0. Review of formulae for sums of random variables.

(a) \( E(X + Y) = E(X) + E(Y) \); and \( E(X - Y) = E(X) - E(Y) \). If \( X \) and \( Y \) are independent, then \( \text{var}(X + Y) = \text{var}(X) + \text{var}(Y) \), and \( \text{var}(X - Y) = \text{var}(X) + \text{var}(Y) \).

(b) In the general case, suppose that \( X_1, X_2, \ldots, X_n \) are independent and are drawn from the same distribution with mean, \( \mu \), and variance, \( \sigma^2 \). Then, for each \( i \), \( E(X_i) = \mu \), and \( \text{var}(X_i) = \sigma^2 \). Let \( T = \sum_{i=1}^{n} X_i \). Then (a) above generalises to:

\[
E(T) = E\left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} E(X_i) = n\mu \; \text{and} \\
\text{var}(T) = \text{var}\left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \text{var}(X_i) = n\sigma^2.
\]

(c) Let \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \). Since \( \bar{X} \) is a linear transformation of \( T \),

\[
E(\bar{X}) = E\left( \frac{1}{n} T \right) = \frac{1}{n} E(T) = \frac{1}{n} n\mu = \mu \; \text{and} \\
\text{var}(\bar{X}) = \text{var}\left( \frac{1}{n} T \right) = \left( \frac{1}{n} \right)^2 \text{var}(T) = \frac{1}{n^2} n\sigma^2 = \sigma^2/n.
\]

(d) The Central Limit Theorem. The sum, \( T \), of \( n \) independent random variables, \( \{X_i\} \), has approximately the Normal distribution. Since \( \bar{X} \) is a linear transformation of \( T \), \( \bar{X} \) also is approximately Normally distributed. From (b) and (c), it follows that \( T \) is distributed as \( N( \mu, n\sigma^2) \), and \( X \text{ bar} \) is distributed as \( N(\mu, \sigma^2/n) \).

1. Sampling distributions

In order to make inferences about a population from sample data, we usually have to perform certain operations on the data. One operation is the linear transformation of one variable, \( X \), into another variable, \( Y \), where \( Y = a + bX \). The mean and standard deviation of \( Y \) are related to those of \( X \) through the formulae: \( \mu_Y = a + b\mu_X \), \( \sigma_Y = b\sigma_X \). In the proofs reported last week, we made much use of the facts that \( E(bX) = bE(X) \) and \( \text{var}(bX) = b^2\text{var}(X) \). Another operation is that of pooling two or more estimates of population parameters, such as \( \mu \), the population mean, and \( \sigma^2 \), the population variance, in order to get the ‘best’ estimate of the parameter. Let \( T_i \) be the sum of the \( n_i \) observations in the \( i \)’th sample, and let \( SS_i \) be the sum of squared deviations in the \( i \)’th sample. Then the ‘best’ estimates of \( \mu \) and \( \sigma^2 \) are, respectively,

\[
\bar{X} = \frac{T_1 + T_2 + \ldots}{n_1 + n_2 + \ldots}, \quad \text{and} \quad s^2 = \frac{SS_1 + SS_2 + \ldots}{n_1 - 1 + n_2 - 1 + \ldots}.
\]

Yet another operation, which was discussed last week, is that of adding two or more variables. (Let us agree that subtracting and averaging are special cases of ‘adding’. ) If, say, we toss a die twice and let \( X_1 \) and \( X_2 \) be the values obtained on the two tosses, we might be interested in the probability distribution of (i) the sum, \( T = X_1 + X_2 \) (total winnings), (ii) the average, \( \bar{X} = (X_1 + X_2)/2 \) (winnings per trial), (iii) the difference, \( D = X_1 - X_2 \) (change in fortunes), and so on. These distributions of sums of variables are referred to as sampling distributions so as to distinguish them from the original distribution of the \( \{X_i\} \), which we call the parent distribution. We use the distribution of \( X_i \) to derive the sampling distribution of sums of the \( X_i \). Very often, we know what the ‘shape’ of the
sampling distribution is (e.g., that the sampling distribution is the Normal curve) but we do not know the mean or the variance of the distribution. The formulae for the mean (i.e., expected value) and the variance of the sums are given above in Sec. 0. The standard deviation of the sampling distribution of a statistic (e.g., \( \bar{X} \)) is called the standard error (s.e.) of the statistic.

Suppose we take a sample of \( n \) observations, \( X_1, X_2, \ldots, X_n \), from a population that has a mean \( \mu \) and a variance \( \sigma^2 \). Let \( T = \sum X_i, \; \bar{X} = \frac{T}{n}, \; SS = \sum X_i^2 - \frac{T^2}{n}, \; s^2 = \frac{SS}{n-1} \). That is, \( \bar{X} \) is the mean of the sample, and \( s^2 \) is the variance of the sample. If we repeatedly draw samples of \( n \) observations and calculate the sample mean and variance, the sample means (\( \bar{X} \)) will vary across samples according to some sampling distribution, and the sample variances (\( s^2 \)) will vary across samples according to some different sampling distribution.

1.1. The sampling distribution of \( \bar{X} \).

In using \( \bar{X} \) to make inferences about the population mean, \( \mu \), we are justified in assuming that the sampling distribution of \( \bar{X} \) is Normal with mean \( \mu \) and variance \( \sigma^2/n \). That \( \bar{X} \) has the Normal distribution follows from the Central Limit Theorem. The variance of \( \bar{X} \) is \( \sigma^2/n \); therefore, the standard error of \( \bar{X} \) is \( \sigma/\sqrt{n} \).

Given these facts about the sampling distribution of \( \bar{X} \), it follows that

\[
Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}
\]

has the distrn, \( N(0, 1) \), i.e., the Standard Normal distrn. Note that, in order to standardise the raw statistic, \( \bar{X} \), we first express the raw statistic as a deviation from its mean, and then divide the deviation by the standard deviation (i.e., standard error) of the raw statistic.

1.2. The sampling distribution of \( \bar{X} \), when \( \sigma \) is unknown.

In order to convert \( \bar{X} \) into a Z-score, as in Sec. 1.1, we need to know the population standard deviation, \( \sigma \). But often we do not know \( \sigma \), and we are forced to use the sample s.d., \( s \), as an estimate of \( \sigma \). Let us use a different symbol, \( t \) (instead of \( Z \)), to define the ratio,

\[
t = \frac{\bar{X} - \mu}{s/\sqrt{n}}.
\]

Now, \( \bar{X} \) is a random variable (i.e., it would vary from sample to sample), and so is \( s \). \( Z \) is a ratio involving only 1 random variable, whereas \( t \) is a ratio involving that same random variable plus a second random variable, \( s \). Therefore, \( t \) has a greater variance than \( Z \). In other words, \( t \) has a different distribution from \( Z \), and it is called the Student-t distribution. The difference between the \( t \) and \( Z \) distributions is pronounced when \( n \) is ‘small’, e.g., \( n < 20-30 \), but is negligible when \( n \) is ‘large’. To summarise, if we are using \( \bar{X} \) to make inferences about \( \mu \) and we do not know \( \sigma \), then there are two cases to consider.

(i) If \( n \) is ‘large’, we can proceed as in Sec. 1.1, after substituting \( s \) for \( \sigma \), and assume that

\[
Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}
\]

has the distrn, \( N(0, 1) \), i.e., the Standard Normal distrn.

(ii) If \( n \) is ‘small’, we define

\[
t = \frac{\bar{X} - \mu}{s/\sqrt{n}}.
\]
and use the fact that $t$ has the Student-t distribution with $n-1$ degrees of freedom (df). This is a new distribution, and it is tabulated in a different format from the others we have used so far.

1.3. The sampling distribution of $s^2$.

In using $SS$ or $s^2$ to make inferences about the population variance, $\sigma^2$, we can assume that the sampling distribution of $SS/\sigma^2$ is chi-square with $(n-1)$ df only if the $X_i$ are Normally distributed. (See the Simulations Handout, Sec. B.) It can be shown that $E(s^2) = \sigma^2$. For this reason, we say that $s^2$ is an unbiased estimate of $\sigma^2$.

2. Confidence Intervals

2.1. Usefulness of confidence intervals (CI).

In the world of practical (e.g., policy) decisions, it is often the case that our motive for estimating a population parameter, such as, $\mu$, $\sigma$, or a proportion, $p$, is that we wish to make a decision, and that our decision ought to be based on the value of the parameter. Further, our decision is likely to be the same if, e.g., $\mu = 50.2$ as if $\mu = 50.9$ or 49.8. In other words, it is often the case that what would be most useful to us in our decision-making is not a point estimate, $\overline{X}$, of $\mu$, but rather an interval estimate of $\mu$, i.e., an interval that we are confident contains the true value, $\mu$. We call such an interval a confidence interval. (This ‘applied’ frame of mind is not very common within social science research.)

Suppose we wish to estimate $\mu$. Then, intuition suggests that we should define a confidence interval (CI) in the form $(\overline{X} - E, \overline{X} + E)$. How wide an interval do we prefer, i.e., what value of $E$ do we prefer? And how confident can we be that a given interval, $(\overline{X} - E, \overline{X} + E)$, contains the true value, $\mu$? These are related questions that we consider next.

Concerning the width of the CI and the associated confidence level, if the CI is very narrow, then we cannot have much confidence that it contains $\mu$. If we make the CI so wide that we are 99.99% confident that it contains $\mu$, the CI would be so wide as to be practically useless. A good compromise would be to aim for a ‘high’ but not ‘extremely high’ level of confidence: 95%, or 90%, or even 97.5%, depending on what or how much is at stake. Concerning the link between the amount of ‘stakes’ and the width of the CI, consider the following scenarios. (a) $\mu$ is the average time to run 40 yards (a-i) among all 25 year-olds, or (a-ii) for a particular athlete that is about to join your team. (b) $\mu$ is the average value of (b-i) my blood pressure, or (b-ii) a critical hormone. The ‘stakes’ are probably higher in (a-ii) than in (a-i), and we are likely to prefer or require a narrower CI in (a-ii) than in (a-i). For the same reason, we would probably require a narrower CI in (b-ii) than in (b-i). How to compute $E$, the half-width of the CI, for a given level of confidence?

2.1. Confidence intervals for $\mu$.

If we know the population s.d., $\sigma$, and we have the sample mean, $\overline{X}$, and sample size, $n$, then we can do more than offer $\overline{X}$ as an estimate of $\mu$. We can offer, as an interval estimate of $\mu$, the interval $\overline{X} \pm w(\sigma/\sqrt{n})$, where $w$ is some number to be determined. The confidence we can have in this interval estimate is equal to the probability that the interval contains the value of $\mu$. This latter probability can be found as follows:

$$P\left[\overline{X} - w(\sigma/\sqrt{n}) < \mu < \overline{X} + w(\sigma/\sqrt{n})\right] = P\left[-w(\sigma/\sqrt{n}) < \mu - \overline{X} < w(\sigma/\sqrt{n})\right]$$

$$= P\left(-w < \frac{\mu - \overline{X}}{\sigma/\sqrt{n}} < w\right) = P(-w < Z < w)$$

From the Normal Table, we see that, if $w = 1$, $P(-1 < Z < 1) = .68$, and we would have 68% confidence that the interval, $\overline{X} \pm (\sigma/\sqrt{n})$, contains $\mu$. If $w = 1.645$, we would have 90% confidence that the interval, contains $\mu$. If $w = 1.96$, we would have 95% confidence that the interval, $\overline{X} \pm 1.96(\sigma/\sqrt{n})$, contains $\mu$. (By the way, we often round up 1.96 to 2, which is why we can say that $P(-2 < Z < 2) \approx 0.95$.)
Scientists often report, e.g., ‘The estimated lifetime (or molecular weight, or toxic concentration) is 0.105, with a standard error of measurement of 0.003’. In this statement, 0.003 is $\sigma / \sqrt{n}$, the standard error of the mean.

2.2. Confidence intervals for $\mu$ when $\sigma$ is unknown.

If we do not know the population s.d., $\sigma$, then we would proceed as in Sec. 2.1 except that (i) we would substitute $s$ for $\sigma$, and (ii) use the appropriate critical values from the $t$-distribution with the appropriate $df$, instead of from the $Z$-distribution. The intervals have the form,

$$X \pm t_{n-1}(s / \sqrt{n}).$$

As discussed in Sec. 1.2., when $n$ is ‘large’, we get similar answers if we use the $Z$-distribution or the $t$-distribution.

2.5. Confidence intervals for $\sigma^2$.

It can be shown that, if the $X_i$’s are Normally distributed, then the sampling distribution of $SS/\sigma^2$ is chi-square with $n-1$ df. Given the $SS$ or $s^2$ from a sample, what is, e.g., a 90% confidence interval for $\sigma^2$? We use ‘symmetric’ confidence intervals $(a, b)$, such that $P(\sigma^2 < a) = P(\sigma^2 > b) = .05$. This implies that $P(a < \sigma^2 < b) = .9$. Given the $df$, $n-1$, and a complete Table of the chi-square distribution, one can find $c_1$ and $c_2$ such that

$$0.05 = P(\chi^2 < c_1) = P\left(\frac{SS}{\sigma^2} < c_1\right) = P\left(\sigma^2 > \frac{SS}{c_1}\right), \text{ and } 0.05 = P(\chi^2 > c_2) = P\left(\frac{SS}{\sigma^2} > c_2\right) = P\left(\sigma^2 > \frac{SS}{c_2}\right).$$

Therefore, $P\left(\frac{SS}{c_2} < \sigma^2 < \frac{SS}{c_1}\right) = 0.9$,

and the 90% confidence interval for $\sigma^2$ is $(SS/c_2, SS/c_1)$. For example, if $SS = 1000$ and $n = 20$, the chi-square Table gives (with 19 df) $c_1 = 10.1$, and $c_2 = 30.1$. Therefore, the 90% confidence interval is $(1000/30.1, 1000/10.1) = (33.2, 99)$. Typically, the confidence intervals for $\sigma^2$ are very wide. However, the ratio of the upper limit to the lower limit is usually between 2:1 and 3:1.

2.6 The Normal approximation to the Binomial distribution. (See previous reading.)

Let $p$ denote the probability that a tossed coin will come down Heads; $1-p = q$ is the probability of Tails. ($p = .5$ if the coin is fair.) If the coin is tossed $n$ times, let $r$ denote the number of Heads, and let $\hat{p} = r/n$ denote the observed proportion of Heads. The exact distribution of $r$ is called the Binomial distrn. However, the approximate distribution of $r/n$ is Normal with mean $np$ and variance $npq$. Also, the distribution of $\hat{p}$ is approximately Normal with mean $p$ and variance $pq/n$. The standard error of $\hat{p}$ is $\sqrt{pq/n}$. It follows that

$$Z = \frac{r - np}{\sqrt{npq}}, \text{ and, on dividing the numerator and denominator by } n,$$

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

are both Standard Normal random variables. We should use the ‘continuity correction’ here whenever $n$ is ‘small’, e.g., less than 50. (You should check to see if using the correction gives a result very different from not using the correction; you will find that for ‘large’ $n$ there is little difference.)

2.7. Confidence intervals for $p$. 
If we do not know the parameter, \( p \), [e.g., \( \Pr(\text{Heads}) \), \( \Pr(\text{Correct}) \), \( \Pr(\text{a randomly chosen person favors Measure 1}) \)], and we have the sample proportion, \( \hat{p} \), and sample size, \( n \), then we can do more than offer \( \hat{p} \) as an estimate of \( p \). We can offer, as an interval estimate of \( p \), the interval \( \hat{p} \pm w \sqrt{\frac{pq}{n}} \), where, as above, \( w \) determines how much confidence we can have in this interval estimate. A minor difficulty we have here is that, because we don’t know \( p \), we don’t know \( pq \). We can solve this problem in either of two ways.

(i) We can use \( \hat{p} \) and \( \hat{q} \) instead of \( p \) and \( q \); the s.e. is then estimated as \( \hat{p} \hat{q} \sqrt{\frac{1}{n}} \).

(ii) My preferred method is as follows. Note that \( pq \leq 0.25 \) for all \( p \), and it is close to 0.25 for a wide range of values of \( p \); e.g., when \( p = 0.3 \) or \( 0.7 \), \( pq = 0.21 \), which is close to 0.25. Therefore, replacing \( pq \) by 0.25 in the formula for the confidence interval would give us a slightly larger interval, i.e., a ‘conservative’ estimate of the confidence interval. I would replace 1.96 (see para. 2.1 above) by the slightly larger 2.0, and \( \sqrt{pq} \) by the slightly larger \( \sqrt{0.25} = 0.5 \), to get:

A conservative estimate of the 95% confidence interval for \( p \) is \( \hat{p} \pm \frac{1}{\sqrt{n}} \).

In Gallup and other polls, the expression, \( \pm \frac{1}{\sqrt{n}} \), is the so-called ‘margin of error’ of the poll. Please check for yourself that the margin is 10%, 4.5%, 3.5%, and 3% when \( n \) is 100, 500, 900, and 1000, respectively. When next you hear, “The latest polls say that 38% of the nation favors the Trade Policy, based on about \( n \) responses”, you can then check that the reported ‘margin of error’ is \( \frac{1}{\sqrt{n}} \).

3. 1- and 2-tailed tests of null hypotheses

The rejection region, \( R \), of a test. Let us consider the question: \( H_0 \) is either true or it is false; if \( H_0 \) is false, what alternative hypothesis, \( H_1 \), am I prepared to accept in place of \( H_0 \)? The answer to this question will define, in part, the ‘shape’ of the rejection region, \( R \). To see what ‘shapes’ \( R \) might have, let us consider the following problems, all involving the use of \( Z \) as our test statistic. (The same logic is used when \( t \) is the test statistic.)

(1) A casino “coin” is or is not biased. Toss it \( n \) times, count the number, \( r \), of “Heads”, and compute the test statistic, \( Z = \frac{r - np}{\sqrt{npq}} \). The null is, \( H_0 : p = 0.5 \); what is our alternative, \( H_1 \)?

A reasonable answer is: If the coin is biased, i.e., if \( H_0 \) is false, the bias can be in favor of ‘heads’ or of ‘tails’. In other words, if \( p \neq 0.5 \), then \( p \) can be greater than 0.5, or less than 0.5, and we have no reason to prefer the ‘\( p < 0.5 \)’ possibility over the ‘\( p > 0.5 \)’ possibility. For this reason, we state the alternative as

\[
H_1 : p \neq 0.5.
\]

In this form of \( H_1 \), we are prepared to (i) accept \( H_0 \) if the \( Z \) score is ‘close’ to 0, or (ii) reject \( H_0 \) if either \( Z \) is large negative (e.g., \( Z = -2 \) or less) or \( Z \) is large positive (e.g., \( Z = 2 \) or more). The rejection region, \( R \), is defined as the set of \( Z \) values that lead us to reject \( H_0 \). In this problem, \( R \) consists of the 2 tails of the \( Z \) distribution:

\[
R : Z < -2 \text{ or } Z > 2.
\]

For this reason we refer to this \( H_1 \) as a 2-tailed alternative, and to the test as a 2-tailed test.

(2) A person may or may not have ESP. Give the person \( n \) Yes/No tasks, count the number, \( r \), of “Correct” answers, and compute the test statistic, \( Z = \frac{r - np}{\sqrt{npq}} \). The null is, \( H_0 : p = 0.5 \) (i.e., the person is guessing); what is our alternative, \( H_1 \)?

A reasonable answer is: If the person is not guessing, i.e., if \( H_0 \) is false, the person must have ESP. In other words, if \( p \neq 0.5 \), then \( p \) must be greater than 0.5. For this reason, we state the alternative as

\[
H_1 : p > 0.5.
\]
In this form of $H_1$, we are prepared to (i) accept $H_0$ if the $Z$ score is ‘close’ to 0, or (ii) reject $H_0$ if $Z$ is large positive (e.g., $Z = 2$ or more). The rejection region, $R$, is defined as the set of $Z$ values that lead us to reject $H \text{ sub } 0$. In this problem, $R$ consists of the upper tail of the $Z$ distribution:

$$R: Z > 2.$$  

For this reason we refer to this $H_1$ as a 1-tailed alternative, and to the test as a 1-tailed test.

(3) A patient’s clinical status is checked before and after treatment by a new drug. ‘Heads’ = patient improves, ‘Tails’ = patient gets worse. Under $H_0$, the drug is ineffective, and a patient would be just as likely to show improvement as to get worse. As before, let $n$ be the number of patients, count the number, $r$, of improvements, and compute the test statistic, $Z = (r - np)\sqrt{npq}$. The null is, $H_0: p = 0.5$ (i.e., the drug is ineffective); what is our alternative, $H_1$?

A reasonable answer is: If the drug is not ineffective, i.e., if $H_0$ is false, the new drug must be effective. In other words, if $p \neq 0.5$, then $p$ must be greater than 0.5. For this reason, we would state the alternative as

$$H_{1a}: p > 0.5,$$

which leads to a 1-tailed R and a 1-tailed test. However, another reasonable answer is: If the drug is not ineffective, i.e., if $H_0$ is false, the new drug may be effective or it may be harmful. In other words, if $p \neq 0.5$, then $p$ may be greater than 0.5 or less than 0.5. For this reason, we would state the alternative as

$$H_{1b}: p \neq 0.5,$$

which leads to a 2-tailed R and a 2-tailed test.

This last example (3) shows that two reasonable statisticians can disagree in their statement of $H_1$ and, therefore, differ in their definition of the rejection region, $R$. We shall see below that, for a given significance level (e.g., $\alpha = .05$), it is easier to reject $H_0$ with a 1-tailed test than a 2-tailed test. In other words, for a given sample and given $\alpha$, a person using a 1-tailed test might be able to reject $H_0$, but one using a 2-tailed test might not be able to reject $H_0$. This sometimes leads to controversy.

The significance level of a test. We want the probability to be small that $Z$ will fall in the rejection region, $R$, if $H_0$ is true; this probability is called the significance level of our test, and it is denoted by $\alpha$.

$$\alpha = P(Z \text{ falls in } R | H_0 \text{ is true}).$$

By convention, we set $\alpha$ equal to a small value, such as .05 or .001, depending on the relative costs of the two types of error we can make. These errors are rejecting $H_0$ when it is true, and retaining $H_0$ when it is false. For a given value of $\alpha$, e.g., .05, $R$ depends on whether our alternative hypothesis is 1- or 2-tailed. For example, if our alternative is $H_{1a}$ and we choose $\alpha = .05$ for our study, then $R$ is the interval $Z < -c$, where $c$ is such that $P(Z < -c) = .05$; i.e., $c = 1.645$. If our alternative is $H_{1b}$ and $\alpha = .05$, $R$ is the interval $(Z < -c \text{ or } Z > c)$, where $c$ is such that $P(Z < -c \text{ or } Z > c) = .05$; i.e., $c = 1.96$.

In general, if $\alpha$ is the significance level of the test, and if $z_\alpha$ denotes the (positive) $z$-score that has an area, $p$, to the right of it [i.e., $p = P(Z > z_\alpha)$], then:

1-tailed test. $R: Z > z_\alpha$.

2-tailed test. $R: Z < -z_\alpha$ OR $Z > z_\alpha$.

Because $z_\alpha < z_{\alpha/2}$ (e.g., $1.645 < 1.96$), it is “easier” to reject $H_0$ when you use a 1-tailed test than when you use a 2-tailed test.

Confidence intervals and 2-tailed tests. We noted in an earlier Handout that a confidence interval for, say, $\mu$, gives the set of values of $\mu$ that are consistent with the data ($n$, $\bar{x}$, $s$); values of $\mu$ that lie outside the
confidence interval are inconsistent with the data and, therefore, can be rejected as implausible. This use of confidence intervals leads to the conclusion: a $100(1 - \alpha)\%$ CI implies a rejection region, $R$, that is the same as the $R$ corresponding to a 2-tailed test having significance level $\alpha$.

4. The 1-sample $z$ and $t$-tests

See Aron & Aron, Chapter 6 and 7.

5. A theoretical note on $Z$ and $\chi^2$

Almost all statistical tests involve the calculation of a standardised test statistic, $Z = (\text{Statistic} - \text{Exp. Value})/(\text{Standard error})$; even the chi-square test involves such a calculation. Let us prove this assertion when there are 2 categories, say, Heads and Tails. On each trial, $p = P(\text{Heads})$, $q = 1 - p = P(\text{Tails})$. There are $n$ trials, $r$ of them come up Heads.

To test a null hypothesis that $p$ has a specific value, we would compute $Z = (r - np)/\sqrt{npq}$, and see if $Z$ is ‘large’ (e.g., greater than 2 in absolute value) or ‘small’. Alternatively, we could use the chi-square goodness-of-fit test that we learned in Week 1. There are $k = 2$ categories; $O_1 = r$, $O_2 = n - r$; and $E_1 = np$, $E_2 = nq$, if the null hypothesis is true. Therefore,

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(r - np)^2}{np} + \frac{(n - r - nq)^2}{nq}$$

$$= \frac{(r - np)^2}{n} \left( \frac{1}{p} + \frac{1}{q} \right) = \frac{(r - np)^2}{npq} = Z^2$$

We have shown that $Z^2 = \chi^2$, i.e., that $Z^2$ has the chi-square distribution with 1 df. More generally, if the $Z_i$’s are independent, standard Normal random variables, then $\sum Z_i^2$ is distributed as $\chi^2$. [One consequence of this analysis of $\chi^2$ into a sum of $k$ variables is that the Central Limit Theorem applies to yield: for $k$ large (e.g., 10 or more), the distribution of $\chi^2$ is approximately Normal. It is proved in the more advanced texts that $E(\chi_i^2) = k$, and $\text{var}(\chi_i^2) = 2k$. Therefore, $\chi_i^2 \sim N(k, 2k)$, for $k$ large.]