Review of Hypothesis Testing

PSYCH 710

September 11, 2017

0 Testing means

0.1 known population mean and variance

Consider the following scenario: The population of a small town was unknowingly exposed to an environmental toxin over the course of several years. There is a possibility that exposure to the toxin during pregnancy adversely affects cognitive development which eventually leads to lower verbal IQ. To determine if this has happened to the children of this town, you use a standardized test to measure verbal intelligence in a random sample of 20 children who were exposed to the toxin. Based on extensive testing with typical children, the mean and standard deviation of the scores in the population of typical children is 100 and 10, respectively. The mean score of your sample was 93. Given this information, should we conclude that our sample of twenty scores was drawn from the population of typical scores?

We will answer this question in what appears to be a roundabout way. We will start by assuming that there was no effect of the toxin, and therefore that our sample of scores was drawn from the population of typical scores. Hence, our **null hypothesis** is that the data were drawn from a population with a mean of 100 ($\mu = 100$) and a standard deviation of 10 ($\sigma = 10$). Next, we have to evaluate whether our observation (i.e., the sample mean is 93) is unusual, or unlikely, given the assumption that the null hypothesis is true. If the observation is unlikely, then we reject the null hypothesis ($\mu = 100$) in favor of the **alternative hypothesis** that $\mu \neq 100$. In your textbook, the null and alternative hypotheses often are displayed thusly:

$$\begin{aligned} H0 : \mu &= 100 \\ H1 : \mu &\neq 100 \end{aligned}$$

How can we determine if our observation is unlikely under the assumption that the null hypothesis is true? Recall that the mean ($\mu_{\bar{Y}}$) and standard deviation ($\sigma_{\bar{Y}}$) of the distribution of *means* – otherwise known as the **sampling distribution of the mean** – are related to the mean (μ) and standard deviation (σ) of the *individual scores* in the population by the equations

$$\mu_{\bar{Y}} = \mu \tag{1}$$

$$\sigma_{\bar{Y}} = \sigma/\sqrt{n} \tag{2}$$

where n is sample size. Therefore, for the current example

$$\mu_{\bar{Y}} = 100 \tag{3}$$

$$\sigma_{\bar{Y}} = 10/\sqrt{20} = 2.236 \tag{4}$$

Next, we note that if verbal IQ scores are distributed Normally, then the distribution of sample means will also be Normal. Moreover, the **Central Limit Theorem** implies the the distribution of sample means will

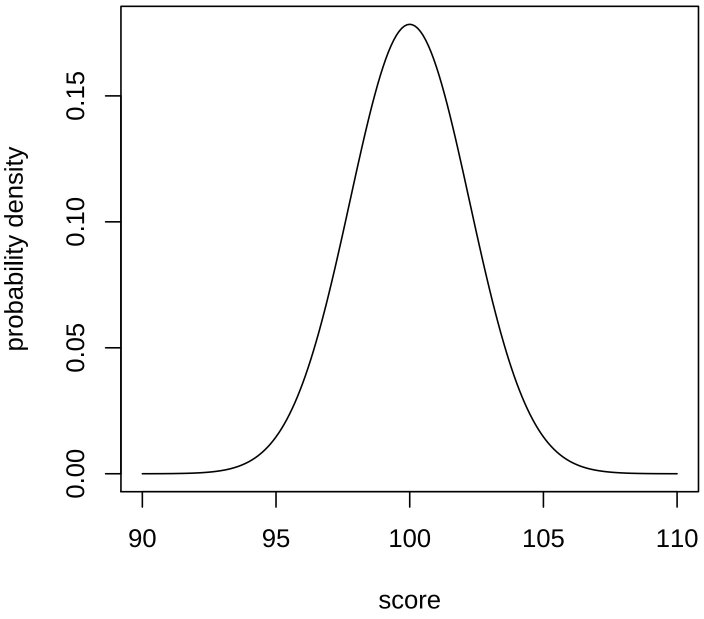


Figure 1: The theoretical sampling distribution of the mean.

be Normal *regardless* of the shape of the population distribution as long as sample size is sufficiently large[1]. We will assume, therefore, that the sample mean is a random variable that follows a Normal distribution.

Finally, we are in a position to determine if our observation ($\bar{Y} = 93$) is unusual (assuming the null hypothesis is true and in the absence of additional information). According to the null hypothesis, sample means will be distributed Normally with a mean of 100 and a standard deviation of 2.236. This distribution — actually, the probability density function — is shown Figure 1. Inspection of Figure 1 indicates that most of the scores drawn from this sampling distribution should be between 95 and 105. In other words, the probability of getting a score that is less than 95, or greater than 105, should be low. In fact, the probability of randomly selecting a score that is less than 95 is equal to the area under the part of the curve in Figure 1 that lies to the left of 95 (i.e., the lower tail of the function; see Figure 2). Similarly, the probability of selecting a score that is greater 105 is equal to the area under the curve that is to the right of 105 (i.e., the upper tail of the function). The joint probability – i.e., the probability of selecting a score that is less than 95 *or* greater than 105 – is equal to the sum of the two individual probabilities. In R[2], the probability of selecting a score that is less than 95 or greater than 105 is calculated using the following commands:

```
> pnorm(95,mean=100,sd=2.236,lower.tail=TRUE)

[1] 0.01267143 # prob of getting a score <= 95

> pnorm(105,mean=100,sd=2.236,lower.tail=FALSE)
```

[1] The key phrase here is "sufficiently large". For some types of population distributions, the sample size must be *very* large for the sampling distribution of the mean to follow a normal distribution. In other words, although the Central Limit Theorem is true, in some cases the sampling distribution of the mean can deviate significantly from normality even with sample sizes that are large relative to those typically used in psychology experiments (Wilcox, 2002).

[2] R is a software environment for statistical computing that is free and runs on several common operating systems. It can be downloaded at `http://cran.r-project.org`.

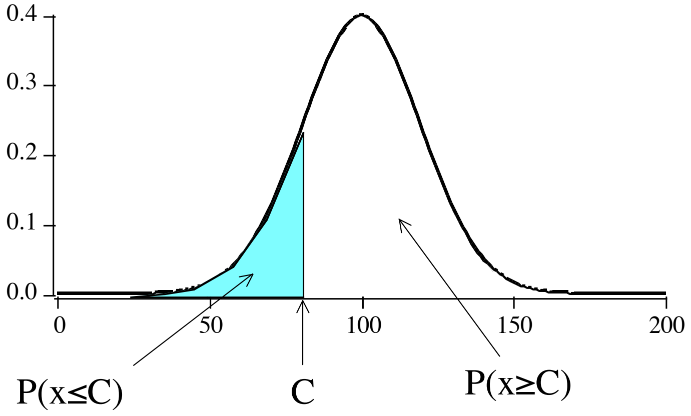


Figure 2: The probability of randomly selecting a value of x that is $\leq C$ – i.e., $P(x \leq C)$ – corresponds to the area under the probability density function that is to the left of C. $P(x \geq C)$ equals the area under the curve that is to the right of C.

```
[1] 0.01267143 # prob of getting a score >= 105

> 0.01267143+0.01267143

[1] 0.02534286 # prob of getting a score <= 95 OR >= 105
```

In other words, the probability of selecting a score that is less than 95 *or* greater than 105 is only 0.025. Because the total probability is 1, the probability of selecting a score that is between 95 and 105 is $1 - 0.025 = 0.975$. Given their relatively low probability, it is reasonable to assert that scores that fall below 95 or above 105 are unusual. By this criterion, our observed mean score of 93 is unusual, and we therefore reject the null hypothesis that $\mu = 100$ in favor of the alternative hypothesis $\mu \neq 100$.

In our example, we considered any score falling below 95 or above 105 to be unusual. It is important to note, however, that getting an unusual score does not necessarily mean that the null hypothesis is false. After all, unusual scores are possible even when the null hypothesis is true. In fact, we *expect* to obtain an unusual score with a probability of .025 when the null hypothesis is true. Hence, it is possible that our decision to reject the null hypothesis is incorrect. This type of error — rejecting the null hypothesis when it is true — is called a **Type I** error. The probability of making this error is determined by the criteria we use to define a score as unusual. In this case, we used criteria (i.e., below 95 or above 105) which would lead to a Type I error 2.5% of the time. The probability of making a Type I error is referred to as α (i.e., *alpha*), and so we would say that the Type I error rate, or α, is .025 for this statistical test.

It is standard practice in Psychology to set α to either 0.05 or 0.01. If we set $\alpha = .05$, then our decision criteria would be 95.6 and 104.4:

```
> qnorm(.025,mean=100,sd=2.236,lower.tail=TRUE) # qnorm... not pnorm

[1] 95.61752 # the probability of getting a score <= 95.6 is 0.025...

> pnorm(95.61752,mean=100,sd=2.236,lower.tail=TRUE)

[1] 0.02499999

> qnorm(.025,mean=100,sd=2.236,lower.tail=FALSE)

[1] 104.3825 # # the probability of getting a score >= 104.38 is 0.025...
```

```
> pnorm(104.3825,mean=100,sd=2.236,lower.tail=FALSE)

[1] 0.02499946

[1] 0.02499999 + 0.02499946 # the JOINT probability of getting a score <=95.6 OR >=104.38

[1] 0.04999945
```

Now, any score that is less than 95.6 or greater than 104.4 leads to the rejection of the null hypothesis. Notice that the range of acceptable scores — which do not cause us to reject the null hypothesis — is smaller than before. In other words, we are more likely to reject the null hypothesis even when it is true. This change in the Type I error makes sense because we increased α from .025 to .05. If we set $\alpha = .01$, then our decision criteria would be 94.8 and 105.2, and any score that is outside that range leads to the rejection of the null hypothesis. Now the Type I error rate, .01, is lower than before.

0.2 standardized scores & z tests

In the previous example, I used a computer to calculate the decision criteria for $\alpha = .05$ and $\alpha = .01$. Before computers were readily available — yes, there was such a time — people looked up the decision values in published tables. It would be impossible to publish tables for every possible case, and therefore people used a table of **standard normal deviates** or z scores. This section shows how to use such a table to conduct a z test.

Any value, Y, can be converted to a standard score using the formula

$$z = \frac{(Y - \mu)}{\sigma} \tag{5}$$

Notice that a z score equals the number of standard deviations that Y is from μ. When Y is drawn from a normal distribution, then z is distributed as a Normal variable with $\mu = 0$ and $\sigma = 1$.

We can convert our observed mean score from the previous example into a z score – $z = (93-100)/2.236 = -3.13$ – which implies that the observed mean is 3.13 standard deviations below the expected value of the mean. Now we want to know if our observed z score is unusual, given the assumption that the null hypothesis is true. If the null hypothesis is true, then z will be between ± 1.96 95% of the time and between ± 2.56 99% of the time. Therefore, using the criteria of ± 1.96 to reject the null hypothesis will yield a Type I error rate of 0.05, whereas the criteria of ± 2.56 corresponds to a Type I error rate of 0.01. Our observed z score falls outside of both sets of criteria, and so the null hypothesis is rejected. It used to be standard practice to indicate which α level was used by writing that the null hypothesis was rejected ($p < .05$) or ($p < .01$). Nowadays, scientists are encouraged to publish the exact p value for their observed statistic. In our case, the probability of drawing a z that was outside the range of ± 3.13 is 0.00175, and so we would report the statistical result by writing "the sample mean differed significantly from 100 ($z = -3.13$, $p = 0.00175$) and so the null hypothesis $\mu = 100$ was rejected".

0.3 t tests

In the previous example, we knew that $\sigma = 10$. But in most cases we we do not know σ, and therefore it must be estimated from the data:

$$\hat{\sigma} = s = \sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{(n-1)}}$$

The estimate, s can be used to calculate t, which is similar to a z score:

$$t = \frac{(\bar{Y} - \mu)}{(s/\sqrt{n})} \tag{6}$$

s^2 is an unbiased estimator of σ^2, which means that the *expected value* of s^2 is σ^2. However, the sampling distribution of s^2 is positively skewed, and therefore, $s^2 < \sigma^2$ more than 50% of the time. Therefore, t is *not* distributed as a standard normal deviate: in fact, **extreme values of t (i.e., much less or greater than zero) will occur more frequently than would be expected if t followed a normal distribution.** Think about the effect this will have on our decision to reject or reject the null hypothesis: extreme values of t are more likely than we would predict if we used the methods described in the previous section. Therefore, the inflation of t, if left uncorrected, would make it more likely that we would mistakenly conclude that our sample mean is unusual. What we need is a more accurate description of the distribution of t.

William Gosset showed that $\frac{(\bar{Y}-\mu)}{(s/\sqrt{n})}$ follows the so-called t distribution (Student, 1908). The t "distribution" is actually a family of symmetrical distributions that are centered on zero (Figure 4). Different members of the t family are distinguished by a parameter called the **degrees of freedom**, which can be thought of as the number of independent pieces of information that are used to estimate a parameter. In this case, n numbers are used to estimate s, but the same numbers also are used to estimate $\bar{Y}$. Because s depends on the prior estimation of $\bar{Y}$, the number of independent pieces of information relevant to the calculation of s is $n-1$. When the degrees of freedom is greater than about 20, the t distribution essentially is identical to the standard normal distribution (i.e., the z distribution). When there are fewer than 20 degrees of freedom, the t distribution has thicker tails than the z distribution (i.e., extreme values are more likely).

The logic of using the t distribution to evaluate the null hypothesis is similar to the logic used in the z test. The only difference is that the range of "typical" values of t generally will differ from ± 1.96 and ± 2.56. The so-called critical values of t are listed in Table A.1 in your textbook. For our sample, $df = 20 - 1 = 19$. If we set $\alpha = .05$, then we can use Table A.1 to determine that the critical values of t are ± 2.09. Let's assume that, for our data, $s = 14$. Therefore,

$$t = \frac{(93 - 100)}{14/\sqrt{20}} = -2.236$$

The observed value of t falls outside ± 2.09, so we reject the null hypothesis that $\mu = 100$ ($t = -2.236, p < .05$). As with the z test, modern practice is to report the exact p value, which is given by most statistical software. Therefore, we would write "the sample mean differed significantly from 100 ($t(19) = -2.236, p = 0.0376$)". The number in the parenthesis is the degrees of freedom.

The following command shows how a t test can be done in R, assuming that the data are stored in the variable `my.scores`, that the null hypothesis is $\mu = 100$, and that it is a two-tailed test (i.e., that the alternative hypothesis is $\mu \neq 100$):

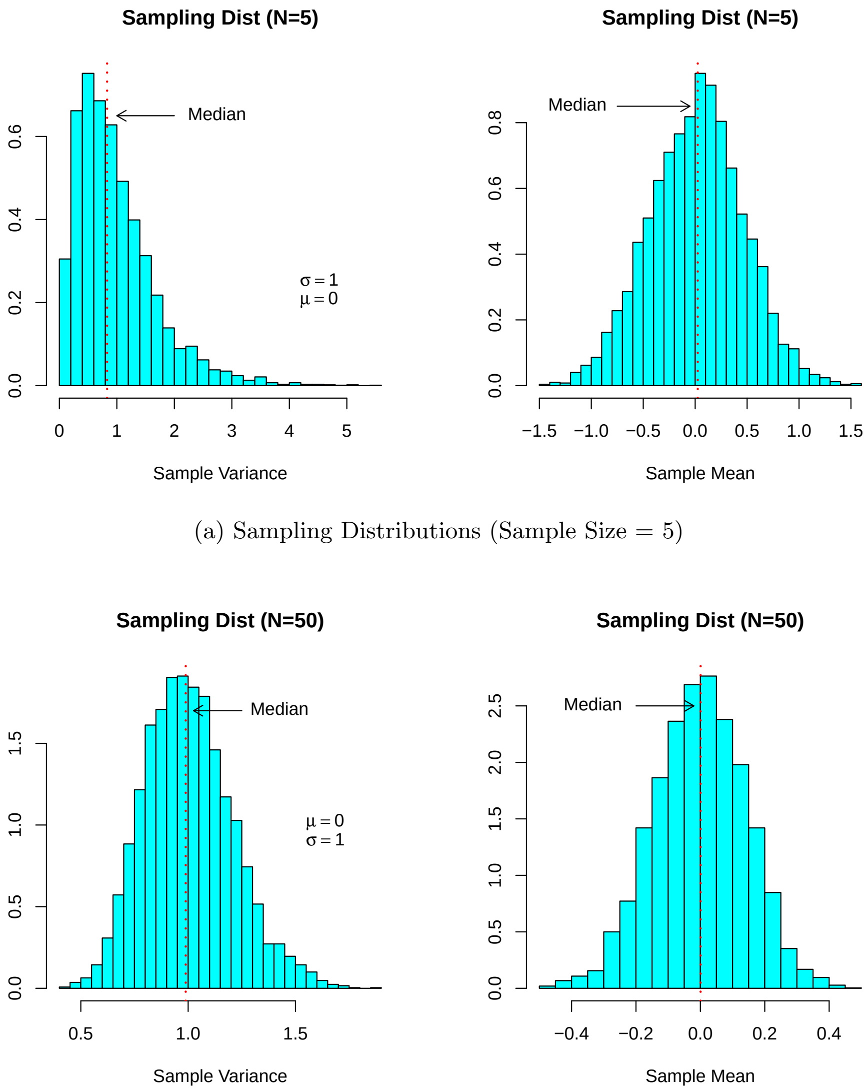
```
> my.scores <- c(96, 94, 102,  86, 100,  92, 105,  86, 111,
  82,  96,  79,  79,  85,  99,  97, 105,  98,  94,  73)
> t.test(x=my.scores,alternative="two.sided", mu=100)
```

0.4 two-tailed vs. one-tailed tests

Our `t.test` example evaluated the null hypothesis that $\mu = 100$. In fact, this hypothesis does not quite capture the essence of the experimental question, which was whether the exposure to the toxin had *reduced* verbal IQ. A better set of null and alternative hypotheses would reflect this directional aspect of the question:

$$\begin{aligned} H0 : \mu &\geq 100 \\ H1 : \mu &< 100 \end{aligned}$$

As in the original example, the null hypothesis is formulated in a way that is opposite to our expectations of the effect of the toxin. Also, the null and alternative hypotheses are mutually exclusive (i.e., only one can be true) and exhaustive (i.e., one *must* be true). The difference between the original hypotheses and this new set is that the latter are directional. Now, only scores that are *less than* 100 tend to favor H1, and



(a) Sampling Distributions (Sample Size = 5)

(b) Sampling Distributions (Sample Size = 50)

Figure 3: Sampling distributions of the variance and the mean for sample sizes of 5 (top) and 50 (bottom). The histograms show the variance and mean of 5,000 samples drawn from a normal distribution with $\mu = 0$ and $\sigma = 1$. The sampling distribution of the variance, but not the mean, is positively skewed, particularly for small samples. The positive skew implies that the median of the variance sampling distribution is less than 1. Hence, although the *mean* of the sample variance is 1, and therefore the sample variance (s^2) is an unbiased estimate of the population variance ($\sigma^2 = 1$), s^2 is less than σ^2 in more than 50% of the samples.

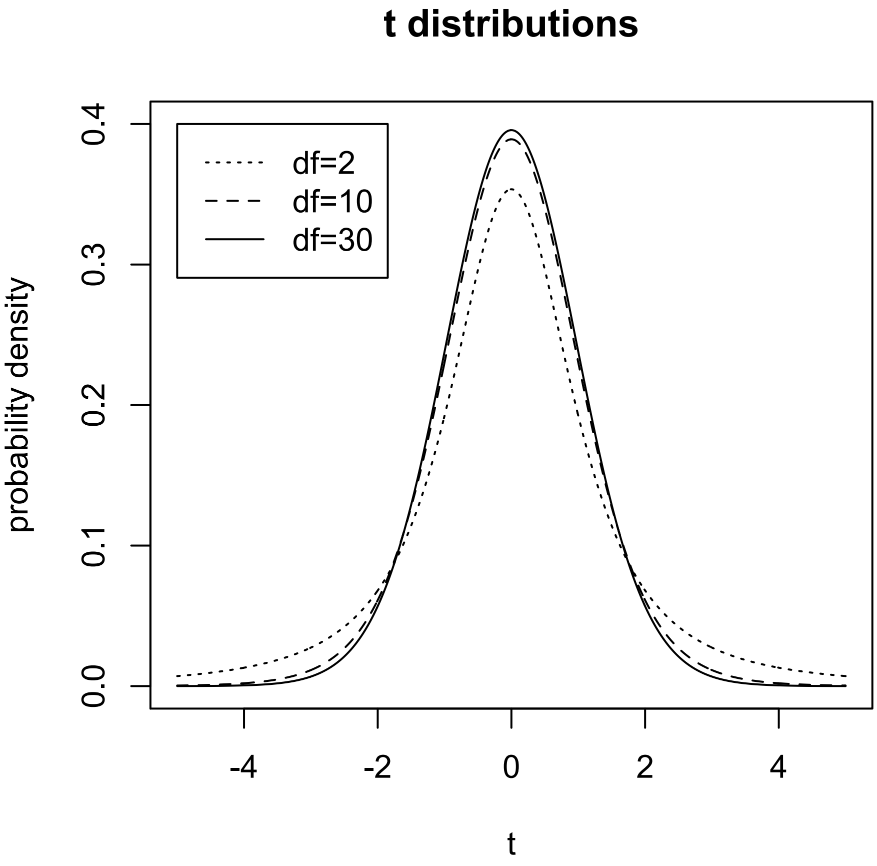


Figure 4: Probability density functions for the t distribution with 2, 10, and 30 degrees-of-freedom.

therefore we need only one criterion to decide if our observed score is unusually *lower* than 100. If we are doing a z test, for example, then the criterion would be -1.64 for $\alpha = 0.05$, and -2.33 for $\alpha = 0.01$. If we were doing a t test in R, we would type

```
> t.test(x=my.scores,alternative="less", mu=100)
```

On the other hand, if we wanted to determine if exposure to the toxin had *increased* IQ, then we would evaluate the hypotheses

$$\begin{aligned} H0 : \mu &\leq 100 \\ H1 : \mu &> 100 \end{aligned}$$

with the R command

```
> t.test(x=my.scores,alternative="greater", mu=100)
```

0.5 Type I vs. Type II errors

We have said that rejecting the null hypothesis when it is true is a Type I error, and that the probability of making such an error is α. Another kind of error occurs when we fail to reject the null hypothesis when it is false. This type of mistake is called a Type II error, and the probability of it occurring is called β. In fact, the two states of the world regarding H0 (i.e., it is either True or False) combined with the two possible decisions we can make regarding H0 mean that there are four possible decision outcomes (see Table 1). Obviously, we would like to maximize our correct decisions, and so we should minimize the probability of making both Type I (α) and Type II (β)errors. Of course, Type I errors can be minimized by adopting very small levels of α. Unfortunately, adopting a small α will (all other things begin equal) lead to an increase in β.

Table 1: Possible outcomes of hypothesis testing.

decision	H0 is True	H0 is False
reject H0:	Type I ($p = \alpha$)	Correct ($p = 1 - \beta$ =power)
do not reject H0:	Correct ($p = 1 - \alpha$)	Type II error ($p = \beta$)

The probability of rejecting H0 when it is false is referred to as the **power** of a statistical test, and it equals $1-\beta$. Obviously, the power of test depends on the size of the effect being studied. In our IQ example, for instance, it would be much easier to detect the effect of the toxin if it reduced IQ by 50 points than if it reduced IQ by only 1 point. Power also generally increases with increasing sample size. Do you see why? (Hint: Think about what happens to the variation among sample means as sample size increases). Finally, the power of a test *declines* as α increases. The power of a t test can be calculated using R's `power.t.test()` command. The following example shows how to use the command to compute the power of a two-sided t test:

```
> power.t.test(n=20,delta=5,sd=10,sig.level=0.05,type="one.sample",alternative="two.sided")

     One-sample t test power calculation

              n = 20
          delta = 5
             sd = 10
      sig.level = 0.05
          power = 0.5644829
    alternative = two.sided
```

The power is 0.56. What does this probability mean? If our sample size is 20, and if the scores are selected from a population that has a standard deviation of 10 and a mean that is 5 less than the mean in the null hypothesis, then the probability is 0.56 that we will (correctly) reject the null hypothesis. So, if the population mean of our IQ scores is 95, then we have a 56% chance of correctly rejecting H0 when we use a sample size of 20. If we evaluate a one-tailed null hypothesis, then the power increases to 0.69:

```
> power.t.test(n=20,delta=5,sd=10,sig.level=0.05,type="one.sample",alternative="one.sided")

     One-sample t test power calculation

              n = 20
          delta = 5
             sd = 10
      sig.level = 0.05
          power = 0.6951493
    alternative = one.sided
```

This increase in power is one very good reason for using one-tailed tests whenever possible.

0.6 what do p-values mean?

A p value indicates the probability of getting our score ($\bar{Y} = 93$), or one that is more extreme (i.e., further from the mean), assuming that the null hypothesis is true. We can write this more formally with the notation that is used to express conditional probabilities: the probability of getting our data (or data that are more extreme) *given* that the null hypothesis (H0) is true is $P(\text{data}|H0)$. Notice that this probability is not quite the same as the probability that H0 is true given our data, or $P(H0|\text{data})$, which is usually what we want to know. **It is vitally important that you understand that these two conditional probabilities are not the same:**

$$P(\text{data}|H0) \neq P(H0|\text{data})$$

In fact, there is no way to derive $P(H0|\text{data})$ from $P(\text{data}|H0)$ in the absence of other information about the *a priori* probability that H0 is true. Another common mistake is to interpret a p value as indicating the probability that a particular finding would be replicated: for instance that a p value of .01 means that if the experiment was conducted many times then a significant result would be obtained 99% of the time. Such an interpretation is incorrect: the probability of obtaining a statistically significant result when the null hypothesis is false depends on the effect size, sample size, and the alpha level. Indeed, given the moderate-to-low power of many psychological experiments, the actual probability of obtaining a significant result typically is much lower than a naïve guess based on the p value. Finally, there is a tendency to interpret statistical tests (incorrectly) in a binary fashion: a significant test is interpreted as showing that an effect is real, whereas a non-significant test indicates the effect is literally zero. These and other issues have been discussed by many statisticians and scientists in what is now a large literature on the use and abuse of p values, as well as the limitations of the null hypothesis testing procedure that I have described in these notes (e.g., Altman and Bland, 1995; Cohen, 1994; Dixon, 2003; Gelman, 2013; Krantz, 1999; Loftus, 1996; Lykken, 1968).

It also is important that you realize that the p-values are correct *only if the assumptions underlying the statistical tests are true.* A t test, for example, assumes that the scores are drawn independently from a normal population. If these assumptions are correct, then the p values obtained with a t test will be correct. If the distribution deviates from a normal distribution, or if the scores are not independent, then the p values may be very misleading. Other statistical procedures — the analysis of variance, for example — make more assumptions about the data. Again, the p values are exactly correct only if *all* of the assumptions are true. If one or more assumptions are false, then the p values are, at best, only approximately correct. In general, it is unlikely that all of the assumptions of a statistical test are strictly true, and so it is unlikely that the p are *exactly* correct. It makes you wonder if p values of 0.04, 0.05 and 0.06 differ in any meaningful way…

0.7 difference between two sample means

0.7.1 two independent samples

So far we have considered how to decide whether a single sample mean was drawn from a population with a mean of μ. More commonly, we want to compare two sample means to each other to decide whether or not they differ. If the population means of the two samples are μ_A and μ_B, then the two-sided null and alternative hypotheses are:

$$\begin{aligned} H0: \mu_A &= \mu_B \\ H1: \mu_A &\neq \mu_B \end{aligned}$$

These hypotheses can be re-written as:

$$\begin{aligned} H0: \mu_A - \mu_B &= 0 \\ H1: \mu_A - \mu_B &\neq 0 \end{aligned}$$

Each sample mean is a random variable drawn from the sampling distribution of the mean. The means of the sampling distributions will be μ_A and μ_B, and the variances will be σ_A^2/n and σ_B^2/n. (We will take advantage of the Central Limit Theorem and assume that the distributions are normal.) Now, let's create a new variable that is the difference between sample means, $d = \bar{Y}_A - \bar{Y}_B$. How is d distributed? It can be shown that the sum (or difference) of two or more Normal variables is also a Normal variable, so d is distributed Normally. Furthermore, the mean of a sum (or difference) of two variables is the sum (or difference) of the two means. Therefore, the mean of the difference distribution equals the difference between the means of the two sampling distributions of the means: $\bar{Y}_d = \mu_{\bar{Y}_A} - \mu_{\bar{Y}_B}$. If the two means are equal, then $\bar{Y}_d = 0$. Finally, the variance of a sum (or difference) of two *independent* random variables is the sum of the two variances: $\sigma_d^2 = \sigma_{\bar{Y}_A}^2 + \sigma_{\bar{Y}_B}^2$. (N.B. The variance of a difference between independent variables also is the *sum*, not the difference, of the two variances.) Therefore, if we know the form of the populations A and B, then statistical theory allows us to predict the form of the distribution between the means of samples drawn from those populations (see Figure 5).

In general, we will not know either σ_A^2 or σ_B^2, and therefore $\sigma_{\bar{Y}_A - \bar{Y}_B}^2$ will have to be estimated from the data with the formula

$$\hat{\sigma}_{\bar{Y}_A - \bar{Y}_B}^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \left[\frac{1}{n_A} + \frac{1}{n_B} \right] \tag{7}$$

where n_A and n_B are the sample sizes of the two groups. The two-group t statistic is

$$t = \frac{(\bar{Y}_A - \bar{Y}_B) - (\mu_A - \mu_B)}{\hat{\sigma}_{\bar{Y}_A - \bar{Y}_B}} \tag{8}$$

with df $= n_A + n_B - 2$. A close comparison of Equations 6 and 8 will reveal an underlying similarity: in both cases, t equals the difference between an observed statistic and a population parameter, divided by the standard error of the statistic:

$$t = \frac{\text{statistic} - \text{parameter}}{\text{estimated standard error of statistic}}$$

This general formula applies to all t tests.

In R, a two-sample t test is done with the `t.test()` command. Notice that the parameter `var.equal` is set to `TRUE`. Using this setting tells R to assume that samples A and B have the same variance.

```
> sample.A <- c(95, 92, 93, 96, 98, 99)
> sample.B <- c(101, 99, 106, 100, 98, 111)
> t.test(x=sample.A,y=sample.B,alternative="two.sided", var.equal=TRUE)
```

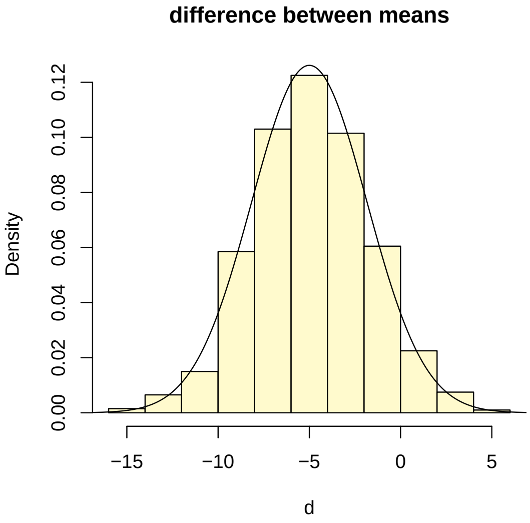


Figure 5: A histogram of differences between 1,000 pairs of sample means drawn from two Normal populations. The populations had the same variance but different means ($\mu_A - \mu_B = -5$). The smooth curve shows the normal distribution of the difference scores that is predicted from statistical theory when $\sigma^2_{\bar{Y}_A - \bar{Y}_B}$ is known.

```
        Two Sample t-test
data:  sample.A and sample.B
t = -3.0031, df = 10, p-value = 0.01327
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.193683  -1.806317
sample estimates:
mean of x mean of y
     95.5     102.5
```

If the variances of the two samples are not equal, then the calculated t statistic will not be distributed as a t variable with $n_A + n_B - 2$ degrees of freedom, and the p value given by `t.test` will be inaccurate. However, the t statistic will be approximately distributed as a t variable with fewer degrees of freedom. R corrects the degrees of freedom and the p value when the parameter `var.equal` is `FALSE`. This is the default setting, so we can use the following command:

```
> t.test(x=sample.A,y=sample.B,alternative="two.sided")
```

```
        Welch Two Sample t-test
data:  sample.A and sample.B
t = -3.0031, df = 7.743, p-value = 0.01764
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.40638  -1.59362
sample estimates:
mean of x mean of y
     95.5     102.5
```

Notice that the degrees of freedom are lower than before and, consequently, the p value is higher. The output is labeled `Welch Two Sample t-test` because it uses the Welch approximation to the degrees of freedom. (The approximation is sometimes referred to as the Welch-Satterthwaite approximation after the two investigators who (independently) derived the formula.) The assumption that the two samples come from populations that have *exactly* the same variance is almost always invalid, so the Welch t test almost certainly is a more valid test of the difference between two means.

0.7.2 two paired samples

Often we want to compare the means of two sets of measurements gathered on the same sample. For example, we might want to compare test performance on twenty subjects both before and after they have undergone some experimental treatment. In this case, we cannot treat the two sets of numbers as independent samples because they come from the same subjects, and therefore the t test illustrated in the previous section is inappropriate. Instead, we convert the pairs of numbers to a single set of difference scores, and perform the t on those numbers. If we want to determine if the two sets of numbers differ, then we would use a t test to evaluate the null hypothesis of no difference (i.e., $\mu_{\text{difference}} = 0$). The degrees of freedom is one less than the number of difference scores. (N.B. In this test, *all* of the numbers in one sample must be paired with a number in the other sample. Unpaired numbers are not included in the analysis).

In R, paired t tests are performed by setting the `paired` parameter in `t.test` to `TRUE`. The following example shows how to use R to do a paired t test.

```
> t.test(x=sample.A,y=sample.B,paired=TRUE)

        Paired t-test
data:  sample.A and sample.B
t = -3.5, df = 5, p-value = 0.01728
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.141164  -1.858836
sample estimates:
mean of the differences
                     -7
```

References

Altman, D. G. and Bland, J. M. (1995). Absence of evidence is not evidence of absence. *BMJ*, 311(7003):485.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12):997–1003.

Dixon, P. (2003). The p-value fallacy and how to avoid it. *Can J Exp Psychol*, 57(3):189–202.

Gelman, A. (2013). P values and statistical practice. *Epidemiology*, 24(1):69–72.

Krantz, D. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, 44(448):1372–81.

Loftus, G. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 5:161–71.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychol Bull*, 70(3):151–9.

Student (1908). The probable error of a mean. *Biometrika*, 6:1–25.

Wilcox, R. R. (2002). Understanding the practical advantages of modern ANOVA methods. *J Clin Child Adolesc Psychol*, 31(3):399–412.