BIM 105
Probability and Statistics for Biomedical Engineers

David M. Rocke

Department of Biomedical Engineering

Hypothesis Tests

- We have seen that a confidence interval for a parameter based on a sample from a population can be thought of as the set of parameter values that is consistent with the data collected, at the specified level of confidence.
- For example, if a 95% confidence interval for the mean was 111 ± 12, or (99, 123), then values for the population mean consistent with the data include 100, 99, 120, but not 98 or 125.
- A hypothesis test is used when we have one specific value of the parameter that we wish to know if it is consistent with the data.
- We could just look to see if it is in the 95% confidence interval, but that does not distinguish between $\mu = 122.5$, which is almost out of the CI, and $\mu = 114$, which is near the sample mean.

- The hypothesis we wish to test is called the *null hypothesis*, largely because we often want to test if the difference between two or more things is null, that is zero.
- We may or may not have a specific alternate hypothesis, but this is often not needed.
- We construct a statistic which under the null hypothesis has a known distribution; that is, we temporarily assume that the null hypothesis is true, and see what that would imply.
- If the value of this statistic is sufficiently unlikely under the null hypothesis, then we conclude that the null hypothesis probably is not true.
- **We can never conclude that the null hypothesis is true, only that it is not shown to be false.**

Example

The mean calcium content of a type of cell growth medium is supposed to be 30 g/L. A sample of 100 batches has a mean concentration of 36 g/L with a standard deviation of 14 g/L. Test the hypothesis that the population mean is actually 30 as specified.

$H_0 : \mu = 30$

Under the null hypothesis,

$$\frac{\overline{x} - 30}{14/\sqrt{100}} \sim N(0,1)$$

$$\frac{36-30}{1.4} = \frac{6}{1.4} = 4.286 \sim N(0,1)?$$

This is too large to have come from a standard normal random variable. Almost all values of a standard normal lie between -3 and 3.

Fixed Level Tests

- We compute a statistic that should be standard normal if the null hypothesis is true.
- If that statistics is either too large (like 3) or too small (like -3), then we will reject the null hypothesis, that is, decide that it is likely not to be true.
- For example, the chance that a standard normal random variable is larger than 1.960 in magnitude is 5%.
- If we choose 5% as the criterion level, then we will reject the null hypothesis whenever the z statistic is larger than 1.960 or smaller than –1.960.

P-Values

- Suppose we had a sample mean level of 33 g/L with a standard deviation of 14 g/L and $n = 100$. To test the null hypothesis that the population mean is 30, we calculate $z = (33 - 30)/1.4 = 2.143$. This is approximately standard normal if the true mean is 30.
- The chance that z exceeds 2.143 is 0.0161. We double this to account for the fact that we might have gotten a sample mean below 30. This is called a *two-sided test*.
- The (two-sided) p-value is then 2(0.0161) = 0.0322
- This means that if the null is true, we would get a value as extreme as this only 3.22% of the time.
- At this point, we may choose to reject the hypothesis based on a p-value threshold like 5%.
- We have two choices: either something unlikely has happened or else the null hypothesis is false.

We Never "Accept the Null"

- If we have a null hypothesis that the population mean is 30, then we know from first principles that it is false at some decimal point. The true mean may be in 30 ± 1 or 30 ± 0.1 or 30 ± 0.01, but it is not in 30 ± 10^{-9}.
- So when we reject the null, it means that we know that the true mean is not 30 or very close to 30.
- If we do not reject the null, that does not mean that we know that the true mean is exactly 30.
- In fact, the confidence interval contains all the values of the true mean consistent with the data.

A sample of 100 batches has a mean concentration of 33 g/L with a standard deviation of 14 g/L. A 95% confidence interval for μ is

$$33 \pm (1.960)(14/\sqrt{100})$$

$$33 \pm 2.744$$

$$(30.256, 35.744)$$

The value 30.256 is the smallest value still in the 95% confidence interval.

If we test the hypothesis that the population mean is actually 30.256, then

$$H_0 : \mu = 30.256$$

Under the null hypothesis,

$$\frac{\overline{x} - 30.256}{14/\sqrt{100}} \sim N(0,1)$$

$$\frac{33 - 30.256}{1.4} = \frac{2.744}{1.4} = 1.960$$

The p-value is 2(0.0250)=0.05 or 5%

The 95% confidence interval is the set of possible null hypotheses for μ such that the p-value is greater than 5%

We can make a 95% confidence interval by inverting the test because the 95% confidence interval is the set of possible null hypotheses for μ such that the p-value is greater than 5%. This happens when the test statistic is larger than 1.960 in absolute value.

$$\frac{33-\mu}{14/\sqrt{100}} < 1.960$$

$$33-\mu < (1.960)(14/\sqrt{100})$$

$$\mu > 33-(1.960)(14/\sqrt{100})$$

$$\frac{33-\mu}{14/\sqrt{100}} > -1.960$$

$$33-\mu < -(1.960)(14/\sqrt{100})$$

$$\mu < 33+(1.960)(14/\sqrt{100})$$

$$33-(1.960)(14/\sqrt{100}) < \mu < 33+(1.960)(14/\sqrt{100})$$

$$30.256 < \mu < 35.744$$

Type I and Type II Errors

- We make a Type I error when we reject a null hypothesis even though it is in fact true.
- By definition, we will always make some Type I errors, but we can explicitly control how many we make.
- If we use a 5% criterion, then we will reject a true null hypothesis 5% of the time.
- We make a Type II error when we fail to reject a false null. We can't calculate this without specifying what false null we are talking about, and it cannot reasonably be so close to the null that it makes no difference.

Type I and Type II Errors

- For example, if the null hypothesis is that the population mean is 30 g/L, we may choose to calculate the Type II error for a concentration of 35 g/L.
- The value of the parameter that we use to calculate the Type II error is sometimes called the *alternate hypothesis*.
- For a given sample size, the *power* of a test is one minus the Type II error. This depends on the alternate hypothesis.
- We would like to have small Type I error and large power, but this may require very large values of n.

Why do we double the tail area?

- Suppose we want to test the hypothesis that the true mean is 30 g/L. A sample of 100 gives a mean of 33 g/L and a standard deviation of 14 g/L.
- If we construct the test statistic, we get z = 2.143, which has a tail area to the right of 0.0161. This happens 1.61% of the time when the null is true.
- The p-value is the fraction of time that we would reject the null if the null is true, and it looks like this would be 1.61%.
- But if the sample mean had been 27 g/L, then the test statistic would have been z = –2.143, which has a tail area of 1.61% to the left.
- So if we use the 1.61% criterion, we reject the null 1.61% of the time with positive z and 1.61% with negative z, for a total of 2(1.61%) = 3.22%

What about one-sided tests?

- The book discusses a one-sided test, used when one expects only deviations on one side of the null.
- This should almost never be used because it is a form of self deception.
- Suppose I think that vitamin C prevents cancer. I might advocate a hypothesis test with null hypothesis that vitamin C has no effect and alternate hypothesis that it helps. Say I use a 5% cutoff.
- I am then promising that no matter how much the evidence seems to show that vitamin C causes cancer, I will not make any conclusion from the data.
- This is obvious nonsense; if the evidence looks strong, I will conclude that vitamin C is harmful.
- But then my chance of being wrong is 5% plus the chance that the evidence of harm is sufficiently great, which might also be 5%.

What about one-sided tests?

- One-sided tests can be used in a decision context. If we test a new manufacturing process for a medical device to see if the defective rate can be reduced from 4% to a lower rate, then we only want to adopt or pursue the new process if the defective rate is lower than 4%. If the defective rate of the new process is the same or higher, then we don't want to use the new process.
- One-sided tests are appropriate for some decision analysis contexts. They are not usually appropriate for science.
- It is not sufficient to say that "we want to know if the mean is less than 12." It must be the case that departures on the other side lead to the same conclusions as if the sample mean is near the true mean.
- Some book problems will assume a one-sided test. You can tell if the alternate is phrased with "greater than" or "less than" instead of "not equal."

Hypothesis Tests Summary

- The test a statistical hypothesis you use a statistic whose distribution is known if the null hypothesis is true.
- Compute the probability that the statistic would be as greater (if positive) or less than (if negative) and double it. This is the (two-sided) p-value.
- Reject the null hypothesis if this p-value is sufficiently small.
- A $100(1 - \alpha)\%$ confidence interval is the set of values of the parameter such that the null hypothesis p-value is less than α.

One-Sample Tests

If we have a sample $x_1, x_2, ..., x_n$ from a population with mean μ

and we have a possible value $\mu = \mu_0$ in mind

and want to see if the data are consistent with that value of μ,

we construct the test statistic

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

and we compare it to the t-distribution with $n-1$ degrees of freedom,

or with the normal distribution if n is large.

If t is positive and $\Pr(T > t) = \alpha$,

then the p-value of the test is 2α.

If t is negative and $\Pr(T < t) = \alpha$,

then the p-value of the test is 2α.

For example, if 18 cells were sized

and the average diameter was 22 μm with a

standard deviation of 3 μm,

test the hypothesis that the true value of μ is 20 μm.

$$\frac{22-20}{3/\sqrt{18}} = \frac{2}{0.707} = 2.828 = t_{17}$$

$\Pr(T_{17} > 2.828) = 0.0058$ so the p-value is 2(0.0058)=0.0116.

The value $\mu = 20$ is not very consistent with the data.

This requires MATLAB to compute the tail area 0.0058.

From Table A.3, we can say that the tail area is less than 0.01 (2.567)

but not as small as 0.005 (2.898), so we can say only that $p < 0.02$.

Proportions

Suppose we have a sample from a production run of a component for a medical device and they are tested to see if they are within specifications. The manufacturing process is delicate, so up to 30% of the components may be defective. If we have 600 components of which 217 are defective, test the hypothesis that $p = 0.30$. For a large value of n, we have that

$X \sim \text{Bin}(n, p) = \text{Bin}(600, 0.30)$ if $p = p_0 = 0.30$

$\hat{p} = X / n = 217 / 600 = 0.36167$

$E(\hat{p}) = p = 0.30$ under the null hypothesis

$V(\hat{p}) = p(1-p)/n = (0.30)(0.70)/600 = 0.00035$

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{217/600 - 0.30}{\sqrt{(0.30)(0.70)/600}} = \frac{0.06167}{0.01871} = 3.296 = z$$

The p-value for this test is 2(0.00049) = 0.00098, so not consistent. We reject the null hypothesis.

Proportions

When making a confidence interval for a proportion, we know that this should be

$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$ or

$\tilde{p} \pm 1.96\sqrt{\tilde{p}(1-\tilde{p})/n}$

where $\hat{p} = x/n$ and $\tilde{p} = (x+2)/(n+4)$

For a hypothesis test, we don't have this problem. The null hypothesis is

$p = p_0$ so under the null, we know what p is, we don't have to estimate it.

Under the null, the variance of $\hat{p}$ is $p(1-p)/n$, and under the null this is known.

$n = 600 \quad x = 217 \quad \hat{p} = 0.36167$

under the null hypothesis that $p = 0.30$

$E(\hat{p}) = p = 0.30$

$V(\hat{p}) = p(1-p)/n = (0.30)(0.70)/600 = 0.00035$

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} = \frac{217/600 - 0.30}{\sqrt{(0.30)(0.70)/600}} = \frac{0.06167}{0.01871} = 3.296 = z$$

The p-value for this test is 2(0.00049) = 0.00098, so not consistent. We reject the null hypothesis.

We still need the normal approximation to make this correct.

Confidence Intervals and Hypothesis Tests

For means, the $100(1-\alpha)\%$ confidence interval is exactly the set of hypothetical values of μ that are not rejected by the usual hypothesis test if we set the cutoff for rejection at α. Otherwise put, a test of the hypothesis that $\mu = \mu_0$ is rejected at a significance level of α exactly when μ does not lie in the $100(1-\alpha)\%$ confidence interval for the mean.

This does not quite work for proportions because the ordinary confidence confidence interval uses a standard deviation of $\sqrt{\hat{p}(1-\hat{p})/n}$ to see if a particular value $p = p_0$ lies in the interval, whereas the test uses $\sqrt{p_0(1-p_0)/n}$ MATLAB's Clopper-Pearson interval (the default) does exactly invert the hypothesis test.

The Chi-Squared Distribution

- The Chi-Squared or Chi-Square distribution is the distribution of a sum of squares of independent standard normals.
- If $X_1, X_2, \ldots, X_k$ are independent standard normal random variables, and if $Y = X_1^2 + X_2^2 + \ldots + X_k^2$, then Y has a $\chi^2(k)$ distribution.
- This is a type of gamma distribution. It is strictly positive. Its mean is k and its variance is $2k$
- One of the uses is to assess the difference between counts and expected counts.

Observed and Expected Counts

- For a Poisson distribution, the expected count is λ. If the actual count is X, and if λ is large, then $(X - \lambda)/\sqrt{\lambda}$ is approximately normal, so $(X - \lambda)^2/\lambda$ is approximately $\chi^2(1)$
- This is of the form $(O - E)^2/E$, where O is the observed count and E is the expected count.
- This is large when the observed value is far from the expected value.
- The χ^2 approximation to this distribution only works well if the expected count is at least largish, conventionally at least 5.
- The expected values can come from a variety of sources depending on the problem, but a two common uses are to test for goodness of fit or to test for independence of two discrete random variables.

Is the Die Fair?

Category	Observed	Expected	$(O - E)^2$	$(O - E)^2/E$	
1	115	100	225	2.25	
2	97	100	9	0.09	
3	91	100	81	0.81	
4	101	100	1	0.01	
5	110	100	100	1.00	
6	86	100	196	1.96	
	600	600		6.12	

6.12 is to be compared to a Chi-Squared distribution. The number of degrees of freedom is 5 since the expected values were fitted to the total of 600. Only large values show a discrepancy. The area to the right tail is 0.4099 (using MATLAB chi2pdf(x,v)) or from Table A.5 we see that 6.12 does not exceed the 10% tail area point of 9.236. We don't have to double the p-value because deviations in any direction generate larger values of the Chi-Squared statistic.

Goodness of Fit

- The previous example is goodness of fit test.
- It can test a specific distribution against a set of data that may or may not have that distribution.
- The data can be discrete, or continuous data can be binned (as one does to make a histogram).
- The degrees of freedom is the number of categories less the number of parameters that are fit to the data.

Homogeneity or Independence

- The chi-squared distribution can also be used to test for homogeneity or independence in two-way table.
- This is when data are classified in two ways and the question is whether the two are related or independent.
- This is different from deciding on independence when we know the probability distribution. In that case, we just check if the joint probability of two events is the product of the individual probabilities.
- When we have data, the question is this: is there a set of probabilities for the events such that they are independent and such that the data are consistent with these probabilities?

Lymphoma Clinical Trials

- Intensive vs. moderate chemotherapy in lymphoma. Adapted from Pocock, S, *Clinical Trials*.
- 273 patients, 138 on one treatment (BP) and 135 on the other (CP)
- Are the treatments provably different in outcome?
- 77/138 = 56%; 90/135 = 67%

	BP	CP	Total
Improvement	77	90	167
No Improvement	61	45	106
Total	138	135	273

	BP	CP	Total
Improvement	77	90	167
No Improvement	61	45	106
Total	138	135	273

We test for independence, meaning that improvement rates on the two treatments is the same. To use the chi-squared test, we need to calculate expected values for each cell in the table. For Improvement on BP, we have 167/273 of the patients are improved and 138/273 are on BP, so if these events were independent the probability would be

$\left(\frac{167}{273}\right)\left(\frac{138}{273}\right)$ and the expected number in that cell would be

$\left(\frac{167}{273}\right)\left(\frac{138}{273}\right)273 = \frac{167\times 138}{273} = 84.41,$ which is (row total)(column total)/total.

We can do the same for the other cells, or subtract from the margins.

	BP	CP	Total
Improvement	77	90	167
No Improvement	61	45	106
Total	138	135	273

$$\chi_1^2 = \frac{(77-84.42)^2}{84.42} + \frac{(90-82.58)^2}{82.58} + \cdots + \frac{(45-52.42)^2}{52.42}$$
$$= 0.6518 + 0.6662 + \cdots + 1.0497$$
$$= 3.3945$$

The p-value is 0.0654 using matlab `chi2cdf(3.3945,1)`.

or from table A.5, the p-value is between 0.05 (3.841) and 0.10 (2.706)

The degrees of freedom starts with the number of cells (4).

The margins (row and column sums) are fixed.

There is then only one cell with some choice to make the margins add up

Once the top left cell is known, the rest are determined.

If there are r rows and c columns, then the df is $(r-1)(c-1)$

Same Problem, Different Test

We could view this as a test of difference of proportions

$p_1 = 77/138 = 0.5580$ *vs.* $p_2 = 90/135 = 0.6667$

$\hat{p} = (77+90)/(138+135) = 0.6117$

$\sqrt{(0.6117)(0.3883)(1/138+1/135)} = 0.05900$

$$z = \frac{p_2 - p_1}{0.05900} = \frac{0.1087}{0.05900} = 1.842$$

p-value = 0.06541

Compare to p = 0.0654 by the chi-squared test.

Chi-squared test can handle more than two by two.

Failures of Four Compressors

- There are four compressors in a plant, each of which has three legs, North, Center, and South.
- We compare piston-ring failures in each leg on each compressor over a year's time.
- The question of interest is whether the four apparently identical compressors have different patterns of failure.
- We can also ask if the compressors have different rates of failure overall.
- Davies, OL and Goldsmith, PL, *Statistical Methods in Research and Production*.

Compressor	North Leg	Center Leg	South Leg	Total
1	17	17	12	46
2	11	9	13	33
3	11	8	19	38
4	14	7	28	49
Total	53	41	72	166

To test for independence, meaning that the failure rates on the three legs is similarly distributed across the four compressors, we need to calculate expected values for each cell in the table. For the north leg of compressor 1, we have that compressor 1 has 46/166 of the total failures and north legs have 53/166, so if these events were independent the probability would be $\left(\frac{46}{166}\right)\left(\frac{53}{166}\right)$ and the expected number of failures in that cell would be

$$\left(\frac{46}{166}\right)\left(\frac{53}{166}\right)166 = \frac{46\times 53}{166} = 14.69,$$ which is (row total)(column total)/total.

Observed

Compressor	North Leg	Center Leg	South Leg	Total
1	17	17	12	46
2	11	9	13	33
3	11	8	19	38
4	14	7	28	49
Total	53	41	72	166

Expected

Compressor	North Leg	Center Leg	South Leg	Total
1	14.69	11.36	19.95	46
2	10.54	8.15	14.31	33
3	12.13	9.39	16.48	38
4	15.64	12.10	21.25	49
Total	53	41	72	166

Observed Expected

Compressor	North Leg		Center Leg		South Leg		Total
1	17	14.69	17	11.36	12	19.95	46
2	11	10.54	9	8.15	13	14.31	33
3	11	12.13	8	9.39	19	16.48	38
4	14	15.64	7	12.10	28	21.25	49
Total		53		41		72	166

$$\chi^2_6 = \frac{(17-14.69)^2}{14.69} + \frac{(17-11.36)^2}{11.36} + \cdots + \frac{(28-21.25)^2}{21.25}$$
$$= 0.3644 + 2.7983 + \cdots + 2.1419$$
$$= 11.7227$$

The p-value is 0.06845 using matlab `chi2cdf(11.7227,6)`.

or from table A.5, the p-value is between 0.05 (12.592) and 0.10 (10.645)

The degrees of freedom is $(r-1)(c-2)$ or 3×2 or 6.

The margins (row and column sums) are fixed.

There are then 6 cells with some choice to make the margins add up

Compressor	North Leg	Center Leg	South Leg	Total
1	17	17	12	46
2	11	9	13	33
3	11	8	19	38
4	14	7	28	49
Total	53	41	72	166

Compressor	North Leg	Center Leg	South Leg	Total
1	17	17		46
2	11	9		33
3	11	8		38
4				49
Total	53	41	72	166

Overall failure rates

- The failures of the three compressors were 53, 41, 72, with a total of 166.
- The expected failure if they were all the same would be 166/3 = 55.33.
- So the chi-squared statistic would be
  $(53 - 55.33)^2/55.33 + (41 - 55.33)^2/55.33$
  $+(72 - 55.33)^2/55.33 = 8.8313$
- This should be referred to a chi-squared distribution with 2 df. The tail area is 0.0121.
- Thus, the failure rates are likely not the same.

Fisher's Exact Test

- The chi-squared test for a contingency table is generally considered accurate if all of the expected counts are at least 5.
- In cases where this is not true, we can use Fisher's exact test. In fact, some prefer to use this always.
- The idea is to line up all possible tables with a given set of row and column sums according to how strong the evidence of association is, then count how extreme the actual data are in this list. This is then doubled for a two-sided test.
- In practice, we weight each table by how likely it is to occur under the null hypothesis.
- This may be hard to compute in practice, and MATLAB can do so only for 2x2 tables, so can't be used for compressor example.
- R and other statistical packages can do this.

Glasses and Crime

- Study comparing health measures of juvenile offenders to a control group. Weindling, AM, Bamford, FN, and Whittall, RA "Health of Juvenile Delinquents," *British Medical Journal*, 292, 1986.
- Compares those in both groups who failed a vision test as to whether they actually wear glasses
- The sample size is too small for a chi-squared test.

	Offender	Non-Offender	Total
Glasses	1	5	6
No Glasses	8	2	10
	9	7	16

Obs	Offender	Non-Offender	Total
Glasses	1	5	6
No Glasses	8	2	10
	9	7	16

Exp	Offender	Non-Offender	Total
Glasses	3.375	2.625	6
No Glasses	5.625	4.375	10
	9	7	16

More Extreme	Offender	Non-Offender	Total
Glasses	0	6	6
No Glasses	9	1	10
	9	7	16

Obs	Offender	Non-Offender	Total
Glasses	1	5	6
No Glasses	8	2	10
	9	7	16

- The p-value from Fisher's exact test is 0.03497
- This is done using
```
[h p stats] = fishertest(glasses)
```
- The measure of extremeness is the odds ratio, the observed value of which is (1/5)/(8/2) = 0.05
- The chi-squared test is not accurate here so should not be used.

Tests in MATLAB

```
Loaded Alkaline Phosphatase data as x (298 obs)
>> xl = log(x)
>> mean(xl)

-0.3632

>> std(xl)

0.5469
>> ttest(xl,-.45)


     1
>> [h,p] = ttest(xl,-.45)

h =

     1

p =

    0.0065
```

```
>> [h,p,ci,stats] = ttest(xl,-.45)

h =

     1

p =

    0.0065

ci =

   -0.4255
   -0.3008

stats =

    tstat: 2.7410
       df: 297
       sd: 0.5469
```

```
For a binomial, we can use the ztest function. Suppose that we
have a sample of
>> p0=0.3
>> phat = 217/600

phat =

    0.3617

>> sp = sqrt(p0*(1-p0)/600)

0.0187

>> [h,p,ci,val] = ztest(phat,p0,sp)

h =

     1

p =

   9.7995e-04

ci =

    0.3250
    0.3983

val =

    3.2962
```

For a contingency table, MATLAB seems to be able to compute the chi-squared test only when the data are given in raw form, not in tabulated form. This example is one where the chi-squared statistic should Not actually be used because the sample size is too small, but here it is.

```
>> offend = [1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 ]
>> glasses = [1 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 ]
>> [table chi2 p labels] = crosstab(offend,glasses)

table =
     2     5
     8     1
chi2 =
    6.1122
p =
    0.0134
labels =
    '0'    '0'
    '1'    '1'
```

Here is Fisher's exact test:

```
>> glasses = [1 5;8 2]

glasses =

     1     5
     8     2

>> [h p stats] = fishertest(glasses)

h =

     1

p =

    0.0350

stats =

              OddsRatio: 0.0500
    ConfidenceInterval: [0.0035 0.7061]
```

Statistical Power

- Hypothesis tests are designed so that if the null hypothesis is true, the chance of the hypothesis being rejected is small. This is called a Type I error.
- For example, if we agree to reject the null when $p < 0.05$, then we make a Type I error only 5% of the time, or 1 time in 20.
- A Type II error is when the null hypothesis is false, but we don't reject the null.
- The chance of this happening depends on the specific alternate.

- Suppose we are measuring the size of fibroblast cells, and we want to test the hypothesis that the mean diameter is 30 μm. Assume we have measured 100 cells.
- We choose $\alpha = 0.05$, which means we reject the null if the p-value is less than 0.05.
- Assume that the population standard deviation is 3 μm, so that the standard error of the mean is $3/10 = 0.30$ μm.
- If the true mean is 30.01 μm, then we will commit Type II errors quite frequently. The chance of rejection will be 0.05013, so a Type II error happens almost 95% of the time.
- On the other hand, if the true mean is 32 μm, then we will reject the null with probability 0.915, so the chance of a Type II error is only 8.5%

- Power is the probability of rejecting a specific alternate hypothesis.
- If the power is not large, then the experiment is not worth performing because it will most likely end up with no conclusion.
- To set up a hypothesis test with specific level α, we only need to know the null hypothesis.
- To determine the power, we need also to have a specific alternate in mind and to have some idea what the standard deviation is.
- We can then use the normal distribution to find the chance of the Type II error and thus the power.

Biotechnology Drug Production

- Using standard genetically engineered yeast cells, a biologically active drug can be produced in concentrations of 3.5 gm per liter.
- A new strain of yeast is to be tested with 80 small batches in which the concentration is measured.
- If the concentration reached 4.0 g/L, then this would be economically important and would justify switching the process.
- Suppose previous experience suggested that the standard deviation of the yield would be 2.1 g/L.
- If a 5% test is used, and if the true yield was 4.0 g/L, what is the chance that this will be detected by rejecting the null?

$H_0 : \mu_0 = 3.5$

$\sigma = 2.1$

$$z = \frac{\overline{x} - \mu_0}{s / \sqrt{n}}$$

Reject the null if $| z | > 1.960$, that is, if either of

$\overline{x} > \mu_0 + 1.960 s / \sqrt{n} = 3.5 + (1.960)(2.1) / \sqrt{80} = 3.5 + (1.960)(0.2348) = 3.5 + (0.4602) = 3.9602$

$\overline{x} < 3.5 - 0.4602 = 3.0398$

Each of these has probability 0.025 under then null, so the chance of a Type I error is 5%

If the true mean is 4.0, then

$\Pr(\overline{x} > 3.9602) = \Pr(Z > (3.9602 - 4.0) / 0.2348) = \Pr(Z > -0.1695) = 0.567$

The chance that $\overline{x}$ is less than 3.0398 is negligible, so the power is 56.7% and the chance of a Type II error is 43.3%.

Sample Size Determination

- If the null hypothesis is that the mean is 3.5, the alternate that we would like to detect is 4 and the standard deviation is 2.1, how big does the sample have to be in order that the usual 5% test has 80% power?
- We saw that a sample size of 80 gives only 56.7% power.
- We could use trial and error, or a formula, or we could use a computer analysis. Mostly, in practice, we use a computer analysis.

We have a process whose mean yield is supposed to be 80 g/L. A new process is compared to the old one, and we wish to conduct a hypothesis test with $\alpha = 0.05$ so that the power is 80% if the yield of the new process is 82. How big does the sample size n need to be if we estimate that $\sigma = 5$?

The cutoff for a 95% two-sided test is $c = 80 + (1.960)(5/\sqrt{n})$

If the new mean is 82, then we need that the chance of being less than c is 0.20, and the value of the standard normal with 0.20 to the left is $-0.8416,$ so the point of a normal with mean 82 and standard deviation $5/\sqrt{n}$ is $82 - (0.8416)(5/\sqrt{n})$ so

$$80 + (1.960)(5/\sqrt{n}) = 82 - (0.8416)(5/\sqrt{n})$$

$$2 = (1.960 + 0.8416)(5/\sqrt{n})$$

$$2 = 14.008/\sqrt{n}$$

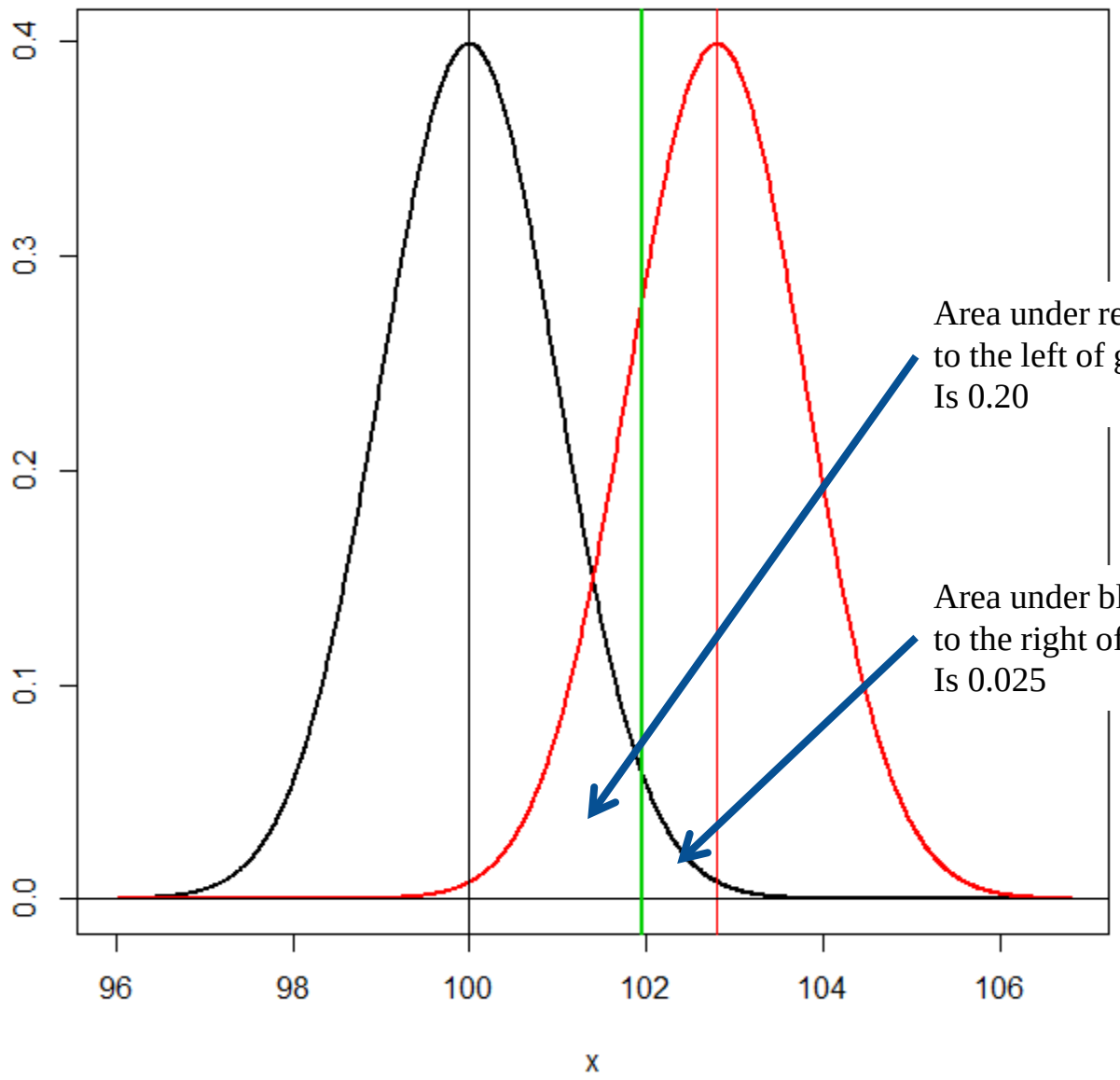
$$n = (14.008/2)^2 = 49.056$$

We need a sample size of at least 50.

If the null hypothesis is $\mu = \mu_0$ and we want a two sided test with $\alpha = 0.05$ so that the power is $100(1-\beta)\%$ when $\mu = \mu_1$, and assuming the standard deviation is σ, then

$$n > \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{(\mu_1 - \mu_0)^2}$$

5% test
80% power



Sample Size and Power in MATLAB

```
power = sampsizepwr(testtype,p0,p1,[],n)
n = sampsizepwr(testtype,p0,p1,power)

testtype: 't', 'z'   p0 = [mu0 sigma0]  p1 = mu1
The test size is always 0.05 unless specified to be otherwise.

>> sampsizepwr('t',[3.5 2.1],4,[],80)
    0.5572

>> sampsizepwr('z',[3.5 2.1],4,[],80)
    0.5674

>> sampsizepwr('t',[80 5],82,0.80)
    52

>> sampsizepwr('z',[80 5],82,0.80)
    50
```

Multiple Testing

- Suppose that we measure 50,000 things (expression of genes, amount of proteins) on each of 10 control samples and 10 treatment samples, say of chondrocytes in a 48-well microplate.
- If we do a 5% test on each of the 50,000 genes, and if the treatment actually does nothing at all, then we will have false positives of around (50,000)(0.05) = 2,500. This can cause many false conclusions that cannot be replicated.
- For significance, the ratio of the difference of means to the standard error of the difference needs to be at least 2 or so.

Bonferroni

- We can prevent this by testing at a smaller value of α.
- The Bonferroni inequality says that if we test at α/k, where k is the number of tests, then there is only a chance of α that even one false positive will occur.
- But $0.05/50000 = 10^{-6}$, and the critical value for a two-sided t-test is not around 2, but around 7, so only enormous differences can be detected.
- We can use improved methods of correcting for multiple testing, of which the most popular is the False Discovery Rate control methods.

Example

- A 2012 paper on the possible effect of dental x-rays on the risk of meningioma, compared a group of patients with the disease to randomly sampled controls matched for age.
- The 'result' that the variable 'ever had a bitewing x-ray' was associated with the disease at a 5% level was widely reported in the media. The paper was published in a top journal and had authors in the medical schools at Harvard, Yale, UCSF, and Duke.
- There were around 50 questions that were tested, and only one was 'significant'.
- With a 5% test and 50 tests, the expected number of false positives is 2.5.
- So the result is not believable.