (136)\\
se_0 &= \sqrt{\frac{\hat{\pi}\,(1-\hat{\pi})}{n}} && (137)\\
\hat{\pi}_1 &= \frac{a}{a+b} && (138)\\
\hat{\pi}_2 &= \frac{c}{c+d} && (139)\\
z^2 &= \frac{\left(\hat{\pi}_1 - \hat{\pi}_2\right)^2}{se_0^2} \sim \chi^2_1. && (140)
\end{aligned}$$

In our case, we have that the test statistic $\chi^2 = 24.566$ is more extreme than the critical value of $\chi^{2,*}_{6,0.05} = 12.592$ and thus provides significant evidence at the 5% level of variable dependency.

$$p\text{-value} = \mathbb{P}\left(\chi^2_6 > 24.566\right) = 0.0004098 < 0.05 = \alpha = \mathbb{P}\left(\chi^2_6 > 12.592\right). \tag{141}$$

   (b) **Fisher's Exact test:**
   You don't need to know how to compute this manually, only that you can use the computer to do it an must do so if an expected cell frequencies is less than 5 in a contingency table. It gives you the p-value so you can make inferences directly from the SPSS output.

Homework

1.1 Homework 1

1. In an experiment on the methods of teaching statistics, researchers studied the effects of taking a lab class. A random sample of 46 of Statistics 1A students was selected. Out of the 46 students, 23 were randomly assigned to an experimental group and the other 23 to a control group. The control group immediately started in week 1 with the lab class. In the experimental group, the lab class was postponed for 4 weeks. At the end of the semester both groups were evaluated on their knowledge of statistics.
Assume that in the population, the test scores of the control group are distributed as a Normal with mean $\mu_X = 56$ and standard deviation $\sigma = 9$. When the lab class is postponed, the population distribution is Normal with mean $\mu_Y = 49$ and standard deviation $\sigma = 9$. Thus, the experimental manipulation leads, on average, to lower test scores.

(a) What is the sampling distribution of the mean score $\bar{X}$ for students assigned to the control group?

Solution Refer to § 4.5 (Agresti) on page 97, if $X \sim \mathcal{N}(56, 81)$, then $\bar{X} \sim \mathcal{N}(56, {}^{81}\!/_{23})$ is the sampling distribution of the mean score for the students assigned to the control group.

(b) What is the sampling distribution of the mean score $\bar{Y}$ for students assigned to the experimental group?

Solution Similarly, if $Y \sim \mathcal{N}(49, 81)$ then $\bar{Y} \sim \mathcal{N}(49, {}^{81}\!/_{23})$ is the sampling distribution of the mean score for the students assigned to the experimental group.

(c) What is the sampling distribution of the difference $\bar{X} - \bar{Y}$ between the mean scores of both groups? Assume that X and Y are independent.

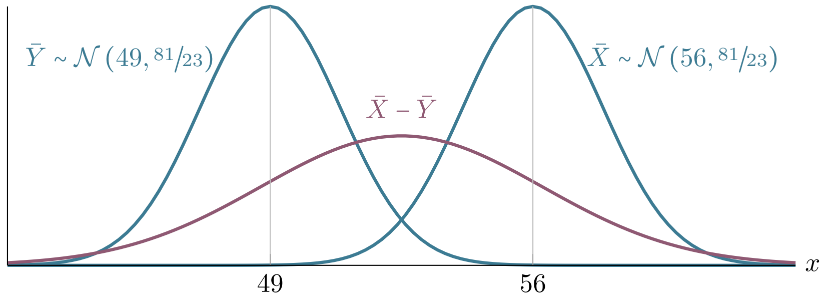


Figure 10: This graph shows the distributions from parts a and b of question 1 (homework 1), with the sampling distribution of the difference in scores overlaid.

Solution You can see the sampling distributions of the control and experimental groups in Figure 10 with the sampling distribution of the difference laid over the top. As both $\bar{X}$ and $\bar{Y}$ are normal and independent, it is easy to relate the mean and standard deviation of their difference:

$$\mu_{\bar{X}-\bar{Y}} = \frac{\sum_{j=1}^{n}\left(\bar{X}_i - \bar{Y}_i\right)}{n} = \frac{\left(\bar{X}_1 - \bar{Y}_1\right) + \ldots \left(\bar{X}_n - \bar{Y}_n\right)}{n} \tag{142}$$

$$= \frac{\overbrace{(56-49) + \ldots (56-49)}^{n \text{ times}}}{n} = \frac{n \times (56-49)}{n} = 56 - 49 = 7. \tag{143}$$

Similarly, for the standard deviation (which we know is the square root of the variance):

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\sigma^2_{\bar{X}-\bar{Y}}} \tag{144}$$

You can remember from 1A that if we have two normally distributed independent random variables, and there is a linear transformation (here the linear transformation is $V = \bar{X} - \bar{Y} = \bar{X} + \left(\bar{Y}\right))$ then $\mu_V = \mu_{\bar{X}} - \mu_{\bar{Y}}$ and $\sigma^2_V = \sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}}$.

$$= \sqrt{\sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}}} = \sqrt{2 \times \frac{81}{23}} = \sqrt{\frac{162}{23}} \tag{145}$$

$$\approx \sqrt{7.0435} \approx 2.654. \tag{146}$$

So, the sampling distribution of the difference between the mean scores of both groups $\bar{X} - \bar{Y} \sim \mathcal{N}\left(7, {}^{162}/_{23}\right)$.

(d) What is the chance that the experiment outlined above, with 23 students in each group, will determine that the mean in the experimental group is greater than the mean in the control group; that is, that the sample means will mislead us about the effectiveness of the lab classes?

Solution This question is a little long and wordy, however it is asking you to calculate the probability that $\bar{Y} > \bar{X}$, which is the same as calculating the probability that some random variable $V < 0$ where $V = \bar{X} - \bar{Y}$. We can use the standard normal distribution and the z-table so solve this:

$$\mathbb{P}\left(\bar{X} < \bar{Y}\right) = \mathbb{P}\left(V < 0 \middle| V = \bar{X} - \bar{Y} \sim \mathcal{N}\left(7, {}^{162}/_{23}\right)\right) \tag{147}$$

$$= \mathbb{P}\left(\frac{V - \mu_V}{\sigma_V} < \frac{0 - \mu_V}{\sigma_V}\right) \tag{148}$$

$$= \mathbb{P}\left(Z < \frac{-7}{2.654}\right) \tag{149}$$

$$= \mathbb{P}\left(Z < -2.64\right) \tag{150}$$

Using Table A, we find that $\mathbb{P}(Z < -2.64) = 0.0041$ which allows us to conclude that the experiment outlined above will determine that the mean in the experimental group is greater than the mean in the control group with an infinitesimally small chance ($< 0.5\%$). This has been plotted in Figure 11.

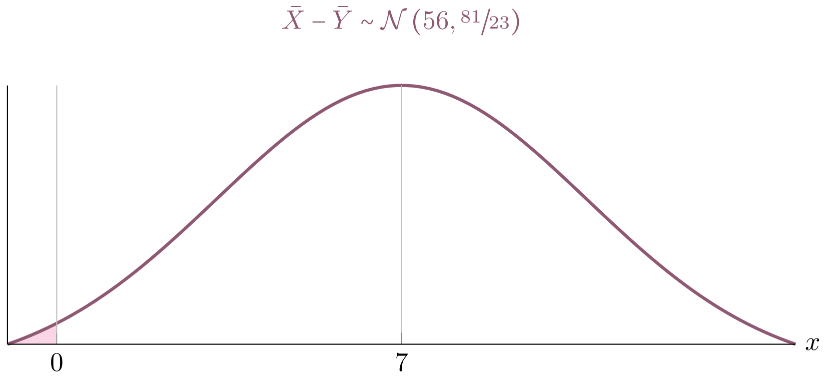


Figure 11: This graph shows the sampling distribution of the difference in scores from parts a and b of question 1 (homework 1). The shaded area on the left (between the axis and zero) represents the probability that the difference is less than zero, which equates to approximately 0.41%.

2. A test to determine a person's blood pressure is not completely accurate. Moreover, a person's blood pressure varies from day to day. Assume that repeated measurements on different days for the same person are normally distributed with a standard deviation $\sigma = 5$. Assume that the measurements are independent.

(a) Marja's blood pressure (diastolic pressure) is measured once, with the result $X = 80$. Compute the 95% confidence interval for her blood pressure.

Solution Here, we have that $n = 1$ so our confidence interval will be subject to a lot of scrutiny, given the relatively large standard error for the sampling distribution of the mean of Marja's blood pressure. The mean of one sample is the value of the single sample taken, thus we have that $\bar{X} = 80$. Furthermore, the standard error is given by $^{\sigma}/_{\sqrt{n}} = 5$. We know from § 5.3 that the general formula for a confidence interval is **estimate ± margin of error**, where the margin of error incorporates the following: if $\bar{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$ then the margin of error is the critical value multiplied by the standard error of the estimate, which is $\sigma/\sqrt{n}$.

$$95\%\text{ CI for }\mu: \qquad \bar{x} \pm z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} = 80 \pm 1.96 \times 5 \tag{151}$$

$$= (70.2, 89.8)\,. \tag{152}$$

It is important to remember this general formula!!! If you wish to narrow the confidence interval (i.e. produce a better range to estimate the true parameter value), then the **best** way to do that is increase your sample size. The other way is to increase your alpha, which leads to a higher probability of Type I errors, so the former method is **better**.

(b) Over the course of five days, daily measurements were taken of Marja's blood pressure. The mean measurement is $\bar{X} = 80$. Compute the 95%-confidence interval for Marja's blood pressure.
Compare the two confidence intervals you computed. Which is smaller and why? What does a smaller confidence interval mean?

Solution Now we have that $n = 5$, so our standard error is $\sigma/\sqrt{n} = 5/\sqrt{5} = \sqrt{5} \approx 2.236$. In part a the margin of error was $1.96 \times 5 = 9.8$, but now we have a new margin of error $1.96 \times \sqrt{5} \approx 4.383$. So, our new margin of error is less than half of our old margin of error, just because we collected an additional 4 samples! We can now compute the new CI:

$$\textit{CI for }\mu: \qquad \bar{x} \pm z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} = 80 \pm 1.96 \times \sqrt{5} \tag{153}$$

$$= (75.6, 84.4)\,. \tag{154}$$

Notice that this confidence interval is narrower than the one we computed in part a? In part a, the CI had a range of 19.6, and in part b we have reduced the range to 8.8 (range of CI = 2× margin of error). The reason that it is narrower is that we have reduced our margin of error. This tells us that we are 95% confident that the true population mean of Marja's blood pressure lies between 75.6 and 84.4.

3. A researcher is conducting a study on the price of lecture books for Psychology classes. A random sample is taken of 12 lecture books that are being used for the study Psychology. Prices (in €) are as follows:

53,95	54,50	99,00	69,00
83,00	77,00	32,00	86,05
86,05	79,00	52,50	69,89

Assume that the price of a lecture book is normally distributed in the population and that the standard deviation (σ_X) in the population of all lecture books is known and equals €20.

(a) Compute $\sigma_{\bar{X}}$, the standard deviation of $\bar{X}$.

Solution We have a sample of size $n = 12$ and the population standard deviation $\sigma_X = 20$, so the standard deviation for the sample mean is the standard error of the estimate, which is:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{20}{\sqrt{12}} \approx €5,77. \tag{155}$$

(b) Given the sample in the Table above, compute the 95%-confidence interval for μ, the mean of the population from which the sample was taken.

Solution Firstly, we use the sample to provide an estimate for μ:

$$\begin{aligned}
\bar{x} &= \frac{\sum_{i=1}^{12} x_i}{12} && (156)\\
&= \frac{1}{12} \times (53,95 + 54,50 + 99,00 + 69,00 + 83,00 + 77,00 && (157)\\
&\qquad +32,00 + 86,05 + 86,05 + 79,00 + 52,50 + 69,89) && (158)\\
&= \frac{841.94}{12} = 70.161\bar{6} \approx €70,16 && (159)
\end{aligned}$$

Now, we can compute the CI:

$$\begin{aligned}
\textit{95\% CI for } \mu: \qquad \bar{x} \pm z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} &= €70,16 \pm 1.96 \times €5,77 && (160)\\
&= (€58,85, €81.47)\,. && (161)
\end{aligned}$$

This CI tells us that the true population mean lies between €58,85 and €81.47 with 95% certainty, based on the sample.

(c) Do you think it is likely that the true mean price of the population of lecture books is less than €90?

Solution Referring back to our computed CI in part b, we can be confident that in 95% of cases the true value does indeed lie below €90 because it is not contained in the CI. Referring back to the previous question, we learnt that the only way to change the range of a CI is to alter the margin of error and the best way to do that is by altering the sample size. If we wished to ensure that €90 was included in our CI, we would need to reduce our sample size to 3 (keeping $\bar{x}$ and α the same):

$$\begin{aligned}
90 &< 70,16 + 1.96 \times \frac{20}{\sqrt{n}} = \text{upper bound for 95\% CI} && (162)\\
\implies n &< \left(\frac{1.96 \times 20}{(90 - 70,16)}\right)^2 \approx 3.904. && (163)
\end{aligned}$$

Now assume that we would rather report the price of the lecture books in US dollars instead of Euros; we know that 1 Euro equals 1.07 dollars. Use your answers to the previous questions in order to answer the following.

(d) Compute the mean price in dollars for the lecture books.

Solution We must transform our random variable X (EURO) to Y (USD) using the following linear transformation $Y = 1.07X$. This results in the following transformation for our sample mean:

$$\bar{Y} = \frac{\sum_{i=1}^{12} Y_i}{12} = \frac{\sum_{i=1}^{12} (1.07X_i)}{12} = 1.07 \times \frac{\sum_{i=1}^{12} X_i}{12} = 1.07 \times \bar{X} = 75.07298\bar{3} \approx \$75.07. \qquad (164)$$

(e) Compute the standard deviation of the mean price for the lecture books in dollars.

Solution Using the same linear transformation:

$$\begin{aligned}
\sigma^2_{\bar{Y}} &= 1.07^2 \times \sigma^2_{\bar{X}} = 38.16\bar{3}. && (165)\\
\implies \sigma_{\bar{Y}} &= \sqrt{38.16\bar{3}} \approx \$6.18 && (166)
\end{aligned}$$

(f) Compute the 95%-confidence interval for the mean price in dollars for the lecture books in the population.

Solution

$$\textit{95\% CI for } \mu: \qquad \bar{x} \pm z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} = \$75.07 \pm 1.96 \times \$6.18 \tag{167}$$

$$= (\$62.96, \$87.18)\,. \tag{168}$$

This CI tells us that the true population mean lies between \$62.96 and \$87.18 with 95% certainty, based on the sample.

(g) Explain what the confidence interval computed in part f means. How would you explain this to someone who does not know what a confidence interval is?

Solution The confidence interval states that, based on a random sample of 12 books drawn from a population which deviates (on average) by \$21.04, the average cost of any particular Psychology lecture book lies between \$62.96 and \$87.18 in 95% of cases. In layman's terms, you could tell a prospective Psychology student that they can expect to pay somewhere between \$62.96 and \$87.18 per book, however there is a 5% chance that the cost could be more or less.

4. A study on depression among twelve-year-old children is planned. The researchers want to compute the 90%-confidence interval for the population mean score on a depression scale. They want the 90% margin of error to be no bigger than 5 points. The standard deviation in the population is 12.

(a) How large does the sample N of twelve-year-old children need to be in order to yield a 90% margin of error of less than 5 points?

Solution Recall that the margin of error in a confidence interval is based on three things: the prescribed acceptance level α, the sample size N, and the standard deviation in the population σ.

$$\text{margin of error } = z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{N}}. \tag{169}$$

If the margin of error must be less than 5 points, we have that

$$5 > z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{N}}. \tag{170}$$

If we now substitute $z^*_{1-\alpha/2} = 1.645$ (for 90%) and $\sigma = 12$, we can now rearrange to solve for N:

$$5 > 1.645 \times \frac{12}{\sqrt{N}} \tag{171}$$

$$\implies N > \left(\frac{1.645 \times 12}{5}\right)^2 = 15.586704. \tag{172}$$

Recall that N is an integer, so if $N \geq 16$, we can be sure that the 90% margin of error is less than 5 points when the population standard deviation is 12. In general, to find the N which limits the margin of error to a particular maximum we calculate

$$N > \left(\frac{z^*_{1-\alpha/2} \times \sigma}{\text{max margin of error}}\right)^2. \tag{173}$$

(b) If they decided they wanted a 95% margin of error, rather than a 90%, to be no larger than 5 points, would they have to survey more participants or fewer? Why?

Solution If the margin of error is changed from 90% to 95%, what has been altered is the prescribed α, i.e. it has decreased. If α decreases, then the critical value increases: $z^*_{1-0.10/2} = 1.645 < 1.96 = z^*_{1-0.05/2}$. Noting (188), we can calculate the new required sample size:

$$N > \left(\frac{z^*_{1-\alpha/2} \times \sigma}{\text{max margin of error}}\right)^2 \tag{174}$$

$$= \left(\frac{1.96 \times 12}{5}\right)^2 = 22.127616 \tag{175}$$

Again, remembering that N is an integer we conclude that the sample size must be 23 or more ($N \geq 23$) in order for the 95% margin of error to remain less than 5 points, where the standard deviation in the population is 12.

If we did not increase the sample size when we changed our α, the margin of error would be:

$$\text{margin of error} = 1.96 \times \frac{12}{\sqrt{16}} = 1.96 \times 3 = 5.88 > 5. \tag{176}$$

1.2 Homework 2

1. Researcher John is interested in the extent to which stress influences human memory. He has run a study in which a random sample of 25 participants has been told that they will be required to give a lecture in front of a large audience. Giving a speech in front of a large audience is considered a stressful event by many people.

While the participants were busy preparing for their presentation, they were asked to complete a memory task. It is known that in "typical" situations (without a lot of stress), the mean number of questions on the memory task is 53. The *observed* mean number of correctly-answered questions for these 25 participants is $\bar{X} = 48$. Assume that the standard deviation of memory scores in the population is 7.

The null hypothesis is that the manipulation did not affect the scores. The alternative hypothesis is that the true mean number of correctly answered questions is smaller than 53; that is, the stress of the preparation for the presentation hurts performance. The hypotheses may be written formally:

$$\begin{aligned} H_0 &: \mu = 53 \\ H_A &: \mu < 53 \end{aligned}$$

(a) Carry out the test and give the p-value.

Solution

First step: extract the important information!!!

$$X \sim (53, 49) \tag{177}$$

In the population, the mean is 53 with standard deviation 7 (variance $7^2 = 49$) - no assumptions about normality.

$$n = 25 \tag{178}$$

$$\bar{x} = 48 \tag{179}$$

A sample of size 25 is drawn from the population and the computed mean of the sample is 48.

$$\bar{X} \sim \mathcal{N}(53, 1.96) \tag{180}$$

The sampling distribution of the mean of X is normally distributed with mean 53 and standard error $7/\sqrt{n} = 7/5 = 1.4$ (variance $1.4^2 = 1.96$) - this conclusion is drawn thanks to the CLT (drawing on the LLN).

Next step: find the p-value! The p-value is the probability of observing the same, or more extreme, sample statistic in an any other sample collection of the same size. So, if we were to collect 100 million samples from this particular population, each of size 25, how likely is it that we would would observe a mean equal to or more extreme than 48. **Remember** that this probability is computed under the null hypothesis, so we are assuming that the population mean is 53.

$$\begin{aligned} p &= \text{probability of observing } \bar{x} \leq 48, && (181) \\ &\qquad \text{when we have assumed } \bar{X} \sim (53, 1.96) && (182) \\ &= \mathbb{P}\left(\bar{X} < 48 \,|\, \mu_{\bar{X}} = 53\right) && (183) \\ &= \mathbb{P}\left(\frac{\bar{X} - 53}{\sqrt{1.96}} < \frac{48 - 53}{\sqrt{1.96}}\right) && (184) \end{aligned}$$

$$= \mathbb{P}\left(Z < -3.57\right) \tag{185}$$
$$< 0.0003 \text{ according to table A.} \tag{186}$$

Given that the p-value is infinitesimally small, we can conclude that it is **extremely unlikely** (at **any** significance level) that we observe a mean of 48 in a population that we suppose has a mean of 53, when the standard error is 1.4. However, we **have** observed a mean of 48, so what we previously supposed **must be incorrect**; we reject H_0 in favour of H_A. It must be, that in the population, the mean number of correctly-answered questions under stress is less than 53.

(b) Explain in words what the p value means. The p value is the probability that...

Solution See part a.

(c) How strong is the evidence that, for this memory test, the mean number of correctly answered questions in stressful situations is smaller that the mean number of correctly answered questions in "typical" situation?

Solution See part a.

2. Exercise 2 from week 3 (statistics 1A) describes a study by a psychologist interested in the relationship between children watching violent TV programs and displays of violent behavior by the children. Assume that the number of aggressive actions towards their friends in the population has a standard deviation of $\sigma = 3$. A random sample has been taken of 18 young children, who watch a lot of violent TV-programs relative to their peers. In this sample, the mean number of aggressive actions towards their peers is $\bar{X} = 4.5$.

(a) Find the 95%-confidence interval for the mean number of aggressive actions in the population of the sample of young children.

Solution The standard error of the sampling distribution of the mean of X is $3/\sqrt{n}$, where X is the number of aggressive actions and n is the size of the sample taken from the population. In our case, we observe a mean of 4.5 in a sample of 18 children.

$$\textit{95\% CI for } \mu: \qquad \bar{x} \pm z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} = 4.5 \pm 1.96 \times \frac{3}{\sqrt{18}} \tag{187}$$
$$= (3.11, 5.89) \tag{188}$$

(b) Suppose it is known that in cases where no violent TV programs are watched, the mean number of aggressive actions towards their peers is $\mu = 3$. Determine H_0 and H_A, then conduct a test at $\alpha = 0.05$ in order to determine whether watching violent TV programs will lead to more aggressive actions by young children.

Solution In order to ensure a test that only makes **appropriate changes to society**, we assume that watching violent TV programs has no affect on the number of aggressive episodes; $H_0 : \mu = 3$. Moreover, we wish to test against this hypothesis by saying that it will result in an increase of aggressive episodes; $H_A : \mu > 3$. Now, we will find the p-value associated with this test: how probable is it to observe 4.5 or more violent episodes if we assume that the mean is 3.

$$p = \mathbb{P}\left(\bar{X} > 4.5 \,|\mu_{\bar{X}} = 3\right) \tag{189}$$
$$= \mathbb{P}\left(\frac{\bar{X} - 3}{3/\sqrt{18}} > \frac{4.5 - 3}{3/\sqrt{18}}\right) \tag{190}$$
$$= \mathbb{P}\left(Z > 2.12\right) \tag{191}$$

You can find the p-value for this z using table A; it is equivalent to finding $\mathbb{P}\left(Z < -2.12\right)$.

$$= 0.0170 \tag{192}$$

(c) On the basis of both the test and the confidence interval, discuss your conclusions about the research question.

Solution According to our observations, in 95% of samples of size 18 (from the population) the true mean lies between 3.11 and 5.89, so we can not assume that the true mean is 3.
We have prescribed an $\alpha = 0.05 > p$, so we reject H_0: the true mean in the population must be more than 3. Under the assumption that the mean is 3, we found that there is less than a 5% chance of observing a mean of 4.5 or more in samples of size 18. Despite this low probability, we still observed a mean of 4.5 so we must conclude that the true mean is not 3, but must be more than 3.

(d) If you were reporting these results to another researcher, which would be more informative, the significance test or the confidence interval? Why?

Solution The confidence interval is the most informative, as it provides information about the true population parameter range to a certain level of confidence. Additionally, information about the strength and direction is given; a wide interval indicates a weak suggestion of the true parameter v.s. a narrow interval which gives a strong suggestion; if the hypothesised parameter lies in the interval, then we believe that this is the true value in the population (accept H_0).
The p-value tells us whether our findings are statistically noteworthy, but does not tell us where to find the true parameter value if it is indeed not what we supposed (if $p < \alpha$, then reject H_0 but we do not have μ).

3. Everyone that uses statistics must understand what the difference is between statistical significance and practical importance. With very large sample sizes, even very small effects can be found to be statistically significant.
Assume that a sleep therapy clinic has observed (during a number of years) that the time taken by adult men to fall asleep is normally distributed with a mean $\mu = 335$ seconds and a standard deviation of $\sigma = 15$ seconds. A reduction of 2 seconds in time taken (men fall asleep 2 seconds earlier) is of no practical use, but this unimportant change can be found to be highly statistically significant. Suppose there has been a study in which the goal was to determine if hypnotherapy was an effective way to make men fall asleep faster. A sample was taken of adult men, and each underwent hypnotherapy. After the hypnotherapy, the mean observed time taken to fall asleep was 333 seconds. Assume that the standard deviation in the population of times was still 15 seconds.
Compute the p value for the significance test in each of the following situations:

(a) 25 male adults are given hypnotherapy, and the mean time taken to fall asleep is 333 seconds.

Solution We have that $n = 25$ so the standard error is $\sigma/\sqrt{n} = 15/5 = 3$.

$$\begin{aligned} p &= \mathbb{P}\left(\bar{X} < 333 \,|\mu_{\bar{X}} = 335\right) && (193)\\ &= \mathbb{P}\left(\frac{\bar{X} - 335}{3} < \frac{333 - 335}{3}\right) && (194)\\ &= \mathbb{P}\left(Z < -0.67\right) && (195)\\ &= 0.2514 && (196) \end{aligned}$$

Here we have that $p > \alpha$ at 5% significance level, so we do not reject $H_0 : \mu = 335$.

(b) 100 male adults are given hypnotherapy, and the mean time taken to fall asleep is 333 seconds.

Solution We have that $n = 100$ so the standard error is $\sigma/\sqrt{n} = 15/10 = 1.5$.

$$\begin{aligned} p &= \mathbb{P}\left(\bar{X} < 333 \,|\mu_{\bar{X}} = 335\right) && (197)\\ &= \mathbb{P}\left(\frac{\bar{X} - 335}{1.5} < \frac{333 - 335}{1.5}\right) && (198)\\ &= \mathbb{P}\left(Z < -1.33\right) && (199)\\ &= 0.0918 && (200) \end{aligned}$$

Here we have that $p > \alpha$ at 5% significance level, so we do not reject $H_0 : \mu = 335$.

(c) 500 male adults are given hypnotherapy, and the mean time taken to fall asleep is 333 seconds.

Solution We have that $n = 500$ so the standard error is $\sigma/\sqrt{n} = 15/10\sqrt{5} \approx 0.671$.

$$\begin{aligned} p &= \mathbb{P}\left(\bar{X} < 333 \,|\, \mu_{\bar{X}} = 335\right) && (201)\\ &= \mathbb{P}\left(\frac{\bar{X} - 335}{0.671} < \frac{333 - 335}{0.671}\right) && (202)\\ &= \mathbb{P}\left(Z < -2.98\right) && (203)\\ &= 0.0014 && (204) \end{aligned}$$

Here we have that $p < \alpha$ at 5% significance level, so we do reject $H_0 : \mu = 335$.

(d) What can you conclude about the relationship between the observed size of the effect of the hypnotherapy and whether it is statistically significant?

Solution As n increases, the standard error ($\sigma/\sqrt{n}$) for the estimate of the mean (μ) decreases, meaning that the data is more centralised around the hypothesised mean of 335 seconds. This is shown in Figure 12, where the shape of the graph becomes less of a smooth bell curve and more pointy. Given that the graph is a probability density function, we can calculate the probability of observing particular values by summing the area under the graph up to that particular valued point. You can see in Figure 12 that the area under the graph to the left of 333 becomes smaller as n increases, so we can conclude that the probability (p-value) becomes smaller as n increases, i.e. the statistical significance of the observed size of the effect of hypnotherapy increases as n increases.

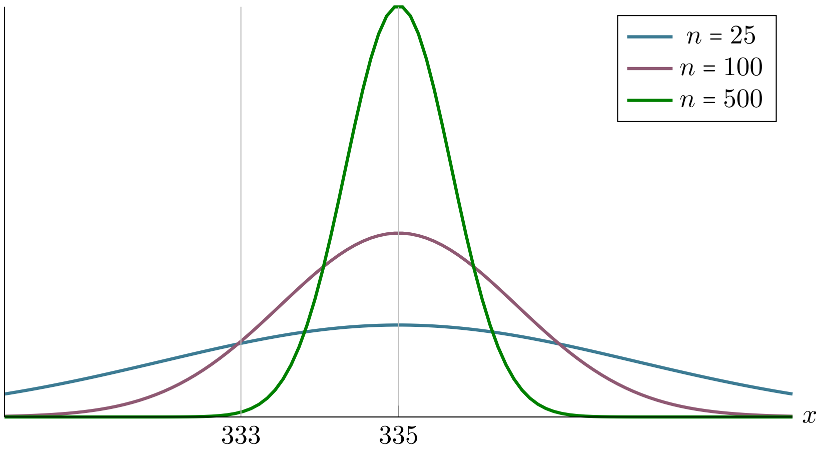


Figure 12: The sampling distributions of the mean of $X \sim \left(335, 15^2\right)$ for samples of size 25, 100 and 500.

1.3 Homework 3

1. In Exercise 3 of Section 2.1 we learned that the time it takes for adult men to fall asleep is distributed as a normal with a mean of $\mu = 335$ seconds and a standard deviation of $\sigma = 15$ seconds. Twenty-five adult men undergo hypnotherapy to decrease the time required to fall asleep. We wish to test the following hypotheses:

$$\begin{aligned} H_0 &: \mu = 335\\ H_A &: \mu < 335 \end{aligned}$$

(a) The mean observed time for the 25 men to fall asleep is 330.1 seconds. Will H_0 be rejected at a significance level of $\alpha = .05$?

Solution We reject H_0 when our sample statistic is more extreme than 95% of the population (note that the defined hypotheses imply a one-sided test):

$$\begin{aligned} z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} &< -z^*_{1-\alpha} && (205) \\ \Longrightarrow \frac{330.1 - 335}{15/\sqrt{25}} &< -1.645 && (206) \\ \Longrightarrow -1.6\bar{3} &< -1.645. && (207) \end{aligned}$$

As the above inequality does not hold (obviously $-1.633 > -1.645$), we can conclude that our statistic is not extreme enough to warrant us rejecting the null hypothesis. If we wish to find the p-value:

$$\begin{aligned} p &= \mathbb{P}\left(\bar{X} < 330.1 \,|\mu_0 = 335\right) && (208) \\ &= \mathbb{P}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{330.1 - 335}{15/\sqrt{25}}\right) && (209) \\ &= \mathbb{P}\left(Z < -1.6\bar{3}\right) && (210) \\ &= 0.0516. && (211) \end{aligned}$$

Again, we do not reject H_0 as we have that $p > \alpha$.

(b) The mean observed time for the 25 men to fall asleep is 330.0 seconds. Will H_0 be rejected at a significance level of $\alpha = .05$?

Solution In the same manner as before, we compare our statistic to the critical z-value:

$$\begin{aligned} z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} &< -z^*_{1-\alpha} && (212) \\ \Longrightarrow \frac{330.0 - 335}{15/\sqrt{25}} &< -1.645 && (213) \\ \Longrightarrow -1.\bar{6} &< -1.645. && (214) \end{aligned}$$

As the above inequality does hold (obviously $-1.667 < -1.645$), we can conclude that our statistic is extreme enough to warrant us rejecting the null hypothesis. If we wish to find the p-value:

$$\begin{aligned} p &= \mathbb{P}\left(\bar{X} < 330.0 \,|\mu_0 = 335\right) && (215) \\ &= \mathbb{P}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{330.0 - 335}{15/\sqrt{25}}\right) && (216) \\ &= \mathbb{P}\left(Z < -1.\bar{6}\right) && (217) \\ &= 0.0475. && (218) \end{aligned}$$

Again, we reject H_0 as we have that $p < \alpha$, but realise that we only reject at a 5% significance level and not at 1%.

(c) State to what extent you believe it is useful to think about results in terms of rigid, accept/reject decisions based on a particular significance level.

Solution How powerful is our test? Firstly, for what value $\bar{x}$ do we certainly reject H_0 (it's somewhere between 330.0 and 330.1 seconds):

$$\begin{aligned} z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} &< -1.645 = -z^*_{1-\alpha} && (219) \\ \Longrightarrow \bar{x} &< \mu_0 - 1.645 \times \frac{\sigma}{\sqrt{n}} = 335 - 1.645 \times 3 = 330.065 && (220) \\ \Longrightarrow &\text{ power = probability of correctly rejecting } H_0 && (221) \end{aligned}$$

Recall that we reject H_0 when $\bar{x} < 330.065$ and we are correct in doing so when the alternative hypothesis is true.

$$= \mathbb{P}\left(\bar{X} < 330.065 \,|\, \mu = \mu_A\right), \tag{222}$$

where μ_A is the true population mean. Suppose $\mu_A = 330.0$

$$\begin{align}
\implies \text{power} &= \mathbb{P}\left(\frac{\bar{X} - \mu_A}{\sigma/\sqrt{n}} < \frac{330.065 - 330.0}{15/5}\right) \tag{223}\\
&= \mathbb{P}\left(Z < 0.021\bar{6}\right) \tag{224}\\
&= 0.5080 \tag{225}
\end{align}$$

So, the probably of a Type II error is around 50%, which is relatively high so it does not seem useful to base our decision on a significance test in this case[2]. In order to increase the power of our test, we need to increase the sample size and in doing so we aim for power of at least 80%:

$$0.80 \approx \mathbb{P}\left(Z < 0.85\right) \tag{226}$$

You can find the above z critical value from Table A.

$$= \mathbb{P}\left(\frac{\bar{X} - \mu_A}{\sigma/\sqrt{n}} < \frac{330.065 - 330.0}{15/\sqrt{n}}\right). \tag{227}$$

So, we are looking for a n which satisfies

$$\begin{align}
\implies 0.85 &= \frac{330.065 - 330.0}{15/\sqrt{n}} \tag{228}\\
\implies n &= \left(\frac{0.85 \times 15}{330.065 - 330.0}\right)^2 \approx 38476.33. \tag{229}
\end{align}$$

This tells us that, if the researchers want to be 80% certain that they are correctly rejecting the null hypothesis, they would require at least 38,477 participants for this study (assuming that the true mean is 330.0 seconds). For a hypothetical μ_A, we require an n that satisfies:

$$n = \left(\frac{0.85 \times 15}{330.065 - \mu_A}\right)^2. \tag{230}$$

The size of the required n is extremely large if the distance between μ_A and 330.065 is less then 1, i.e. $329.065 < \mu_A < 331.065$, but is decreasing if the distance is more than 1. This also gives you a bit of an idea about power: obviously μ_A is **fixed in the population** but a researcher might have a hypothesised mean μ_0 is stupidly too far away from μ_A in order to overflate the power of their test.

We can also look at this from the perspective of confidence intervals, which we learnt in the previous homework is more informative than a p-value.

$$\begin{align}
90\% \text{ CI for } \mu: \qquad \bar{x} \pm z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} &= 330.1 \pm 1.645 \times \frac{15}{5} \tag{231}\\
&= \left(325.165, 335.035\right). \tag{232}
\end{align}$$

If we have a sample mean of 330.1 seconds, then the 90% CI (90%, because the hypothesis test is one-tailed) for μ contains $\mu_0 = 335$.

$$\begin{align}
90\% \text{ CI for } \mu: \qquad \bar{x} \pm z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} &= 330.0 \pm 1.645 \times \frac{15}{5} \tag{233}\\
&= \left(324.065, 334.935\right). \tag{234}
\end{align}$$

Here we see that $\mu_0 = 335$ is not in the CI (only just), but if we were to round up to the nearest second then it would be included. If we were to increase our n, then $\mu_0 = 335$ might not be in the 90% CI for either part a or b. This question is just trying to show you the silly-side of significance testing.

2. In Section 2.2.1, we discussed a neuropsychologist studying about information processing in parents with closed head injury (CHI). The neuropsychologist used a number of reaction time tasks that are assumed to require different forms of information processing. His hypothesis is that CHI patients process information

[2] This paper might be of interest `https://www.jstor.org/stable/449153`, or this one `https://academic.oup.com/ptj/article/79/2/186/2837119`.

more slowly than healthy people.
The test scores for healthy adults in the mirror-reading task are normally distributed with mean of 3000ms and standard deviation 400ms. The neurologist administers the task to 16 randomly selected CHI patients, at least 1 year after their accident. The researcher wonders how easy it would be to determine that the CHI patients were slower if they were much slower than the healthy participants. He chooses a 5% significance level, and "much slower" to him is 3300ms.

[Extract important information:]
$X \sim \mathcal{N}\left(3000, 400^2\right)$: test scores for healthy adults in the mirror-reading task.
$n = 16$ of CHI patients.

$$\begin{aligned} H_0 &: \mu = 3000 \\ H_A &: \mu > 3000 \end{aligned}$$

$\alpha = 0.05$ significance level.
$\bar{X} \sim \mathcal{N}\left(3000, 400^2/16\right)$: sampling distribution of the mean of X.

(a) Find the power of the test under the alternative hypothesis that μ for the CHI patients is 3300ms.

Solution The neuropsychologist considers rejecting H_0 if the sample statistic is more extreme than 3164.5ms:

$$\bar{x} > \mu_0 + z^*_{1-\alpha} \times \frac{\sigma}{\sqrt{n}} = 3000 + 1.645 \times \frac{400}{\sqrt{16}} = 3164.5. \tag{235}$$

Given that we are told the true population mean is 3300ms, how powerful is our test?

$$\text{power} = \mathbb{P}\left(\frac{\bar{X} - \mu_A}{\sigma/\sqrt{n}} > \frac{3164.5 - 3300}{400/\sqrt{16}}\right) = \mathbb{P}\left(Z > -1.355\right) = 0.9123. \tag{236}$$

This tells us that, given a sample size of 16 we will correctly reject the null hypothesis 91.23% of the time.

(b) Does the experiment have a good chance of finding a significant effect if the mean score for the CHI patients is truly 3300ms?

Solution You will find a significant effect is the p-value is less that α. This question is asking you for the probably of rejecting H_0 given that H_0 is false, i.e. what is the power. This question is hoping to trick you, but it's the same as part a and only hopes that you explore the theory a little deeper.
In general, researchers like to ensure that the power of their test is around 80% and in our case we have $\tilde{9}1\%$.

(c) Explain why the experiment will give much clearer results if the researcher were to take a larger sample.

Solution We reject the null hypothesis at 5% significance level when

$$\bar{x} > \mu_0 + z^*_{1-\alpha} \times \frac{\sigma}{\sqrt{n}} = 3000 + \frac{658}{\sqrt{n}}. \tag{237}$$

Assuming that the true population mean is 3300ms, the power of our test is

$$= \mathbb{P}\left(\frac{\bar{X} - \mu_A}{\sigma/\sqrt{n}} > \frac{\left(3000 + \frac{658}{\sqrt{n}}\right) - 3300}{400/\sqrt{n}}\right) \tag{238}$$

$$\text{power} = \mathbb{P}\left(Z > 1.645 - 0.75\sqrt{n}\right) \tag{239}$$

$$= \begin{cases} \mathbb{P}\left(Z > -1.355\right) = 0.9123, & n = 16; \\ \mathbb{P}\left(Z > -2.105\right) = 0.9821, & n = 25; \\ \mathbb{P}\left(Z > -2.855\right) = 0.99785, & n = 36; \\ \mathbb{P}\left(Z > -3.605\right) \approx 1, & n = 49. \end{cases} \tag{240}$$

If n increases, then the standard error of the estimate of μ decreases: $\bar{X}$ estimates μ such that $\bar{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$, where $\sigma/\sqrt{n}$ is the standard error of the estimate. In our case, $\bar{X} \sim \mathcal{N}\left(3000, 400^2/n\right)$ and in Figure 13 you can observe how increasing n causes the sampling distribution of the mean to become more centralised around μ, whether it be 3000ms or 3300ms.

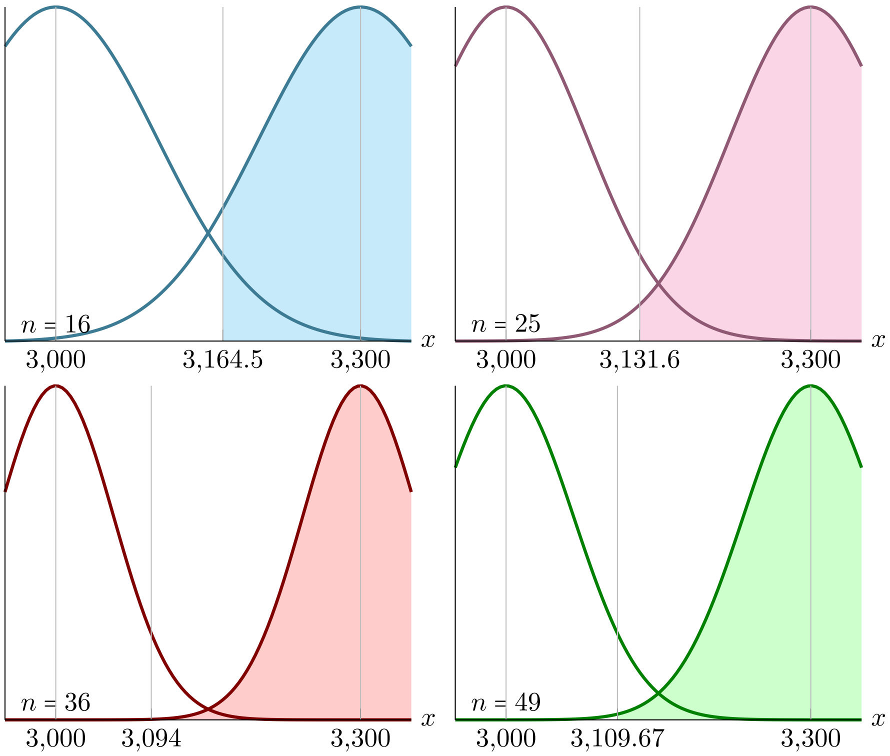


Figure 13: The shaded area represents the power of the test for increasing values of n, where the random variable $\bar{X}$ is distributed normally with standard error $400/\sqrt{n}$ ($\alpha = 0.05$; $H_0 : \mu = 3000$ms; $H_a : \mu > 3000$ms; true mean $\mu = 3300$ms). As n increases, the "cut-off value" decreases resulting in an increase of statistical power (caeteris paribus).

(d) Suppose that the alternative hypothesis is farther away from H_0; say, for instance, μ is truly 3500 for the CHI patients. Will the power than be higher or lower than the value you found in Question 2a?

Solution In the same way that we calculated power in (261):

$$\text{power} = \mathbb{P}\left(\frac{\bar{X} - \mu_A}{\sigma/\sqrt{n}} > \frac{\left(3000 + \frac{658}{\sqrt{n}}\right) - 3500}{400/\sqrt{n}}\right) \tag{241}$$

$$= \mathbb{P}\left(Z > 1.645 - 1.25\sqrt{n}\right) \tag{242}$$

$$= \begin{cases} \mathbb{P}\left(Z > -3.355\right) = 0.9996, & n = 16; \\ \mathbb{P}\left(Z > -3.5\right) \approx 1, & n > 16. \end{cases} \tag{243}$$

Fixing n and α, the power of the test increases dramatically if the true mean is much farther away from the hypothesised mean than if they are close.

(e) Another way to determine whether the sample size is large enough to yield worthwhile results is to determine whether the 90% confidence interval will be narrow enough. (90%, because the hypothesis test is one-tailed). The researcher would like the confidence interval to be more narrow than 60ms. Is his sample size large enough?

Solution If the researcher would like the confidence interval to be more narrow than 60ms, then they would like to ensure that the margin of error is no more than 30ms:

$$\text{margin of error} = z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} < 30 \tag{244}$$

$$\implies 1.645 \times \frac{400}{\sqrt{n}} < 30 \tag{245}$$

$$\implies n > \left(\frac{1.645 \times 400}{30}\right)^2 \approx 481.071. \tag{246}$$

The sample size of 16 is too small to ensure that the confidence interval be narrower than 60ms; a sample size of at least 482 is needed.

(f) Explain the relationship between the width of the confidence interval and how worthwhile the experiment is.

Solution If the researcher wants to ensure that their experiment is worthwhile then they need to ensure that the probabilities of Type I or II errors are maintained at an acceptable level, which are typically $\alpha = 0.05$ and $\beta = 0.20$. The value of α will dictate which critical value is used for the width of the interval and β will dictate the appropriate sample size.

$$\text{width of the CI} = 2 \times z^*_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}. \tag{247}$$

$$1 - \beta = \mathbb{P}\left(Z < \left| z^*_{1-\alpha} - \frac{|\mu_0 - \mu_A|}{\sigma/\sqrt{n}} \right|\right). \tag{248}$$

The researcher prescribes $\alpha = 0.05$ and $\beta = 0.20$.

$$\implies \text{width of the CI} = 2 \times 1.96 \times \frac{\sigma}{\sqrt{n}}. \tag{249}$$

The researcher can increase or decrease their sample size to ensure an appropriate width, or alternatively they can ensure that their requirements for power are met:

$$\implies 0.80 = \mathbb{P}\left(Z < \left| 1.645 - \frac{|\mu_0 - \mu_A|}{\sigma/\sqrt{n}} \right|\right) \tag{250}$$

Checking table A we note that $\mathbb{P}(Z < 2.05) \approx 0.8$.

$$\implies \left| 1.645 - \frac{|\mu_0 - \mu_A|}{\sigma/\sqrt{n}} \right| = 2.05 \tag{251}$$

$$\implies n = \left(\frac{|\mu_0 - \mu_A|}{\sigma \times (2.05 - 1.645)}\right)^2 \tag{252}$$

To summarise, a researcher ensures that their experiment is worthwhile (reducing probability of Type I and II errors) by setting the sample size to restrict the width of the confidence interval.

3. The scores of men on the Chapin Test for Social Insight (CTSI; [Chapin, 1942]) are normally distributed with mean 25 and standard deviation 5. Assume a sample size of $N = 16$. A study is run to determine whether women have a lower average score on the CTSI than men. We let μ be the mean of the population of women, and we wish to test the following hypotheses:

$$\begin{array}{lcl} H_0 & : & \mu = 25 \\ H_a & : & \mu < 25 \end{array}$$

We will assume that the standard deviation of the population of women is also 5. For $\alpha = 0.05$, H_0 can be rejected when $\bar{x} < 22.95$.

(a) Compute the probability of a Type I error. In other words, compute the probability that one will reject H_0 if the mean of the population of women is exactly $\mu = 25$.

Solution Under H_0, the probability of a Type I error is equal to α.

$$\alpha = \mathbb{P}\left(\bar{X} < 22.95 \,|\mu = 25\right) \tag{253}$$

$$= \mathbb{P}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < \frac{22.95-25}{5/\sqrt{16}}\right) \tag{254}$$
$$= \mathbb{P}(Z < -1.64) \tag{255}$$
$$= 0.0516 \tag{256}$$

(b) Compute the probability of a Type II error when $\mu = 22$. In other words, compute the probability that one will not reject H_0 if the mean of the population of women truly is $\mu = 22$.

Solution We don't reject H_0 if $\bar{x} \geq 22.95$, so

$$\beta = \mathbb{P}\left(\bar{X} > 22.95 \,|\, \mu = 22\right) \tag{257}$$
$$= \mathbb{P}\left(\frac{\bar{X}-\mu_A}{\sigma/\sqrt{n}} > \frac{22.95-22}{5/\sqrt{16}}\right) \tag{258}$$
$$= \mathbb{P}(Z > 0.76) \tag{259}$$
$$= 0.2236. \tag{260}$$
$$\Longrightarrow \text{power} = 1-\beta = 0.7764. \tag{261}$$

(c) Compute the probability of a Type II error if $\mu = 20$.

Solution

$$\beta = \mathbb{P}\left(\bar{X} > 22.95 \,|\, \mu = 20\right) \tag{262}$$
$$= \mathbb{P}\left(\frac{\bar{X}-\mu_A}{\sigma/\sqrt{n}} > \frac{22.95-20}{5/\sqrt{16}}\right) \tag{263}$$
$$= \mathbb{P}(Z > 2.36) \tag{264}$$
$$= 0.0091. \tag{265}$$
$$\Longrightarrow \text{power} = 1-\beta = 0.9909 \tag{266}$$

(d) Using your computation of the Type II error rate of the test, find the power if $\mu = 22$.

Solution Power $= 1 - \beta = 0.7764$.

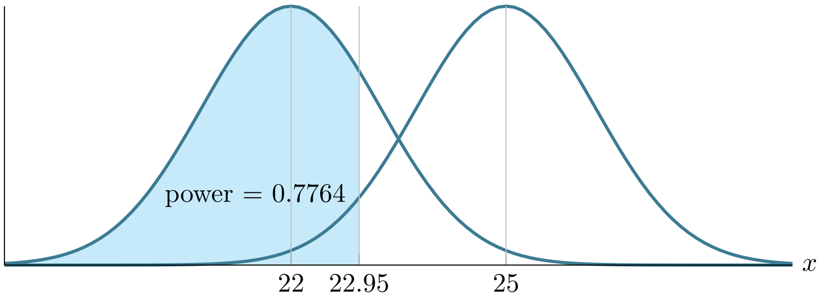


Figure 14

(e) Using your computation of the Type II error rate of the test, find the power if $\mu = 20$.

Solution Power $= 1 - \beta = 0.9909$.

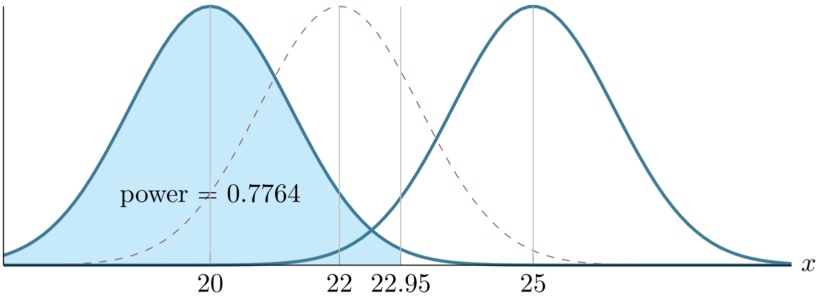


Figure 15

4. If σ is unknown and we must estimate it, we generally use the t statistic. Suppose the t statistic for the test

$$\begin{aligned} H_0 &: \mu = 25 \\ H_a &: \mu < 25 \end{aligned}$$

based on $N = 10$ observations, has a value of $t = -4.45$.

(a) What are the appropriate degrees of freedom for this t test?

Solution Degrees of freedom is just allocation in a loose sense: in this case, we have 10 observations and once we have allocated 9 of these observations with a place number, there is no freedom for the 10th allocation. Hence, we have 9 degrees of freedom.

(b) Using the t-distribution Critical Values Table, find the p value resulting from the t test.

Solution Check table B at the back of the Agresti book, and you will see that $\mathbb{P}\left(|t|_{\text{df}=9} < 4.297\right) = 0.9980$, so this means that our p-value must be less than $(1 - 0.9980)/2 = 0.001$ as our statistic is more extreme than the critical value. This can be seen in Figure 16.

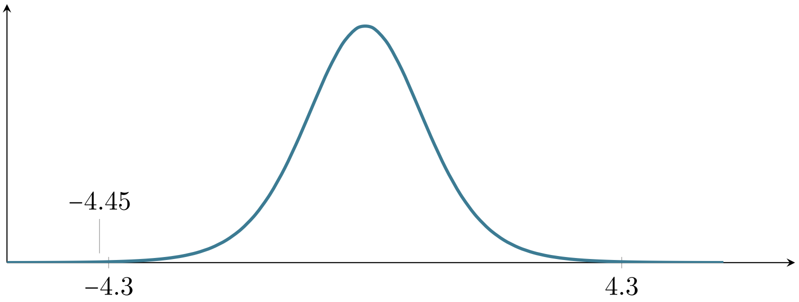


Figure 16: For the t-distribution with 9 degrees of freedom, a test statistic of -4.45 is significant at every level as the (two-sided) critical value for $\alpha = 0.002$ is -4.297.

1.4 Homework 4

1. The data in the following table consists of scores of 20 children with neurological disorders. The two tests, A and B, are both tests of spatial perception ([Efron and Tibshirani, 1993]). The scales of the two tests are the same: they both have a range of 0 to 50. Psychologists often use two versions of a test that they hope measure the same thing, because this gives them the opportunity to test people multiple times. Psychologists often want to compare scores over time, but if researchers give the same test twice, any improvements may be due simply to memory of the previous test. If we have two tests, one thing we might want to know is whether the means of the scores are different.

Person	Test A	Test B	Person	Test A	Test B
1	48	42	11	45	34
2	36	33	12	14	22
3	20	16	13	6	7
4	29	39	14	0	15
5	42	38	15	33	34
6	42	36	16	28	29
7	20	15	17	34	41
8	42	33	18	4	13
9	22	20	19	32	38
10	41	43	20	24	25

Table 8

By computing the differences between Test A and Test B for each person, we can see that the mean difference between Test A and B (Test A minus Test B) equals −0.55, and that the standard deviation of the difference is 6.886.

(a) Suppose that we would like to show that the scores on Test A and B in the population are, on average, the same. A naive researcher would like to use a significance test for this. Give the hypotheses that the researcher will use, H_0 and H_A; describe the meaning of the parameters you use in your hypotheses.

Solution The information that we are given is that $\bar{d} = -0.55$ and $s_d = 6.886$, where $D = A - B$ is the random variable denoting the difference between the random variables A and B which are the scores on test A and B, respectively. Also, $n = 20$ and each test has a maximum score of 50. The naive researcher might be testing:

$$\begin{array}{rcl} H_0 & : & \mu_d = 0 \\ H_A & : & \mu_d \neq 0. \end{array}$$

It is crucial to understand why we have selected a two-tailed significance test here! Both tests are supposedly equal in their outcomes, so we do not assume that either test provides a different answer. Moreover they both have the same maximum number of questions, so we do not need to think about scaling or standardising.

The hypothesis is that the mean difference in the population is the same $\mu_D = \mu_A - \mu_B$ meaning that the tests do measure the same thing. The alternative is that they are not the same and they measure two different things.

(b) Use a graphical method to check for outliers in the data, and to check that the distribution of difference scores is not too skewed. Describe any gross deviations from normality. Also, describe how you find the estimate of variance for the test.

Solution The box plot for this data set is given in Figure 17, which displays the lack of outliers and minimal skewing to the right. Based on the box plot, we do not believe that there are any gross deviations from normality in the population[3]. You estimate the variance for this test with the squared standard error:

$$\sigma^2 \sim \frac{s_d^2}{n} = \frac{6.886^2}{20} \approx 2.371. \tag{267}$$

[3] Review §6.2 (Agresti) on page 163-4 where it discusses robustness for violations of normality assumption.

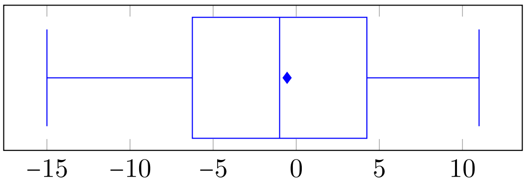


Figure 17: This box plot represents the difference in scores $D = A - B$ for the data given in Table 8. The five-number summary is −15, −6.25, −1, 4.25, 11 and the IQR is 10.5, thus there are no outliers. The blue diamond displays the sample mean −0.55, which indicates a slight skewing to the right.

(c) Perform the test. Is the null hypothesis rejected at a 5% significance level?

Solution We know the usual formula for the t statistic:

$$t = \frac{\text{Estimate of parameter} - \text{Null hypothesis value of parameter}}{\text{Standard error of estimate}} \tag{268}$$

$$= \frac{\bar{d} - \mu_{d0}}{se} = \frac{-0.55 - 0}{\sqrt{2.371}} \approx -0.357 \tag{269}$$

We can now calculate the p-value (recall that df$= n - 1 = 19$):

$$p = \mathbb{P}\left(|t_{19}| > 0.357\right) = 0.725026. \tag{270}$$

Without a calculating tool, you won't be able to find such a p-value. If we have a look at table B, we note the critical t-value of ±2.093 for $\alpha = 0.05$ which is far greater than our calculated statistic. This is equivalent to finding a $p > \alpha$, so we do not have significant findings and thus do not reject H_0.

(d) Why is a significance test not an appropriate method for demonstrating that the means are approximately the same?

Solution It is evident from Figure 18 that the individual distributions are skewed, and it may be more appropriate to consider the median than the mean in this case. In Stats2, you will learn about the Kruskal-Wallis and Wilcoxon tests which allow for significance testing between medians (non-parametric rank/signed-rank test). In our case, our n is too small for us to be sure that the skewing does not violate the robustness of the two-sided t-test for significance. In light of this, we would be better off looking at the confidence interval for the difference in means[4].

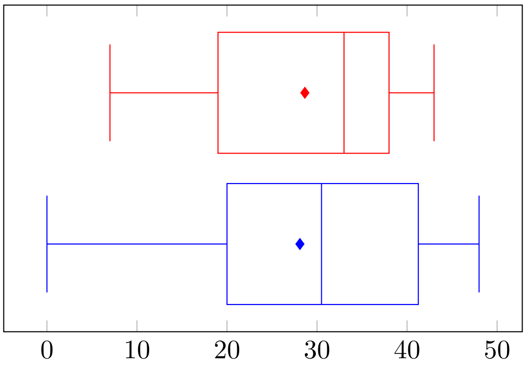


Figure 18: These box plots represent the scores for the data given in Table 8. The five-number summaries are (0, 20, 30.5, 41.25, 48) (IQR = 21.25) and (7, 19, 33, 38, 43) (IQR = 19), respectively, thus there are no outliers. The diamonds display the sample means 28.1 and 28.65, which indicate skewing to the left for both distributions.

[4] The 95% CIs for median of Test A and Test B nearly agree, however their 95% CIs for mean of Test A and B are quite different and the CI for Test B can be wholly contained in the CI for A. Further reading on this topic: `http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTest-Means-Proportions/BS704_HypothesisTest-Means-Proportions_print.html`, `https://www.ucl.ac.uk/child-health/short-courses-events/about-statistical-courses/research-methods-and-statistics/chapter-8-content-8`, and `http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/tests-of-means/why-use-paired-t/`.

(e) Compute a 95%-confidence interval for the mean difference between Test A and Test B.

Solution

$$\text{95\% CI for } \mu : \qquad \bar{d} \pm t^*_{1-\alpha/2,\,\mathrm{df}} \times se = -0.55 \pm 2.093 \times \sqrt{2.371} \tag{271}$$

$$= (-3.77, 2.67)\,. \tag{272}$$

2. In this exercise, all the situations require inference about means. Determine for every situation whether it requires (1) a one sample test, (2) a matched pairs test, or (3) a test for two independent samples. The procedures in §7.1 of the book are (1) and (2). In §7.2, the procedures for (3) are discussed.

(a) An education researcher is interested in the best way to organize a calculus book for primary school. She would like to determine which is more effective: including assignments before, or alternatively after, introducing a new subject. She creates two versions of a section of the book: one version with assignments first, and one version with assignments second. The two versions of text are used for teaching different groups of children (one group gets one version, the other gets the other version). After the class, the students' scores on a calculus test are compared.

Solution Go to page 193 of your Agresti book, and you will see this exact example:

> *Suppose you plan to use an experimental study to analyze whether a tutoring program improves mathematical understanding. One study design administers an exam on math concepts to a sample of students both before and after they go through the program. The sample of exam scores before the program and the sample of exam scores after the program are then dependent, because each sample has the same subjects. Another study design randomly splits a class of students into two groups, one of which takes the tutoring program (the experimental group) and one of which does not (the control group). After the course, both groups take the math concepts exam, and mean scores are compared. The two samples are then independent, because they contain different subjects without a matching between samples.*

We use (3) a test for two indepdent samples.

(b) A different researcher approaches the problem from a different angle. She makes two versions of a text on each of two different calculus topics. The two versions for each topic are one with assignments first, and one with assignments second. The researcher samples some children, and these children are taught both topics. Each child gets one topic with the "assignments first" version, and one topic with the "assignments second" version. For every child, the test scores of the two versions are compared to determine whether one version worked better than another.

Solution Akin to part a, this is (2) a matched pairs test.

(c) To evaluate a new analysis technique, a chemical engineer uses a reference sample with a known concentration of a certain chemical. Using this new analysis technique, the engineer takes 20 measurements of the concentration in the reference sample. Because the analysis technique is not perfect, there is some variance in the measurements. The researcher wishes to compare the mean measured concentration to the known value.

Solution This requires (1) a one sample test.

(d) Another chemical engineer wants to evaluate the same new analysis technique. She has no reference sample, but she wishes to compare the new technique to an old technique she is more familiar with. She wants to test whether the new and the old methods agree. She has 10 samples with unknown concentrations of a certain chemical. For each sample, she uses the new technique and the old technique.

Solution This requires (3) a test for two independent samples.

3. A psychologist wants to determine whether there is a difference between the academic habits and attitudes of first-year and second-year psychology students at the RuG. He uses the Questionnaire of Study Habits and Attitude (QSHA), a survey that measures the self-reported motivation, study habits and attitude of students towards their study. He takes a random sample of 25 first-year students and 29 second-year students, and administers the QSHA to everyone. The following table shows the mean and standard deviation of the scores on the QSHA in both samples.

Group	n	$\bar{x}$	s
First-years	25	127	4.89
Second-years	29	131	5.65

(a) Perform a t test while assuming different population variances to determine whether first-year psychology students score, on average, different from second-year-psychology students on the QHSA. (Use df = $\min(n_1 - 1, n_2 - 1)$; form the hypotheses, compute the statistic with the correct number of degrees of freedom, compute the p-value based on $\alpha = 0.05$).

Solution In our case, df = $\min(24, 28) = 24$ and we formulate our hypotheses as:

$$\begin{aligned} H_0 &: \mu = \mu_2 - \mu_1 = 0 \\ H_A &: \mu \neq 0 \end{aligned}$$

So, we are using a two-tailed t-test for two independent samples without assuming homoskedasticity. Using the formula for unpooled variance (check page 199 of Agresti):

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{4.89^2}{25} + \frac{5.65^2}{29}} \approx 1.434 \tag{273}$$

$$\implies t = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1)}{se} = \frac{(131 - 127) - (0)}{1.434} \approx 2.789 \sim t(24). \tag{274}$$

We can see from table B that a t statistic of 2.789 is significant up to a 1% level with 24 degrees of freedom, which already tells us that $p < 0.05$. Indeed, the critical value for 24 degrees of freedom is 2.063899 which is less than our statistic. We can calculate the p-value more precisely as

$$p = \mathbb{P}(|t_{24}| > 2.789) = 0.010186. \tag{275}$$

The null hypothesis is rejected at 5% significance level, thus there must be a difference in academic habits and attitudes of first-year and second-year psychology students at the RuG.

(b) Compute the 95%-confidence interval for the difference in QSHA between first-year and second-year psychology students.

Solution Using the formula on page 199 (Agresti):

$$95\% \text{ CI for } \mu: \quad (\bar{x}_2 - \bar{x}_1) \pm t^*_{1-\alpha/2, df} \times se = 4 \pm 2.064 \times 1.434 \tag{276}$$

$$= (1.04, 6.96). \tag{277}$$

Evidently, $\mu_0 = 0$ is **not** contained in the 95% CI, which agrees with the inference stemming from our p-value calculated in part a.

(c) What do you conclude about the difference in QHSA scores between first and second year students on the basis of your previous answers?

Solution The p-value indicates that we have significant results, and the 95% CI suggests that a second-year psychology student scores higher on the QSHA, in general, than a first-year student.

4. Does cocaine use by pregnant women lead to low birth-weight in their newborns? In order to answer this question, the birth weights of babies of women that have been tested positive for cocaine on a drug screening test may be compared with the birth weights of babies of women that have been tested negative or that have not been tested. The latter group is labeled “Not Positive”. The following table shows the descriptive statistics from the two samples.

Group	n	$\bar{x}$	s
Positive	134	2733g	599g
Not Positive	5974	3118g	672g

(a) Determine the hypotheses and perform the appropriate significance test to explore possible differences in birth weight.

Solution First we formalise our hypotheses:

$$\begin{aligned} H_0 &: \quad \mu = \mu_{NP} - \mu_P = 0 \\ H_A &: \quad \mu > 0. \end{aligned}$$

That is, we assume that there is no difference in birth weights of babies whose mothers tested positive (P) and not positive (NP), or otherwise that the former group leads to lower birth weights.

We will use a right-tailed t-test for two independent samples and investigate our summary statistics for equal variances assumed and not assumed.

$$s = \sqrt{\frac{(n_P - 1)\, s_P^2 + (n_{NP} - 1)\, s_{NP}^2}{(n_P - 1) + (n_{NP} - 1)}} \tag{278}$$

$$= \sqrt{\frac{(134 - 1) \times 599^2 + (5974 - 1) \times 672^2}{(134 - 1) + (5974 - 1)}} \tag{279}$$

$$= \sqrt{\frac{2745031765}{6106}} \approx \sqrt{449563.01425} \approx 670.495. \tag{280}$$

$$\implies se = s\sqrt{\frac{1}{n_P} + \frac{1}{n_{NP}}} = 670.495\sqrt{\frac{1}{134} + \frac{1}{5974}} \approx 58.568. \tag{281}$$

$$\implies t = \frac{(\bar{x}_{NP} - \bar{x}_P) - (\mu_{NP} - \mu_P)}{se} = \frac{385 - 0}{58.568} \approx 6.574. \tag{282}$$

If we assume equal variances in the population, then the standard error of our estimate $\bar{x}$ for μ is $\tilde{5}$8.6g which yields a t statistic of $\tilde{6}$.57 indicating that there is a significant difference in means between the groups. We draw this inference from table B, which advises us that a statistic more extreme than 3.174 is significant at any α level (df = $n_P + n_{NP} - 2$). Suppose that we do not assume homoskedasticity, then

$$se = \sqrt{\frac{s_P^2}{n_P} + \frac{s_{NP}^2}{n_{NP}}} = \sqrt{\frac{599^2}{134} + \frac{672^2}{5974}} \approx \sqrt{2753.211} \approx 52.47105. \tag{283}$$

$$\implies t = \frac{385 - 0}{52.47105} \approx 7.3374. \tag{284}$$

If we assume unequal variances in the population, then the standard error of our estimate $\bar{x}$ for μ is $\tilde{5}$2.5g which yields a t statistic of $\tilde{7}$.34 indicating that there is a significant difference in means between the groups (df = $\min\{n_P - 1, n_{NP} - 1\}$, or see the previous section on unpooled variance for the exact calculation). The Agresti book seems to suggest that you exclusively do not assume equal variances, which is not always the case for unequal sample sizes. There are official ways of confirming equal variance in the population, however a general rule of thumb is

$$\max\left\{s_1^2, s_2^2\right\} \leq 4 \times \min\left\{s_1^2, s_2^2\right\} \tag{285}$$

$$\implies \frac{\max\left\{s_1^2, s_2^2\right\}}{\min\left\{s_1^2, s_2^2\right\}} \leq 4 \tag{286}$$

$$\implies \left(\frac{\max\left\{s_1, s_2\right\}}{\min\left\{s_1, s_2\right\}}\right)^2 \leq 2^2. \tag{287}$$

This says that the largest sample standard deviation must be smaller than 2 times the smallest sample standard deviation. In other words, the ratio of variances must be bounded by 4 (see the section above on unpooled variance). In our case, we have that the ratio of sample variances is 0.7945, which is certainly less than 4 and suggests that we were correct in initially assuming

that the population variances are similar. In Stats 2, you will learn more official ways of testing for equal variances in the population, e.g. F-test.
We reject the null hypothesis and conclude that a baby whose mother tested positive for cocaine whilst pregnant will have a lower birth weight than a baby whose mother did not.

(b) Find the 95%-confidence interval for the mean difference in birth weight.

Solution Keeping in mind that we previously asserted equal variances in the population, i.e. df $= n_P + n_{NP} - 2 = 6106$ and $se = 58.6$g, then

$$95\% \text{ CI for } \mu: \qquad \bar{x}_{NP} - \bar{x}_P \pm t^*_{1-\alpha/2,df} \times se = 385 \pm 1.96 \times 58.6 \tag{288}$$

$$= (270.144, 499.856)\,. \tag{289}$$

$$\Longrightarrow \mathbb{P}\,(270.144 \leq \mu_{NP} - \mu_P \leq 499.856) \approx 0.95. \tag{290}$$

Given our sample, there is a probability of 0.95 that the true mean difference in groups is more than $\tilde{2}$70g but less that $\tilde{5}$00g.

(c) What do you think can be concluded from the study? Which result do you find more informative for drawing conclusions, the significance test or the confidence interval?

Solution See the ends of part a and b.

5. We discussed the study of the use of Cognitive Behavior Therapy among children with anxiety disorders ([Nauta et al., 2003]) in previous practicum classes in Statistics 1A. In their study, CBT was evaluated by measuring the degree of anxiety of each child before, directly after, 3 months after, and 15 months after the therapy. Anxiety was measured using interviews of both the child and one of the parents. For every child, then, there is a measurement of anxiety as reported by the child, and a measurement of anxiety as reported by the parents. Because we are interested primarily in the anxiety experience of the child, we will use the reported anxiety of the child as our measure of anxiety. Possible anxiety scores are whole numbers ranging from 0 (no anxiety) to 100 (maximum anxiety). Over 10 years ago, the anxiety questionnaire was validated among Dutch children with an anxiety disorder. The mean score on anxiety of children with an anxiety disorder (prior to treatment) was 38. The first question is whether children diagnosed with an anxiety disorder now have a different mean score from 10 years ago. Our second question is whether CBT helps children with anxiety disorders in the short term. Thus, we will compare anxiety scores before and immediately after the therapy.
Open in SPSS the file `4 - Anxiety Therapy.sav`.
The variable `anxc1` is the child's anxiety measurement before the therapy; the variable `anxc2` is the child's anxiety measurement immediately after the therapy.
The first question is whether children with anxiety disorders have a different mean score on the anxiety questionnaire than 10 years ago. To answer this question, we must examine the scores of the children on the anxiety questionnaire pre-CBT (`anxc1`).

(a) What are the mean and the standard deviation of the scores on `anxc1`? On how many children's scores are this mean and standard deviation based?

Solution See Figure 19.

(b) Make a histogram of the scores on `anxc1`. How would you describe this distribution? Sketch the histogram or attach a printout to your homework.

Solution See Figure 19.

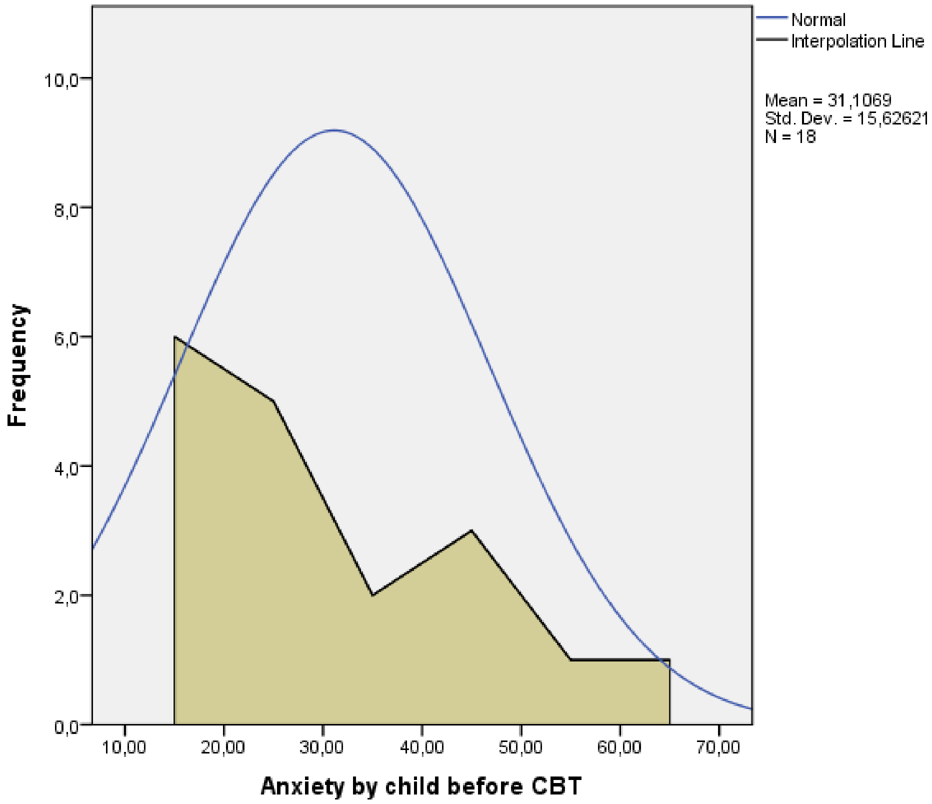


Figure 19: This (centred) histogram of the data from `anxc1` shows that it is a right-skewed distribution, with $\bar{x} = 31.1069$, $s = 15.62621$, and $n = 18$.

1.5 Homework 5

1. Repeat the comparison of the QSHA scores of the first- and second-year psychology students in the homework Exercise 3 from last week, only this time use the pooled-variance t test.

   (a) Perform the significance test and determine the p value.

Solution Begin by computing the pooled variance:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)} \tag{291}$$

$$= \frac{24 \times 4.89^2 + 28 \times 5.65^2}{24 + 28} = \frac{1467.7204}{52} \approx 28.2254. \tag{292}$$

So our pooled standard deviation is $\sqrt{28.2254} \approx 5.313$ allowing us to compute the standard error:

$$se = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 5.313 \times \sqrt{\frac{1}{25} + \frac{1}{29}} \approx 1.45. \tag{293}$$

$$\implies t = \frac{131 - 127}{se} = \frac{4}{1.45} \approx 2.759 \sim t(52). \tag{294}$$

We can see from table B that a t-statistic which is more extreme than 2.009 is significant at the 5% level (noting that our null hypothesis is two-sided) for 50 degrees of freedom. In our case we have $n_1 + n_2 - 2 = 52$ degrees of freedom, so it is not possible to directly read the critical t-value but we can assume that 2.759 is a significant t-value. Using an online calculator, $t^*(52) = 2.006647$ meaning that our p-value must be less than 0.05 as it is less than our statistic. Indeed,

$$p = \mathbb{P}(|t_{52}| > 1.952) = 0.007985. \tag{295}$$

   (b) Compute the 95% confidence interval for the difference between the QHSA scores of first- and second-year psychology students.

Solution In the previous part, we calculated the critical t-value for 52 degrees of freedom as 2.006647, so our 95% CI for μ_d is

$$\bar{x}_1 - \bar{x}_2 \pm t^*_{52} \times se = 4 \pm 2.006647 \times 1.45 = 4 \pm 2.91 = (1.09, 6.91)\,. \tag{296}$$

The probability that the true value of the parameter for the difference in means in the population lies between 1.09 and 6.91 is 0.950.

(c) Compare your results with the results in homework Exercise 3 of last week. How well do the results of the different two-sample t procedures correspond?

Solution The samples are of similar (not equal) size, which accounts for how similar the Standard Errors are (1.434 for unpooled v.s. 1.45 for pooled). The main difference is the degrees of freedom used for the test: in the previous homework, we use 24 which resulted in a fatter tailed distribution, larger critical value and thus larger margin of error for the confidence interval.

2. Due to a shooting incident in a Rotterdam supermarket where a member of the supermarket staff was shot, the supermarket wanted to determine whether the public thought supermarkets were safe. The following question was asked: "Supermarkets should close at night for security reasons". Of the 500 respondents, 156 answered with "Yes", and 344 with "No".

(a) Compute the sample proportion that said "Yes", and estimate the standard deviation of that proportion.

Solution The sample proportion that said "yes" is the number of people who said "yes" divided by the total number of people in the sample:

$$\hat{\pi} = \frac{156}{500} = 0.312. \tag{297}$$

The estimate for the standard deviation of the proportion (or standard error of the estimate) is given by the square root of the proportion that said "yes" times the proportion that said "no" divided by the sample size:

$$se = \sqrt{\frac{\hat{\pi}\,(1 - \hat{\pi})}{n}} = \sqrt{\frac{0.312 \times 0.688}{500}} = \sqrt{0.000429312} \approx 0.02072. \tag{298}$$

(b) Use your answer to the previous question to find a 95% confidence interval for the proportion people that advocate closing supermarkets at night.

Solution

$$\begin{aligned} \text{95\% CI for } \pi\text{:}\ \hat{\pi} \pm 1.96 \times se &= 0.312 \pm 1.96 \times 0.02072 && (299)\\ &= (0.2714, 0.3526)\,. && (300)\end{aligned}$$

The true proportion of the population that believe supermarkets should be closed at night for security reasons lies between 0.2714 and 0.3526 with probability of 0.95. As this does not constitute a majority vote, we decline to close supermarkets at night for security reasons.

3. A matched-pairs experiment is done that compared the taste of "regular" fresh coffee with the taste of Senseo coffee. Every participant tasted two unmarked cups of coffee, one of each kind of coffee. The tastings were done in a random order. Each participant was asked to tell which one of the two cups they preferred. Of the 60 participants, 26 preferred regular coffee, and 34 preferred the Senseo coffee. Assume that p is the chance that a randomly taken person prefers "regular" fresh coffee over Senseo.

(a) Test the claim that a majority of people prefer regular coffee. Note the z statistic and the p-value.

Solution We will consider rejecting the assumption that there is no preference in favour of a majority of people preferring "regular" fresh coffee over Senseo if the p-value is less than our prescribed α. We formulate our hypotheses:

$$\begin{aligned} H_0 &: \quad p = 0.5 \\ H_a &: \quad p > 0.5. \end{aligned}$$

Next, we calculate the standard error of the estimate under the null hypothesis in a sample size of 60:

$$se_0 = \sqrt{\frac{p_0\left(1 - p_0\right)}{n}} = \sqrt{\frac{0.5^2}{60}} \approx 0.065. \tag{301}$$

Finally, we calculate our z-value:

$$z = \frac{\hat{p} - p_0}{se_0} = \frac{{}^{26}\!/_{60} - 0.5}{0.065} \approx -1.026. \tag{302}$$

Visualise this z-value on the standard normal distribution, and recall that $\tilde{6}8\%$ of the population data lies between -1 and 1. A two-sided p-value would be roughly $(100 - 68)\% = 32\%$, so the right-tailed p-value would be roughly $(50 + 68/2)\% = 84\%$ and the left-tailed would be roughly $(100 - 84)\% = 16\%$. This allows us to already deduce that we will not be rejecting our null hypothesis. Keeping this in mind, we calculate the p-value as

$$p = \mathbb{P}\left(Z > -1.026\right) \tag{303}$$

As the standard normal distribution is symmetric about the mean of zero, we can solve this as

$$= \mathbb{P}\left(Z < 1.026\right) = 0.847554, \tag{304}$$

which agrees with our theoretical calculation.

(b) Is this result significant, using $\alpha = 0.05$?

Solution $p > \alpha$, so yeah.

(c) Compute a 90% confidence interval for p.

Solution

$$\begin{aligned} \text{90\% CI for } p\text{: } \hat{p} \pm 1.645 \times se = {}^{26}\!/_{60} \pm 1.645\sqrt{\frac{26 \times 34}{60^3}} \quad & (305) \\ \approx 0.4\bar{3} \pm 1.645 \times 0.064 \quad & (306) \\ = \left(0.3281, 0.5386\right). \quad & (307) \end{aligned}$$

The true value of the chance that a randomly selected person from the population that prefers "regular" fresh coffee over Senseo lies between 0.3281 and 0.5386 with probability of 0.9. Note that $p_0 = 0.5$ is contained in the 90% CI.

(d) Using your two answers above, form your conclusion about whether a majority of people prefer fresh coffee.

Solution It appears that they do not care either way, as we do not reject the null hypothesis that the chance a randomly selected person prefers "regular" fresh coffee over Senseo is 50%.

4. Suppose that in a study on depression among children, the researchers wanted to show a relationship between the depression among children and the composition of the family. The researchers made a distinction between one- and two-parent families. The following table shows the results of the study. The variable X refers to the number of children with a diagnosis of depression, and n refers to the total number of families of that type.

Family type	n	X
One-parent families	74	6
Two-parent families	226	13

(a) For the one-parent families, compute the proportion of children suffering from depression.

Solution Let π_1 denote the proportion of children in one-parent families suffering from depression (note that this is conditional on the number of parents!!), so $\hat{\pi}_1 = {}^6\!/_{74} = 0.\overline{081} \approx 0.0811$. That means that the sample proportion of children in one-parent families not suffering from depression is $(1 - 0.0811) = 0.\overline{918} \approx 0.9189$.

(b) For the two-parent families, compute the proportion of children suffering from depression.

Solution Let π_2 denote the proportion of children in two-parent families suffering from depression, so $\hat{\pi}_2 = {}^{13}\!/_{226} \approx 0.0575$. That means that the sample proportion of children in two-parent families not suffering from depression is $(1 - 0.0575) \approx 0.9425$.

(c) Compute a 95%-confidence interval for the difference between these proportions.

Solution 95% CI for $\pi_1 - \pi_2$:

$$\hat{\pi}_1 - \hat{\pi}_2 \pm 1.96 \times \sqrt{\frac{\hat{\pi}_1 \times (1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2 \times (1-\hat{\pi}_2)}{n_2}} = 0.02356 \pm 1.96 \times \sqrt{0.001007 + 0.015488} \quad (308)$$

$$= 0.02356 \pm 1.96 \times 0.12843 \quad (309)$$

$$= (-0.228, 0.275)\,. \quad (310)$$

Note that we used unpooled standard error? This is because we do not assume homoskedasticity in the population as it is not plausible.

(d) On the basis of your computations above, what are your conclusions regarding depression and family composition?

Solution The true value of the difference between the proportions lies near and about zero, and ranges from −0.228 to 0.275 with probability 0.95. This allows us to conclude that there is no real difference in the chances a child will suffer from depression whether they have one- or two-parent families. You could loosely conclude that a child who has a one-parent family is not more likely to develop depression than a child who has two parents, and thus single parents do a “good job” at ensuring their childrens’ happiness.

5. Eisses and Kluiter’s did a study on depression symptoms among nursing home residents [Eisses and Kluiter, 2002]. The researchers found that the percentage of elderly people undergoing treatment for depression is lower in Drenthe than in the rest of the Netherlands. It was suspected that the staff in Drenthe nursing homes needed better training to detect depression symptoms, so a study was conducted to evaluate the effect of training. Six nursing homes took part in the study. The staff from the nursing homes were assigned to either a control condition (no training) or an experimental condition (training to detect depression symptoms). Three nursing homes were assigned to the control condition, and three to the experimental condition.
Before and after the training, the residents of the nursing home were interviewed by an independent clinical psychologist. Based on the reports of individual residents, the psychologist determined whether that resident was suffering from depression. Eisses and Kluiter used the independent psychologist’s diagnoses to compare to the staff’s diagnoses, but in this practicum we will only analyze the report of the independent psychologist.
Open the file `5 - Depression Diagnosis.sav`.

Variable	Description
`participant`	Participant number
`exp`	1=control group, 2=experimental group
`depdiaB`	depression diagnosis before staff training: 1=yes, 2=no, 9=missing
`depdiaA`	depression diagnosis after staff training: 1=yes, 2=no, 9=missing

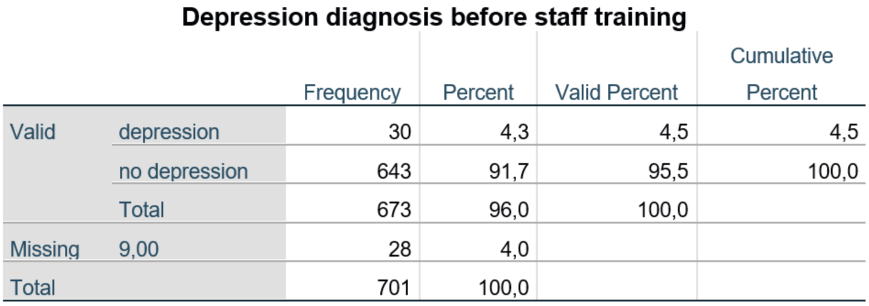
Some data is missing, due to the fact that the residents moved between interviews.
One question of interest is: what proportion of residents of nursing homes in Drenthe suffer from depression?

(a) To answer this question would you use the scores on the pre-training measure, the post-training measure, or both? Explain your answer.

Solution The scores on the pre-training measure would be most useful in answering this question, as we do not want that the additional staff training provides a confounding variable for us to solve. By doing so, we negate the allocation to control or experimental group and can simply estimate the proportion of residents of nursing homes in Drenthe suffering from depression.

(b) Compute the proportion of residents suffering from depression before the training. Look closely at the frequency table. How many residents are diagnosed as depressed before the training? What percent is depressed?

Solution The first thing to notice is that 28 of 701 participants have missing data, so keeping that adjustment in mind the sample proportion is 0.045 even though there are 30 residents with depression in total. With respect to the total 4.3% have depression, and when accounting for the missing data we have 4.5%. The frequency table is given in Figure 20.



Depression diagnosis before staff training

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	depression	30	4,3	4,5	4,5
	no depression	643	91,7	95,5	100,0
	Total	673	96,0	100,0	
Missing	9,00	28	4,0		
Total		701	100,0		

Figure 20: SPSS frequency table output for the proportion of residents of nursing homes in Drenthe suffering from depression prior to initiating staff training in detecting symptoms of depression.

These data will be used again in next week's practicum.

1.6 Homework 6

1. Appleton, French and Vanderpump [Appleton et al., 1996] reported the results of a large epidemiological study of women in Great Britain. In the beginning of the 1970s, many women in Great Britain were asked about their health and health-related habits. Twenty years later, the women were contacted again to determine who had lived and who died. The table below shows the results.

	Smoker	Non-smoker
Dead	139	230
Alive	443	502
Total	582	732

(a) Analyze the data, using plots and descriptive statistics. Use SPSS and/or R (hint: type in "?prop.test" in R and pay attention to the examples below) for your analysis. Describe what you did and sketch a copy of any informative plot you made.

Solution For obvious reasons, let's denote "success" as being alive, and failure as being "dead" with respect to table 8.6 from Agresti. So we have $\hat{\pi}_1 = 443/582 \approx 0.76117$, $\hat{\pi}_2 = 502/732 \approx 0.68579$, $\hat{\pi} = (443+502)/(582+732) = 945/1314 \approx 0.71918$ and $se_0 = \sqrt{\hat{\pi}\left(1-\hat{\pi}\right)\left({}^{1}\!/_{n_1} + {}^{1}\!/_{n_2}\right)} = \sqrt{0.71918\left(1-0.71918\right)\left({}^{1}\!/_{582} + {}^{1}\!/_{732}\right)} \approx 0.02496$, which results in the z-statistic $\left(\hat{\pi}_1 - \hat{\pi}_2\right)/se_0 = (0.76117 - 0.68579)/0.02496 \approx 3.02$. This z-statistic already indicates that $p < \alpha = 0.05$ and we will reject the null hypothesis ($H_0 : \pi_1 = \pi_2$) and squaring gives $z^2 = 9.12$. If you're using `R`, input the following code:

```
alive <- c(443,502)
total <- c(582,732)
prop.test(alive,total)

      2-sample test for equality of proportions with continuity correction

  data:  alive out of total
X-squared = 8.7515, df = 1, p-value = 0.003093
alternative hypothesis:  two.sided
95 percent confidence interval:
0.02555629 0.12519578
sample estimates:
prop 1 prop 2
0.7611684 0.6857923
```

You can use a bar graph of the sample percentages to display this information.

(b) Based on the data in the table, what do you conclude?

Solution We assume that the population variables π_1 and π_2 are independent, i.e. smoking does not negatively influence the health of a woman (during that time period), and would expect that $\hat{\pi}_1 \approx \hat{\pi}_2$ in any sample from said population.

	Smoker	Non-smoker
Dead	23.88%	31.42%
Alive	76.12%	68.58%
Total	100%	100%

2. In a longitudinal study of sexual orientation by [Golombok and Tasker, 1996], children of homosexual mothers were compared to children of heterosexual single mothers. The researchers asked the children about their sexual orientation when they were about 24 years old. The following table shows the children's sexual orientations as a function of the mothers' sexual orientations.

		Mother	
		Homosexual	Heterosexual
Child	Homosexual/bisexual	2 (Expected: $\underline{1.1\overline{1}}$)	0 (Expected: $\underline{0.8\overline{8}}$)
	Heterosexual	23 (Expected: $\underline{23.8\overline{8}}$)	20 (Expected: $\underline{19.1\overline{1}}$)
	Total	25	20

The SPSS file with these data is `6 - Golombok and Tasker.sav`.

(a) Compute the expected frequencies under the assumption that the two factors are independent with SPSS. Note them in the previous table.

Solution You can calculate the expected frequencies using Bayes' Rule for Independence. First we denote M_H as the event that the mother is homosexual and M_S as the event that the mother is heterosexual. Similarly, C_H denotes the event that the child is homo/bisexual and C_S denotes the event that the child is heterosexual. The marginal probabilities can be found by adding the row or column frequencies and dividing by the sample size, e.g. $\mathbb{P}\left(M_H\right) = 25/45$.

$$\mathbb{P}\left(A \cap B\right) = \mathbb{P}\left(A\right) \times \mathbb{P}\left(B\right). \tag{311}$$

The above formula is Bayes' Rule for Independence.

$$\implies \mathbb{P}(M_H \cap C_H) = \mathbb{P}(M_H) \times \mathbb{P}(C_H) = \frac{25}{45} \times \frac{2+0}{45} = \frac{50}{45^2} \approx 2.47\%. \tag{312}$$

Assuming that the two factors are independent, we find that there is a 2.47% chance that a homosexual mother will have a child that is homo/bisexual as well. We now know the probability of such an event, so multiplying it by the sample size will give the expected value.

$$\implies \mathbb{E}(M_H \cap C_H) = 45 \times 2.47\% = \frac{50}{45} = 1.1\bar{1}. \tag{313}$$

The expected number of homo/bisexual children who also have a homosexual mother is less than the observed value of 2, and is 1.11. The other probabilities are calculated in the following way:

$$\mathbb{P}(M_S \cap C_H) = \mathbb{P}(M_S) \times \mathbb{P}(C_H) = \frac{20}{45} \times \frac{2+0}{45} = \frac{40}{45^2} \approx 1.98\%. \tag{314}$$

$$\mathbb{P}(M_H \cap C_S) = \mathbb{P}(M_H) \times \mathbb{P}(C_S) = \frac{25}{45} \times \frac{23+20}{45} = \frac{1075}{45^2} \approx 53.09\%. \tag{315}$$

$$\mathbb{P}(M_S \cap C_S) = \mathbb{P}(M_S) \times \mathbb{P}(C_S) = \frac{20}{45} \times \frac{23+20}{45} = \frac{860}{45^2} \approx 42.47\%. \tag{316}$$

Which gives us the following expected values:

$$\implies \mathbb{E}(M_S \cap C_H) = 45 \times 1.98\% = \frac{40}{45} = 0.8\bar{8}. \tag{317}$$

$$\mathbb{E}(M_H \cap C_S) = 45 \times 53.09\% = \frac{1075}{45} = 23.8\bar{8}. \tag{318}$$

$$\mathbb{E}(M_S \cap C_S) = 45 \times 42.47\% = \frac{860}{45} = 19.1\bar{1}. \tag{319}$$

(b) Is a χ^2 test appropriate for these data? Why or why not?

Solution No. We need the expected frequencies in each cell to be at least 5, or at least 10 for df=1.

(c) Use SPSS to perform the χ^2 test and examine the output. Note it below. Notice that SPSS informs you about the number of cells with a frequency below five.

Solution

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{0,i} - f_{e,i})^2}{f_{e,i}} \tag{320}$$

$f_{0,i}$ denotes the *observed* frequency and $f_{e,i}$ denotes the *expected* frequency in cell i, where i is the number of rows r times the number of columns c so here $i = 2 \times 2 = 4$.

$$= \frac{(2-1.11)^2}{1.11} + \frac{(0-0.89)^2}{0.89} + \frac{(23-23.89)^2}{23.89} + \frac{(20-19.11)^2}{19.11} \tag{321}$$

$$= \frac{0.89^2}{1.11} + \frac{0.89^2}{0.89} + \frac{0.89^2}{23.89} + \frac{0.89^2}{19.11}. \tag{322}$$

$$\approx 1.678. \tag{323}$$

The p-value is calculated as follows:

$$p = \mathbb{P}\left(\chi_1^2 > 1.678\right) = 0.19519. \tag{324}$$

You can enter the following into R to achieve the same output:

```
x <- matrix(c(2,23,0,20), nrow=2)
chisq.test(x,correct=FALSE)

                    Pearson's Chi-squared test
```

```
data:  x
X-squared = 1.6744, df = 1, p-value = 0.1957
```

The use of `correct=FALSE` removes Yates' continuity correction.
As $df = (r-1) \times (c-1) = 1$, we can also look at z^2: first calculate the standard error under the null hypothesis of $\pi_1 = \pi_2$:

$$\bar{\pi} = \frac{2}{45} \approx 0.0\bar{4}. \tag{325}$$

$$\implies se_0^2 = \bar{\pi}\left(1-\bar{\pi}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right) \tag{326}$$

$$= 0.0\bar{4} \times 0.9\bar{5}\left(\frac{1}{25}+\frac{1}{20}\right) = \frac{2\times 43}{45^2} \times \frac{9}{100} = \frac{43}{11250} \approx 0.0038\bar{2}. \tag{327}$$

$$\implies z^2 = \frac{\left(\hat{\pi}_1 - \hat{\pi}_2\right)^2}{se_0^2} \tag{328}$$

$$= \frac{\left({}^2/_{25} - {}^0/_{20}\right)^2}{0.0038222} = \frac{72}{43} \approx 1.6744. \tag{329}$$

We would have a significant result if the statistic z^2 is more extreme than $1.96^2 = 3.8416$ for $\alpha = 0.05$, which it isn't. The p-value is calculated as follows:

$$p = \mathbb{P}\left(\chi_1^2 > 1.674\right) = 0.195724. \tag{330}$$

Neither p-value is significant at $\alpha = 0.05$ level - we do not reject the null hypothesis that they are independent

(d) The SPSS output also yields the result of Fisher's exact test. Note the result of this test, along with the relevant statistics.

Solution You can calculate this yourself using the hypergeometric distribution:

$$p_1 = \frac{2!43!25!20!}{2!0!23!20!45!} \approx 0.303030 \tag{331}$$

As 2 is the greatest possible number for this cell, keeping the marginal frequencies the same (other possibilities are 1 and 0), this is the right-tailed p-value. Next, you need to calculate the p-values for different extremes *without altering the marginal frequencies*, i.e. altering only the cell frequencies:

$$p_2 = \frac{2!43!25!20!}{1!1!24!19!45!} \approx 0.50505. \tag{332}$$

$$p_3 = \frac{2!43!25!20!}{0!2!25!18!45!} \approx 0.191919. \tag{333}$$

If you added all the p-values together, you would get the left-tailed p-value[5]. If you add together the p-values of all combinations that have lower probabilities than that of the observed data (so p-values less than $0.303030 = p_1$), you get the two-tailed p-value. The only p-value that is smaller than p_1 is p_3 (so for $\{(0,2),(25,18)\}$), hence we add p_1 and p_3 together to get the two-tailed p-value:

$$p = 0.30303 + 0.191919 = 0.494949. \tag{334}$$

You can do this in R:

```
x <- matrix(c(2,23,0,20), nrow=2)
exact2x2(x,tsmethod="minlike")

        Two-sided Fisher's Exact Test (usual method using minimum likelihood)

data:  x
p-value = 0.4949
alternative hypothesis:  true odds ratio is not equal to 1
95 percent confidence interval:
```

```
0.2316   Inf
sample estimates:
odds ratio
   Inf
```

Note that the confidence interval given is unbounded as it tends to ∞? This is because of the inclusion of the zero in our table: the odds ratio is $OR = (2/23)/(0/20) = (2/23) \times (20/0) = \infty$ and everyone knows you can't divide by zero. The computation of the 95% Confidence Interval for the odds ratio is as follows:

$$se_{\ln(OR)} = \sqrt{\frac{1}{2} + \frac{1}{0} + \frac{1}{23} + \frac{1}{20}} = \sqrt{\infty}. \tag{335}$$

95% CI for odds ratio:

$$\begin{aligned} \exp\left\{\ln(OR) \pm 1.96 \times se_{\ln(OR)}\right\} &= \exp\left\{\ln(\infty) \pm 1.96 \times \sqrt{\infty}\right\} && (336)\\ &= \left(\infty \times e^{-1.96\times\sqrt{\infty}}, \infty \times e^{1.96\times\sqrt{\infty}}\right) && (337)\\ &= (0.2316, \infty). && (338) \end{aligned}$$

This issue is not present when you don't have zero values in the table - a rearrangement of the columns does not aid us any further as we still have the zero in the odds ratio and the computation of the standard error.

(e) Fisher's exact test was developed by the famous statistician R.A. Fisher. It is designed to test the same null hypothesis as the χ^2 test, but it does not have the same problems with low cell counts that the χ^2 test has. In this class you will not learn the details of how to compute the test by hand, but notice that SPSS gives a p value for the test every time you compute a χ^2 test. What do you conclude on the basis of Fisher's exact test?

Solution Fisher's exact test starts by computing the probabilities of *all possible outcomes which do not alter the marginal frequencies* (p_1, p_2 and p_3), and adding all of these equates to the probability of all events, thus equals 1. For this reason, it is called *exact* as you can exactly calculate the probability of any event using the actual (hypergeometric) sampling distribution rather than a normal approximation. The sampling distribution is not symmetric for low n, meaning that left-, right-, and two-tailed p-values do not add up to 1.

(f) It is also true that confidence intervals computed using small sample sizes can be quite inaccurate. What implications does this have for designing studies using two-way data?

Solution They are inaccurate due the lack of central tendency and presence of skew, also having a zero in the computation of the odds ratio (and therefore confidence interval) is really hard to work with, as written above.

[5] You can play around with the following applet to see what I mean with left- and right-tailed p-values: `https://istats.shinyapps.io/FisherExact/`.