Session 2 - Hypothesis Testing

Brian Leroux

Wednesday, January 16, 2019

Outline

1. Hypothesis testing for a population proportion
2. Issues in the interpretation of hypothesis tests
3. Hypothesis testing for a population mean with known variance
4. Hypothesis testing for a population mean with unknown variance

1. Hypothesis testing for a population proportion

- Hypothesis Testing vs Confidence Intervals
- 1-Sample Test for a Population Proportion (Binomial Test)
- Null and Alternative Hypotheses
- Type I and Type II Errors
- Rejection Rules
- Power
- Using the Normal Approximation for a Test of a Proportion

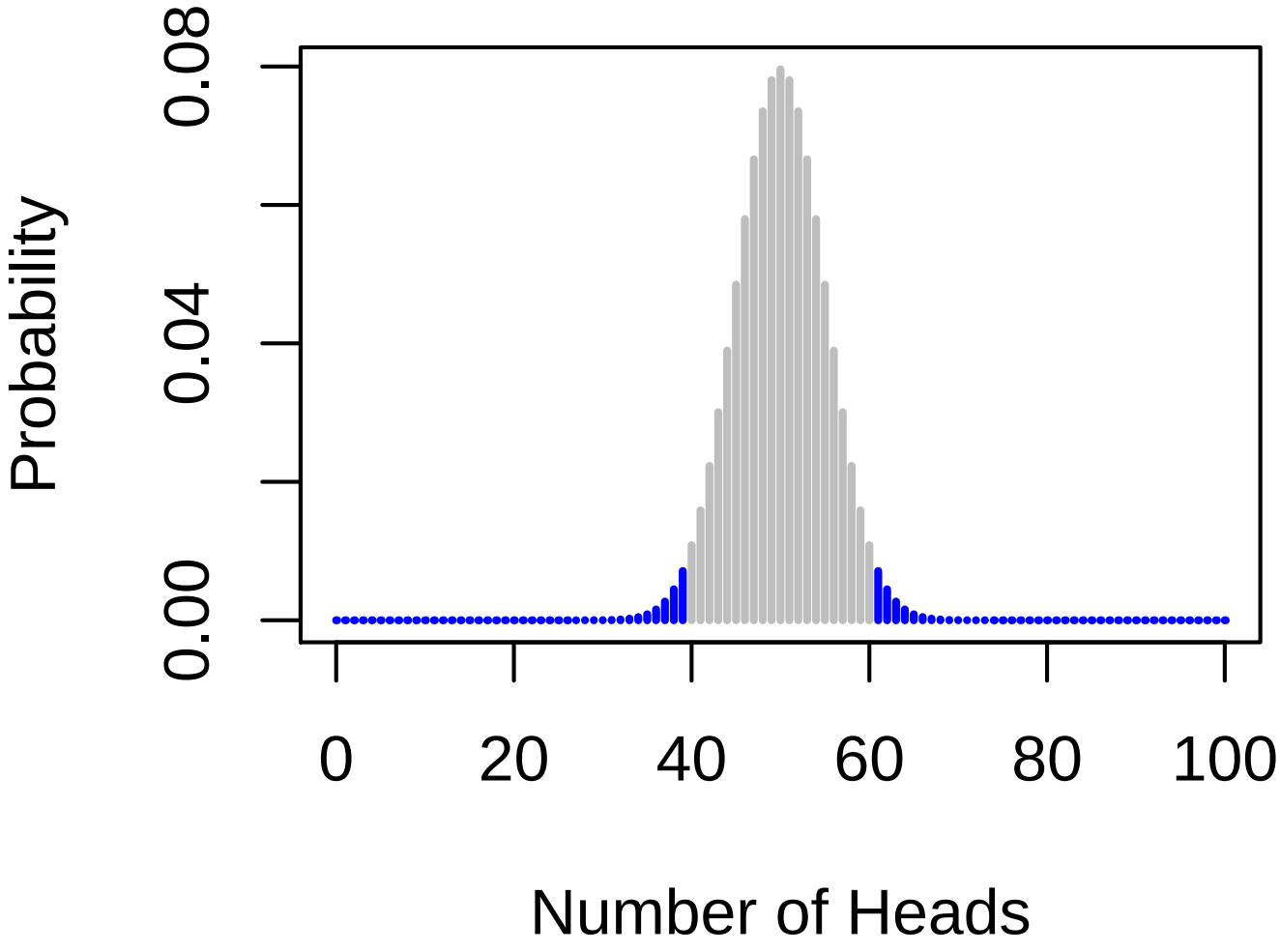
Hypothesis testing versus confidence intervals

Hypothesis tests are used to answer questions about a population (or about differences between 2 or more populations). Whereas a confidence interval gives a range of plausible values for a population parameter, hypothesis tests give a yes/no answer to a question (hypothesis) about the parameter.

Example: suppose we want to test whether a coin is fair (equal probability of heads or tails). We could toss the coin a large number of times and record the number of heads. A confidence interval could be used to describe a range of plausible values for the probability of heads. However, if we need to make a decision as to whether or not the coin is fair, a hypothesis test will be more useful.

Is the coin fair?

Toss a coin $n = 100$ times. Let X be the number of heads. **If the coin is fair**, we expect $X = 50$, and the sampling distribution of X is Binomial($n = 100, p = 0.5$). A value in the blue region can be interpreted as evidence that the coin is not fair.



A Decision Rule

A graph of the Binom(100,0.5) distribution (previous slide) suggests that we should expect to get between 40-60 heads most of the time. Therefore, we could make a decision rule as follows.

> *If* $X < 40$ *or* $X > 60$, *declare that the coin is not fair.*
>
> *If* $40 \leq X \leq 60$, *declare that we* ***do not have evidence*** *that the coin is unfair.*

Note: in the latter case we do **not** make the claim that the coin is fair, only that we **don't have evidence that it is unfair**. The reason is that the probability of a head might be different than 0.5, but *not enough different to be detected by this experiment*.

This test procedure is called a *1-sample test for a population proportion*, or is sometimes called the *Binomial test for a proportion*.

Null and Alternative Hypotheses

In hypothesis testing terminology for the coin tossing experiment, we say that the **null hypothesis** is that the probability (p) of a head is 0.5. In mathematical notation this is written as

$$H_0 : p = 0.5$$

The **alternative hypothesis** is that the probability is not equal to 0.5, i.e.,

$$H_1 : p \neq 0.5$$

Note: This is an example of a *2-sided alternative hypothesis* because it specifies values on both sides of 0.5 (the null hypothesis value). Most hypothesis tests are 2-sided. However, 1-sided hypothesis tests are sometimes more appropriate (we will see examples later).

To Reject or Not To Reject

To conduct a hypothesis test we need to design a **rejection rule**, which tells us whether or not to reject the null hypothesis. Thus there are two possible outcomes of a hypothesis test:

1. Reject the null hypothesis
2. Do not reject the null hypothesis

If we do not reject the null hypothesis, is it ok to say that we accept it?

No!

It is wrong to say that we accept the null hypothesis when we fail to reject it.

There is a useful analogy with a courtroom trial: the accused is either found "guilty" or "not guilty" but a judge never rules that the accused is "innocent".

Rejection Rule for the Coin Tossing Experiment

Rejection Rule:

If $X < 40$ or $X > 60$, reject the null hypothesis.[1]

[1] By default, if $40 \geq X \leq 60$ we do not reject the null hypothesis.

Type I and Type II Errors

There are two possible true states of nature and two possible decisions so 4 possible situations.

True State of Nature	Reject H_0	Do Not Reject H_0
H_0 is True	Type I Error	Correct Decision
H_0 is False	Correct Decision	Type II Error

Type I Error: rejecting a null hypothesis when it is true

Type II Error: not rejecting a null hypothesis when it is false

Type I Error Probability

The type I error probability is the probability of rejecting H_0 when it is true.

> ***Type I Error Probability*** $=$ *Probability of rejecting the null hypothesis when it is true*

The Type I error probability is also called the *significance level*. It is usually denoted by α.

In design of experiments, we wish to control the type I error probability. We often design the experiment so that the type I error probability is equal to a small number, with 0.05 being the most common choice based on historical precedent. Other values such as 0.025 or 0.01 are often commonly used.

R.A. Fisher and the 0.05 Level of Significance

THE ARRANGEMENT OF FIELD EXPERIMENTS

R. A. FISHER, Sc.D.,
Rothamsted Experimental Station.

"When is a Result Significant?"

called the verge of significance; for it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." This level, which we may call the 5 per cent. point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials. To

Type II Error Probability

The type II error probability is the probability of *not* rejecting H_0 when it is false.

> ***Type II Error Probability*** $=$ *Probability of not rejecting the null hypothesis when it is false.*

The Type II error probability is sometime denoted by β and usually described in terms of its complementary probability, i.e., the probability of not making a type II error, also called the **power**.

> ***Power*** $=$ *Probability of rejecting the null hypothesis when it is false.*

In experimental design, we first control our type I error at the desired level (e.g., 0.05), and then we try to achieve a high value for the power, e.g., 0.8 or 0.9, or higher, depending on the context. Achieving high power typically requires selecting a sufficiently large sample size.

Type I Error Probability for the Coin Tossing Experiment

For the coin tossing experiment, the rejection rule is the event $\{X < 40 \text{ or } X > 60\}$. (The set of values $X < 40$ or $X > 60$ is called the *rejection region* for X.) The type I error probability (α) is the probability that X falls in the rejection region, given that H_0 is true (i.e., $p = 0.5$). This is calculated using the binomial distribution as follows.[2]

$$
\begin{aligned}
\alpha &= \text{P(Type I Error|p=0.5)} \\
&= \text{P}(X < 40 \text{ or } X > 60 | p = 0.5) \\
&= P(X = 0|p = 0.5) + P(X = 1|p = 0.5) + \cdots + P(X = 39|p = 0.5) \\
&\quad + P(X = 61|p = 0.5) + P(X = 62|p = 0.5) + \cdots P(X = 100|p = 0.5) \\
&= 0.035
\end{aligned}
$$

[2] Note on mathematical notation: P(event | condition) here is interpreted as the probability of the event (e.g., $X = 0$) under the given condition ($p = 0.5$). This is different than a conditional probability in which the condition is a random event rather than a statement about the value of a parameter.

Calculating the Type I Error Probability

```
n=100
p0=0.5
rejection_region = c(0:39,61:100)
sum(dbinom(rejection_region, size=n, p=p0))
```

```
[1] 0.0352002
```

This is considered a sufficiently small significance level. Note that in this case we cannot achieve a significance level exactly equal to 0.05. If we try to widen the rejection region to $X < 41$ or $X > 59$ then the significance level becomes too large (based on the 0.05 convention):[3]

```
n=100
p0=0.5
rejection_region = c(0:40,60:100)
sum(dbinom(rejection_region, size=n, p=p0))
```

```
[1] 0.05688793
```

[3] Note that we added values of 40 and 60 so that the new rejection region is still symmetric around the mean value of 50. This is important because values of equal distance away from 50 represent equivalent evidence against the null hypothesis.

Calculation of Power for the Coin Tossing Experiment

Power is the probability of rejecting the null hypothesis when it is false. The null hypothesis can be false in an infinite number of ways (any value of p different than 0.5 makes the null hypothesis false).

Therefore, power is calculated under a selected hypothetical alternative hypothesis. For example, if $p = 0.7$, what would be the power of our test? The power is the probability of rejecting H_0 under the assumption that $p = 0.7$ rather than 0.5 as used to calculate α.

$$
\begin{aligned}
\text{Power} &= P(\text{Reject } H_0 | p = 0.7) \\
&= \text{P}(X < 40 \text{ or } X > 60 | p = 0.7) \\
&= P(X = 0 | p = 0.7) + \cdots + P(X = 39 | p = 0.7) \\
&\quad + P(X = 61 | p = 0.7) + \cdots P(X = 100 | p = 0.7) \\
&= 0.979
\end{aligned}
$$

```
sum(dbinom(c(0:39,61:100),size=100,p=0.7))
```

```
[1] 0.9790114
```

Experimental Design

When designing an experiment to test an hypothesis, we first insist on a desirable level for the Type I Error probability, and then we try to achieve a desirably level for the power. This involves selecting an appropriate sample size.

For example, suppose that we wanted to be more conservative with our coin tossing experiment, i.e.,. we want to reject H_0 only about 2% of the time when it is true, and to have at least 98% power.

To achieve this, first make the rejection region smaller (to reduce α), e.g., $X < 39$ or $X > 61$. Then α becomes

```
sum(dbinom(c(0:38,62:100),size=100,p=0.5))
```

```
[1] 0.02097874
```

Thus, we have met our objective of α being approximately 0.02. But what is the impact on the power?

Impact of reducing α on the power

We reduced the chance of a type I error by making the rejection region smaller but this also reduces the power.

```
sum(dbinom(c(0:38,62:100),size=100,p=0.7))
```

```
## [1] 0.966021
```

Now the power has been reduced and is less than the desired 98%. This is what typically happens. There is a trade-off between significance level and power.

To achieve our goals for both significance level and power we would have to increase the sample size and make a new rejection region. To facillitate the required calculations, we take advantage of the Central Limit Theorem and the normal approximation to the binomial distribution, which we will consider next.

Using the Normal Approximation to the Binomial Distribution

By the Central Limit Theorem we know that the binomial distribution can be well approximated by a normal distribution if n is large enough. Therefore, we can approximate the type I error probability using a normal approximation.

Under the null hypothesis, X has a Binomial(n, p) distribution, which has mean np and variance $np(1-p)$. Therefore, if we standardize X by transforming it into a Z-score as follows

$$Z = \frac{X - np}{\sqrt{np(1-p)}},$$

the variable Z will have an approximate standard normal distribution, i.e., N(0,1), with mean 0 and variance 1, **if the null hypothesis is true**.

We can take advantage of this fact to

1. approximate the type I error probability (α) for a given rejection region, or
2. determine what the rejection region should be for a given value of α.
3. determine the sample size required for a given α and power (power calculations to be covered in a later session)

Using the Normal Approximation to Approximate α

Consider the coin tossing experiment with $H_0 : p = 0.5$ and $n = 100$. In that case, Z becomes

$$Z = \frac{X - 50}{5}.$$

The rejection region $X < 40$ or $X > 60$ can be written as $X - 50 < 10$ or $X - 50 > 10$, i.e., $|X - 50| > 10$, or $|Z| > 2$.

$$\text{Rejection Region: } \{X < 40 \text{ or } X > 60\} \quad = \quad \{|Z| > 2\}.$$

This makes it simply to calculate the type I error probability. Because Z has an approximate normal distribution under H_0 the significance level of the test is equal to the two-sided tail probability beyond 2 for the normal distribution:

```
2*(1-pnorm(2))
```

```
## [1] 0.04550026
```

This gives a reasonable approximation to the value of 0.035 calculated using the binomial distribution.

Using the Normal Approximation to Derive the Rejection Rule

Suppose that a manufacturing process produces a proportion of defective items equal to $p_0 = 0.04$. After a change in the process is made it is desired to know if the defective proportion has changed. To evaluate this we define the null and alternative hypotheses as $H_0 : p = p_0$ and $H_1 : p \neq p_0$. If X is the number of defectives in a sample of $n = 500$ items, the rejection region can be defined in terms of the Z-score $Z = (X - np_0)/\sqrt{np_0(1-p_0)} = (X - 20)/4.38$. We can use the standard normal approximation to Z to define the rejection rule.

If we want $\alpha = 0.05$, our rejection rule would be $|Z| > 1.96$. This is equivalent to $X < 20 - 1.96 \times 4.38$ or $X > 20 + 1.96 \times 4.38$, or equivalently, $X < 11.4$ or $X > 28.6$. Because X takes on integer values this is equivalent to $X \leq 11$ or $X \geq 29$. Now we can calculate the exact significance level using the binomial probabilities.

```
sum(dbinom(c(0:11,29:500),size=500,p=0.04))
```

```
[1] 0.0509379
```

This is very close to the theoretical value of 0.05 from the normal distribution.

2. Issues in the Interpretation of Hypothesis Tests

- Why it is wrong to accept the null hypothesis
- The confidence interval as complement to the hypothesis test
- The p-value
- The abuse of p-values

Why it is Wrong to Accept the Null Hypothesis

The two possible outcomes of an hypothesis test are:

Reject the null hypothesis, or

Do not reject the null hypothesis

When H_0 is not rejected, the result is sometimes incorrectly interpreted as "we accept the null hypothesis". This is a dangerous interpretation for a few reasons. For one, the null hypothesis specifies a single value for a population parameter (e.g., probability of a defective is equal to 0.04). It is rare that any experiment could be used to specify the value of a population parameter exactly. In experimentation there is always uncertainty.

Because of the inherent uncertaintly in any experiment, it is more appropriate to conclude "we do not reject H_0" rather than "we accept H_0". It might seem awkward to use a double negative ("not reject"); however, it is an important nuance in the proper interpretation of the results.

Recall the analogy with a courtroom trial: the accused is either found "guilty" or "not guilty" but a judge never rules that the accused is "innocent".

Example

In the coin tossing example, suppose that we observed $X = 55$ out of $n = 100$ heads. This is not exactly equal to the 50 we expected, but it is not enough of a discrepancy to cause us to doubt our null hypothesis that the coin is fair. Note that 55 is not in the rejection region defined as $X < 40$ or $X > 60$. Thus, we would conclude that we don't have any evidence that the coin is unfair.

But have we proved that the coin is in fact a fair coin? No, because the probability of a head could be p=0.55 or some other value close to 0.5. Hence, instead of "accepting"" the null hypothesis and concluding that the coin is fair, we should state only that there is no evidence that it is unfair.

The confidence interval is a useful complement to the hypothesis test and helps with interpretation of results when we fail to reject the null hypothesis.

Using the Confidence Interval to Aid with Interpretation

A confidence interval provides additional information beyond the hypothesis test. In the above example, the 95% confidence interval for the probability of a head is (0.45, 0.65). This shows clearly that there is a lot of uncertainty about the true value of the probability. The value 0.5 is a plausible value but so are values quite different from 0.5.

```
phat=0.55
n=100
se=sqrt(phat*(1-phat)/n)
phat-1.96*se
```

```
[1] 0.4524912
```

```
phat+1.96*se
```

```
[1] 0.6475088
```

Relationship between confidence intervals and hypothesis tests

In general, we can interpret a confidence interval as the set of all values of the population parameter that would *not* have been rejected by the corresponding hypothesis test.

For example, 0.52 is a value in the confidence interval; if we did a test of the null hypothesis $H_0 : p = 0.52$, we would **not** reject this H_0.

The confidence interval and hypothesis test complement each other. In most applications, it is a good idea to report the results of both.

The P-Value

The p-value is another additional piece of information that is useful for interpreting the results of an hypothesis test. As described up to now, an hypothesis test results in a dichotomous decision: either reject H_0 or do not reject H_0.

The *p*-value provides a quantitative assessment of the evidence against the null hypothesis. It is defined as follows:

> *The p-value is the probability, calculated under the null hypothesis, of the test statistic being as extreme or more extreme than the value that was observed.*

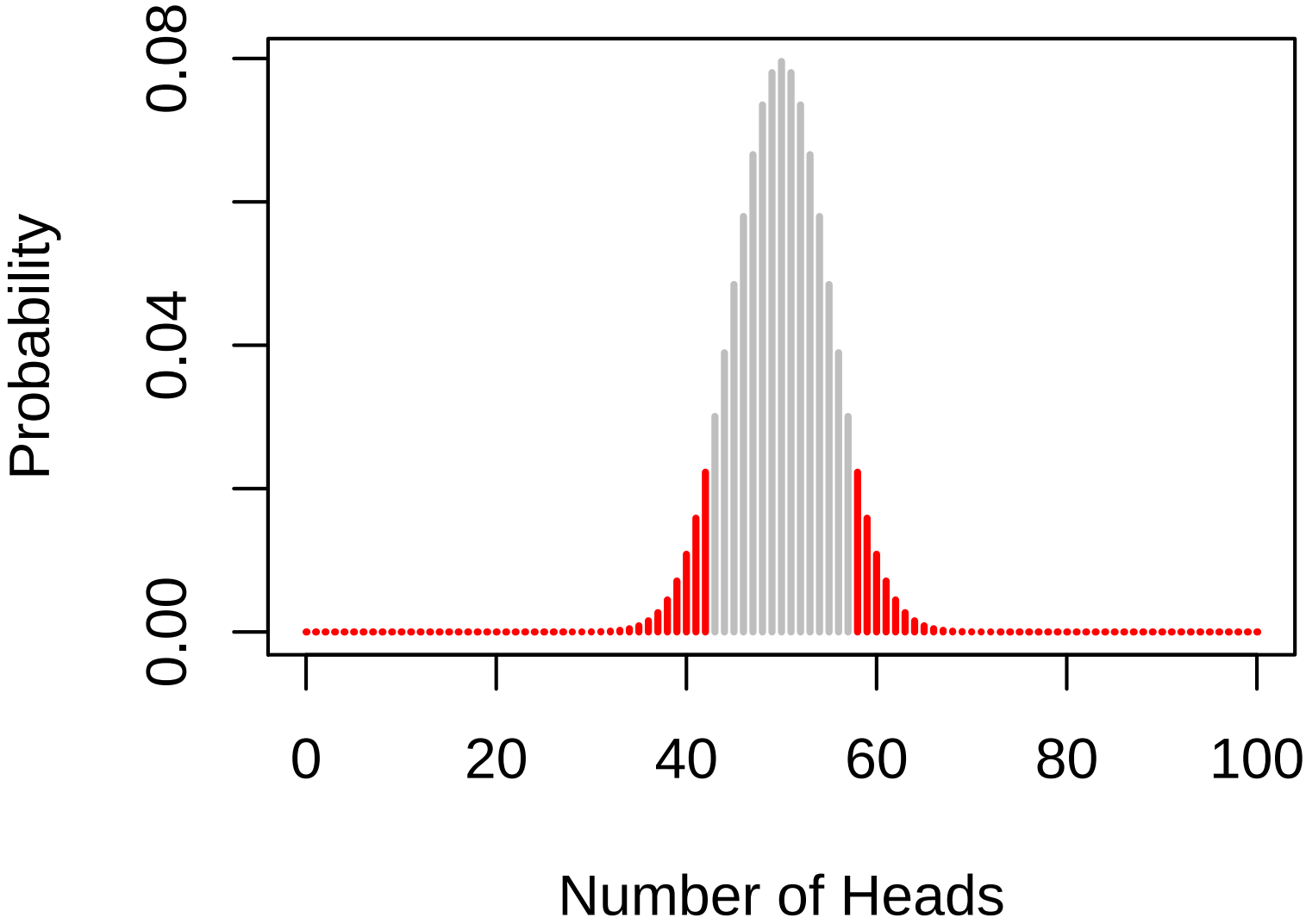
Key Properties of the p-Value

1. The p-value is a probability statement about the test statistic.
2. "as extreme or more extreme" is interpreted as indicative of as great or greater departures from the null hypothesis.
3. The p-value is calculated **under the null hypothesis**

For the coin-tossing experiment: 1) the test statistic is X, the number of heads; 2) "as estreme or more extreme" means "as far away or further away from 50"; and 3) the p-value is calculated with a probability of a head equal to 0.5.

Illustration of the p-value

The Binomial(n=100,p=0.5) distribution is displayed below, which is the distribution of X under H_0. If the observed result is $X = 44$, the p-value is the probability shaded in red.



Calculation of the p-value

For the observed result $X = 44$ heads in $n = 100$ tosses of a coin, the p-value is calculated as follows.

```
sum(dbinom(c(0:44,56:100),size=100,p=0.5))
```

```
[1] 0.271253
```

A p-value of 0.27 means that there is about a 27% chance of getting a result as or more extreme than 44 heads in 100 tosses, if the coin is fair. This probability is not small enough to cause us to question the null hypothesis. Note that this agrees with our original decision rule, which was to reject the null hypothesis only if $X < 40$ or $X > 60$.

Relationship between the *p*-value and the significance level

In the example we saw that we get the same interpretation whether we use the *p*-value or the rejection region to make our decision about whether or not to reject the null hypothesis. Recall that our rejection region ($X < 40$ or $X > 60$) had probability 0.035 under the null hypothesis. We called this number the significance level of the test and use the notation α for it.

Any value of X not in the rejection region (e.g., $X = 44$) will **by definition** have a *p*-value larger than 0.035, andn any value of X in the rejection region will have a *p*-value smaller than 0.035. This is a general rule:

> *$p < \alpha$ if and only if X lies in the rejection region with significance level α.*

What the *p*-value is not

The *p*-value is sometimes interpreted as "the probability that the null hypothesis is true." This is wrong! For one thing, we are treating the null hypothesis as a state of nature, which is either true or false, but we are not associating probabilities with it. The random quantity is not the hypothesis but the test statistic.

The *p*-value is a probability about the test statistic, calculated **under the assumption that the null hypothesis is true**. It is a "what if" calculation, i.e., "what if the null hypothesis is true, is our observed result unlikely or not?".

Abuse of p-values

In many fields of scientific research, there is a problem with overuse of p-values (and hypothesis testing in general)

One aspect of this problem is called the multiple-testing problem: if a large number of hypothesis tests are performed, then the null hypothesis will be rejected in some cases just by chance (approximately 5% of them if $\alpha = 0.05$).

There are specialized methods to mitigate the multiple-testing problem, but no completely satisfying solution.

The abuse of p-values is one factor that has contributed to a crisis in reproducibility in scientific research.

A crisis of reproducibility in scientific research?

Ioannidis, JPA: Why most published research findings are false. PLOS MEDICINE 2(8):696-701, 2005.

"There is increasing concern that most current published research findings are false."

"... a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes and analytical modes; when there is greater financial and other interest and prejudices and when more teams are involved in a scientific field in chase of statistical significance."

"Simulations show that for most study designs and settings it is more likely for a research claim to be false than true."

3. Hypothesis Testing for a Population Mean

- 1-Sample Test for a Population Mean with Known Variance
- Using the CLT
- Simulation studies of the performance of the test

1-Sample Test for a Population Mean with Known Variance

Example: suppose that we want to test whether or not there is an effect of a psychiatric medication on weight. We define the null hypothesis as $H_0 : \mu = 0$, where μ is the population mean weight loss in pounds, i.e., the mean weight loss in a hypothetical (infinite) population of patients who might take this drug in the future. Because we don't know if the drug can cause weight gain or weight loss, we set the alternative hypothesis to be $H_1 : \mu \neq 0$ (i.e., H_1 means that $\mu < 0$ or $\mu > 0$).

Our experiment will consist of measuring weights in pounds of a sample of n patients before they start taking the drug and then again a fixed time (e.g., 3 months) after starting the drug. The variable used to conduct the test is the change in weight
($X =$ weight at 3 months $-$ pre-treatment weight). Then μ is the expected value of X, and H_0 says that the mean weight change is 0.

The Assumption of Known Variance

The variable X will have a certain variance (σ^2). Assume for now that we know the value of the variance, say $\sigma^2 = 25$. For example, this might be a reasonable assumption if we have done many similar experiments in the past and the variance in weight change has been consistent. (It is more common in practice to use the sample variance as an estimate of σ^2, which has an impact on the test procedure as we will see later.)

Deriving the Rejection Rule for the Weight Loss Experiment

Because we are interested in determining if the population mean weight change is equal to 0 or not, a natural decision rule is to reject H_0 if the sample mean $\bar{X}$ is larger than we might expect. How do we calculate how large a value of $\bar{X}$ to expect? Use the CLT!

Under H_0, if n is large enough, $\bar{X}$ will have a normal distribution with mean equal to 0 (because $\mu = 0$ under H_0) and standard deviation equal to the SE, i.e., $\sigma/\sqrt{n}$, where σ is the population standard deviation of weight change.

Therefore, by the well-known property of the normal distribution, $\bar{X}$ will lie within $\pm 1.96 \times SE(\bar{X})$ with probability 0.95:

$$P(-1.96 \times \sigma/\sqrt{n} < \bar{X} < 1.96 \times \sigma/\sqrt{n}) \approx 0.95$$

Rejection Rule for the Weight Loss Experiment

Restating the above property, we have that
$P(\bar{X} < -1.96 \times \sigma/\sqrt{n}$ or $\bar{X} > 1.96 \times \sigma/\sqrt{n}) \approx 0.05$.

Therefore, we will have type I error probability (significance level) $\alpha = 0.05$ for the following rejection rule:

$$\text{Reject } H_0 : \mu = 0 \text{ if } \bar{X} < -1.96 \times \sigma/\sqrt{n} \text{ or } \bar{X} > 1.96 \times \sigma/\sqrt{n}$$

Equivalently,

$$\text{Reject } H_0 : \mu = 0 \text{ if } |\bar{X}| > 1.96 \times \sigma/\sqrt{n}$$

For example, with $\sigma^2 = 25$ and a sample size of 100, so that $1.96\sigma/\sqrt{n} \approx 1$, the rejection region would become $\bar{X} < 1$ or $\bar{X} > 1$, i.e., $|\bar{X}| > 1$. We call the value 1 the "critical value" for the test statistic.

Example

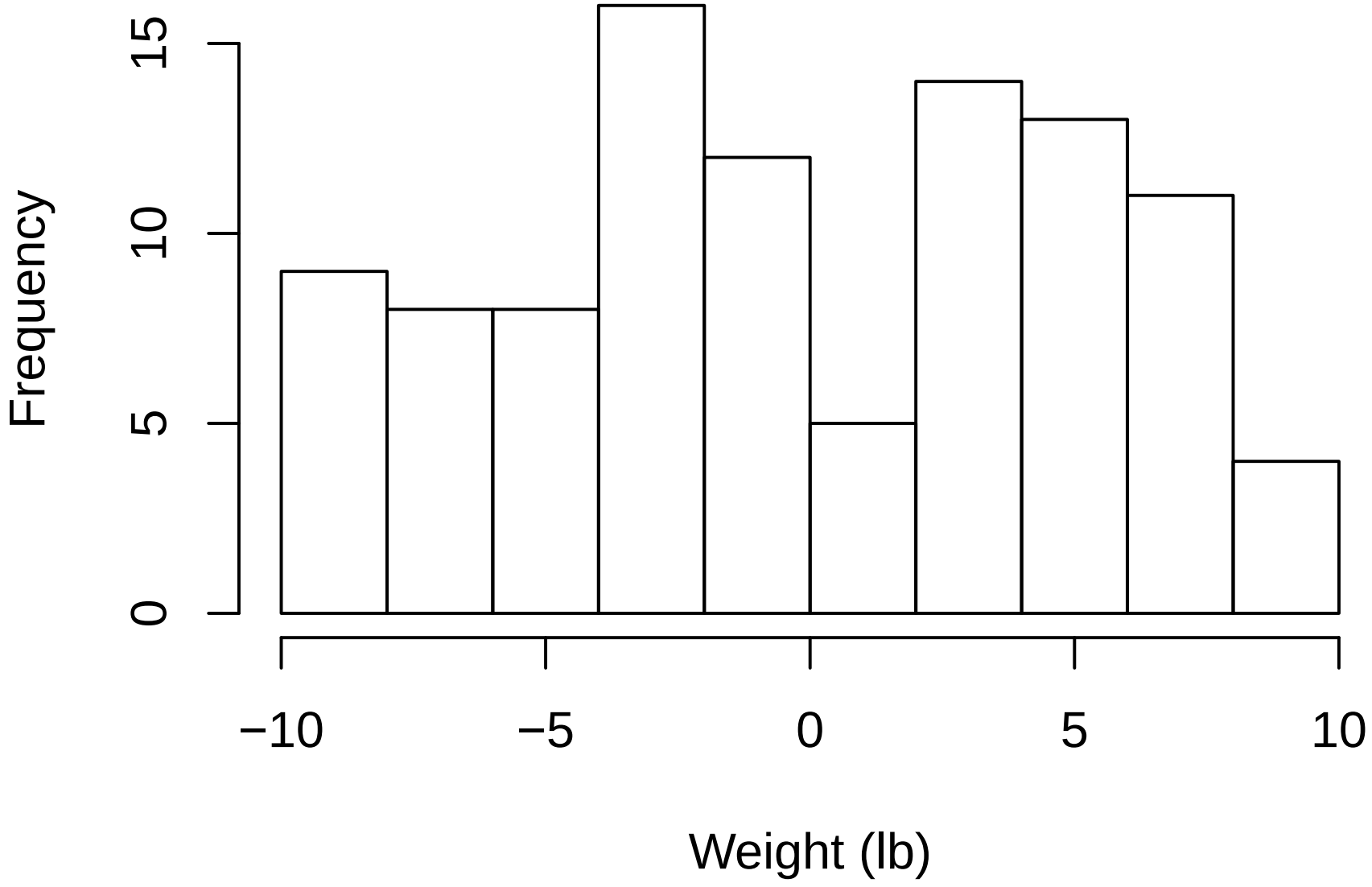
Suppose that with a sample size of $n = 100$ we get the following observations of weight change (recorded to the nearest pound):

```
  [1]  -5  -3   1   8  -6   8   9   3   3  -9  -6  -6   4  -2   5   0
 [20]   6   9  -6   3  -7  -5  -2 -10  -2   7  -3   0   2   0  -6   7
 [39]   4  -2   6   3   6   1   1   6 -10   0   5   4   0   7  -1  -5
 [58]   0   3  -2   8  -4  -1  -3   3  -5   0   5  -8   8  -3   7  -3
 [77]   7  -2   6   9  -1   4  -2  -3   5  -6   4  -8  -5  -7  -5  -9
 [96]   6  -1  -2   6   2
```

Then $\bar{x} = 0.39$ and so we would not reject the null hypothesis.

Illustration of the Data

Remember - always look at the data! Does the distribution accurately reflect the conclusion of the test? Are there any outliers that might have undue influence on the results?



Simulation Study of Type I Error

How do we know if our test really has the significance level, (in this example $\alpha = 0.05$)?

If the sample size is large enough (and our assumed variance is correct), we can be confident because of the CLT. But what if we have $n = 20$? Do a simulation study to check it out.

Because i don't know what the distribution of weight change is i will try a few different types of distribution to see what i get. Here are two examples;

1. A uniform distribution on the interval (-10, 10) with values rounded to the nearest pound.
2. A distribution that takes on only two values: -5 with probability $1/2$ and 5 with probability $1/2$.

Note: both of these distributions have $\sigma = 5$ as we are assuming.

Simulation under the Uniform Distribution with $n = 100$

```
set.seed(123456)
n=100
critical.value=1
reps=200
xbar=rep(NA,reps)
for(i in 1:reps) xbar[i]=mean(round(runif(n,-10,10),0))
summary(xbar)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.2500 -0.3325 -0.0300  0.0279  0.3900  1.5700
```

```
mean(abs(xbar)>critical.value)
```

```
[1] 0.06
```

This is what we would expect (a value close to 0.05).

Simulation under the Uniform Distribution with $n = 20$

For $n = 20$ the critical value is $1.96 \times 5/\sqrt{20} = 2.2$ so the rejection region is $|X| > 2.2$.

```
set.seed(123456)
n=20
critical.value=2.2
reps=200
xbar=rep(NA,reps)
for(i in 1:reps) xbar[i]=mean(round(runif(n,-10,10),0))
summary(xbar)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.8500 -1.0120 -0.1250 -0.0625  0.7500  3.9500
```

```
mean(abs(xbar)>critical.value)
```

```
[1] 0.08
```

This is not quite as close to 0.05 because the sample size is not large enough to give a very good approximation for the assumed distribution.

Simulation under the Dichotomous Distribution with $n = 100$

```
set.seed(123456)
n=100
critical.value=1
reps=200
xbar=rep(NA,reps)
for(i in 1:reps) xbar[i]=mean(sample(c(-5,5),size=n,replace=T))
summary(xbar)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.300  -0.300   0.000   0.019   0.300   1.400
```

```
mean(abs(xbar)>critical.value)
```

```
[1] 0.04
```

Again, a good match to theory, because of the large sample size.

Simulation under the Dichotomous Distribution with $n = 20$

```
set.seed(123456)
n=20
critical.value=2.2
reps=200
xbar=rep(NA,reps)
for(i in 1:reps) xbar[i]=mean(sample(c(-5,5),size=n,replace=T))
summary(xbar)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -2.50   -0.50    0.00    0.01    1.00    3.00
```

```
mean(abs(xbar)>critical.value)
```

```
[1] 0.02
```

This time the type I error probability is too low. As before, the problem is that the sample size is too small for the assumed distribution. This can make the type I error probability either too large or too small.

Calculation of Power

In a later class we will discuss how to do power calculations. For now, we will use simulation to estimate power. Consider the weight loss experiment. As with simulations of the type I error probability we need to consider various options for the alternative hypothesis. For example, suppose we want to know the power of our experiment to detect an average weight gain of 1 pound. We could do this by simulating from a uniform distribution on the interval (-9, 11).

```
set.seed(123456)
n=100
critical.value=1
reps=200
xbar=rep(NA,reps)
for(i in 1:reps) xbar[i]=mean(runif(n,-9,11))
summary(xbar)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.2763  0.6292  0.9901  1.0280  1.3980  2.5530
```

```
mean(abs(xbar)>critical.value)
```

```
[1] 0.49
```

The power is only 49% which would not be considered adequate. By convention, at least 80% power is considered acceptable in most applications; however, higher values such as 90% or even higher are required in some experimental settings.

To achieve a desired level of power we typically need to increase the sample size. This can be done using simulation studies as above. However, there are some approximate formulas that are typically used as short cuts. We will consider these when we cover the topic of power calculations in a later class.

Summary: 1-Sample Test for a Population Mean with Known Variance

Based on a random sample of size n we wish to test the hypothesis that the mean is equal to 0 against the alternative hypothesis that the mean is not 0, i.e.,

$$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0.$$

Assuming that the population variance is known to be equal to σ, we use the following rejection rule:

$$\text{Reject } H_0 \text{ if } |Z| = \frac{|\bar{X}|}{\sigma/\sqrt{n}} > 1.96.$$

This test is valid **if the sample size is sufficiently large**.

The large sample size ensures that the sampling distribution of $\bar{X}$ (and hence of Z) will be approximately normal, regardless of the distribution of the variable in the population.

4. Hypothesis Testing for a Population Mean with Unknown Variance

- Using the sample variance when the population variance is unknown
- The T-Test

Using the sample variance when the population variance is unknown

In practice, we typically do not know σ, so we substitute the sample standard deviation in place of σ to define the test statistic as follows:

$$Z = \frac{|\bar{X}|}{s/\sqrt{n}}$$

Note: the test statistic is sometimes denoted by T rather than Z to indicate that we are using the sample SD in place of the population SD. However, we will continue to use Z for now.

The rejection rule is

$$\text{Reject } H_0 : \mu = 0 \text{ if } Z = \frac{|\bar{X}|}{s/\sqrt{n}} > C,$$

where C is called the *critical value* for the test.

Choosing the critical value

How do we choose C? There are 2 cases to consider:

1. If n is quite large we can continue to use the value from the normal distribution (1.96 for a significance level of 0.05). This is because the large sample not only makes the sampling distribution of $\bar{X}$ approximately normal, but it also ensures that the sample SD will be a good approximation to the population SD (so we can ignore the fact that we use s in place of σ).
2. If n is not that large, the use of s in place of σ introduces additional variability into the test statistic; in this case, we cannot use the values from the normal distribution.

Example

Consider the data set of simulated weight changes. The mean, standard deviation, and test statistic are as follows:

```
m=mean(x)
m
```

```
[1] 0.39
```

```
s=sd(x)
s
```

```
[1] 5.372761
```

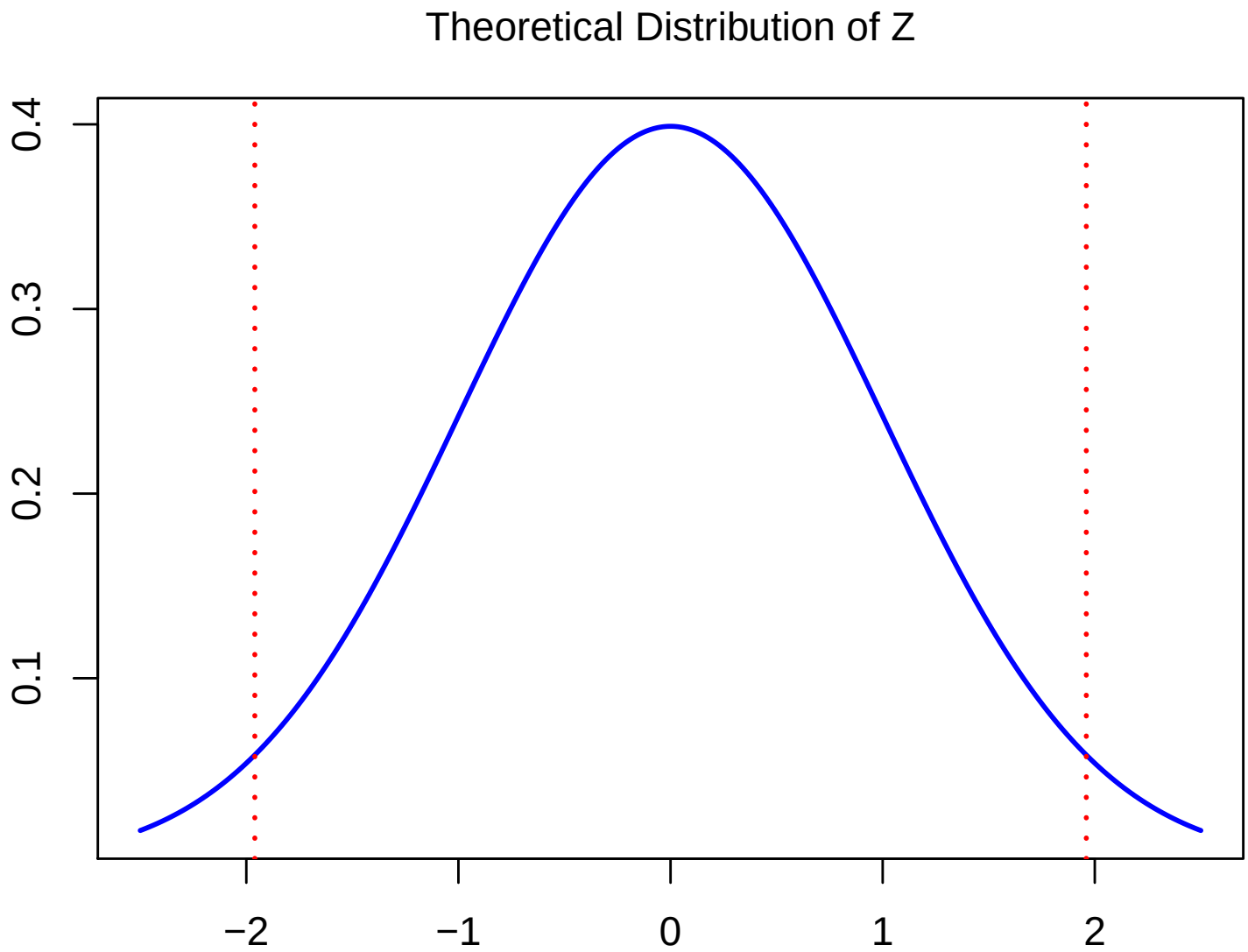
```
z=m/(s/sqrt(n))
z
```

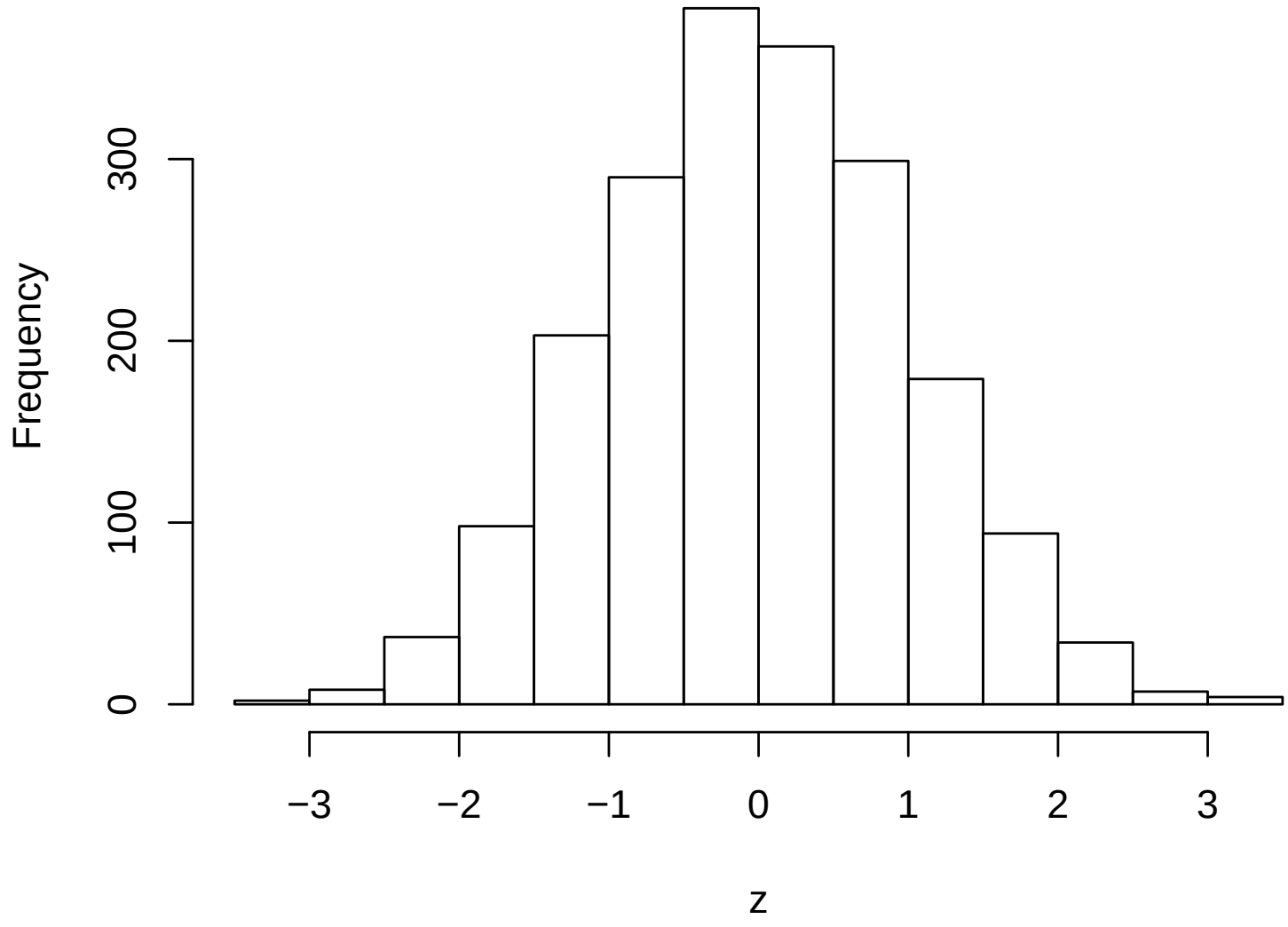
```
[1] 0.7258837
```

Note that the SD is quite close to the value 5 that we had assumed previously. Assuming that we are in case (1), i.e., that n is large enough, then we would not reject the null hypothesis because $|Z|$ is not larger than 1.96. The same conclusion was reached previously when we had assumed $\sigma = 5$.

The Distribution of Z Under the Null Hypothesis with Large n

With large n, the distribution of Z is approximately equal to the standard normal distribution. The tail probability beyond ± 1.96 is equal to 0.05.




Simulated Distribution of Z with Large n
Frequency
0
100
200
300
−3
−2
−1
0
1
2
3
z

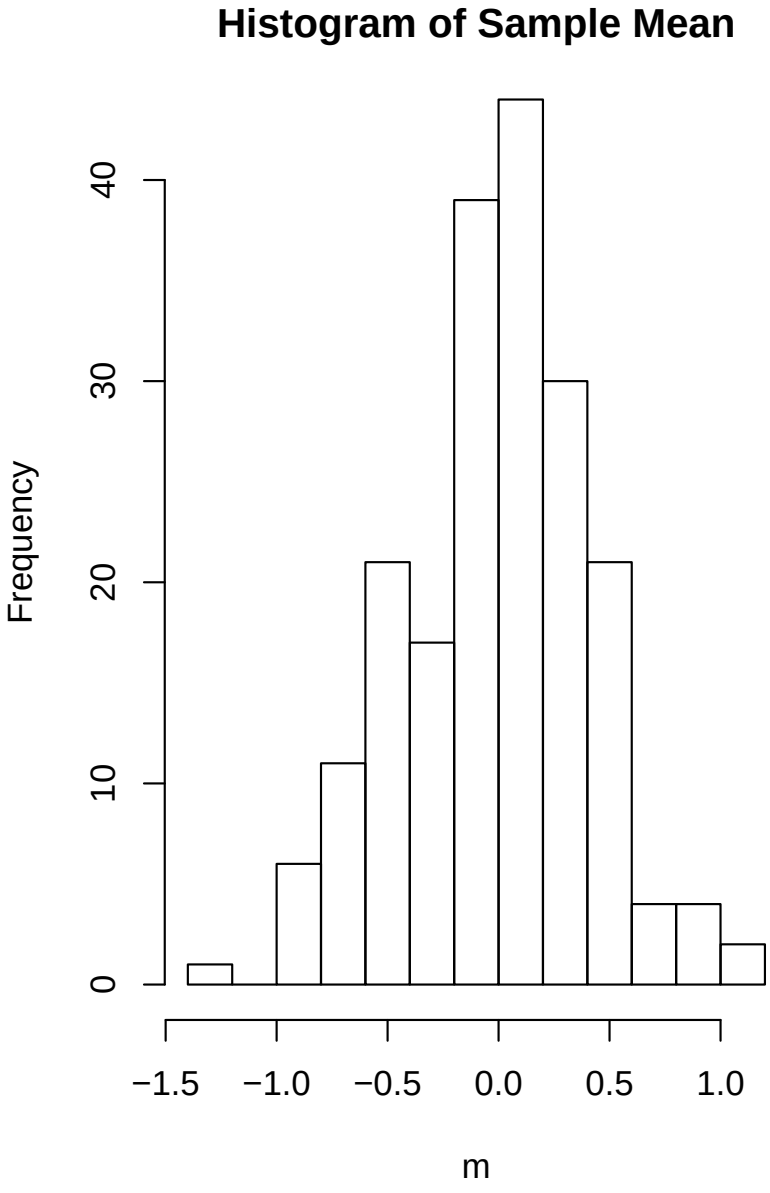
What happens if the sample size is not large enough?

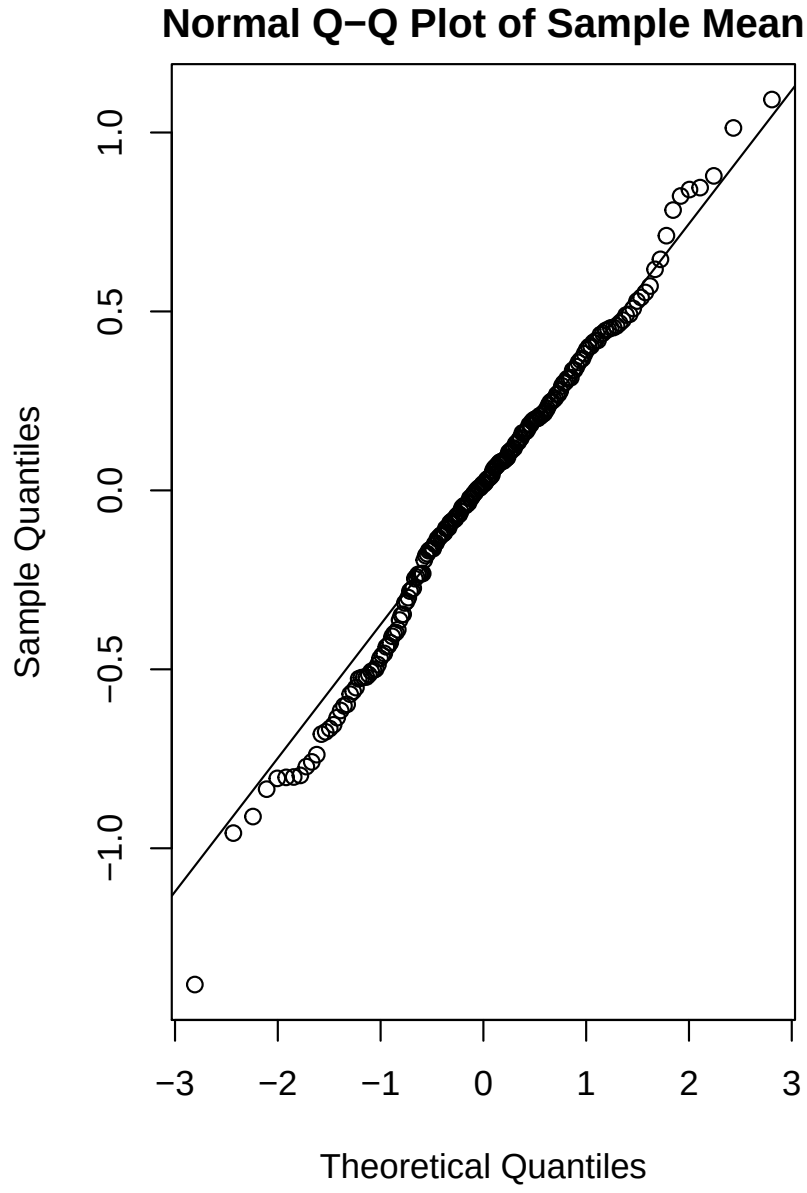
1. If the population distribution is not normal and n is not large enough for the Central Limit Theorem to take effect, then the sampling distribution of the sample mean (and hence that of Z) will not be approximately normal.
2. If the population distribution is not normal, then n might be large enough for the Central Limit Theorem to take effect, *but* not large enough for s to be a good approximation of σ.
3. *Even if the population distribution is normal*, if n is not large enough, the distribution of Z might not be normal because s is not close enough to σ.

Thus, the requirement on the sample size is a little more stringent than before.

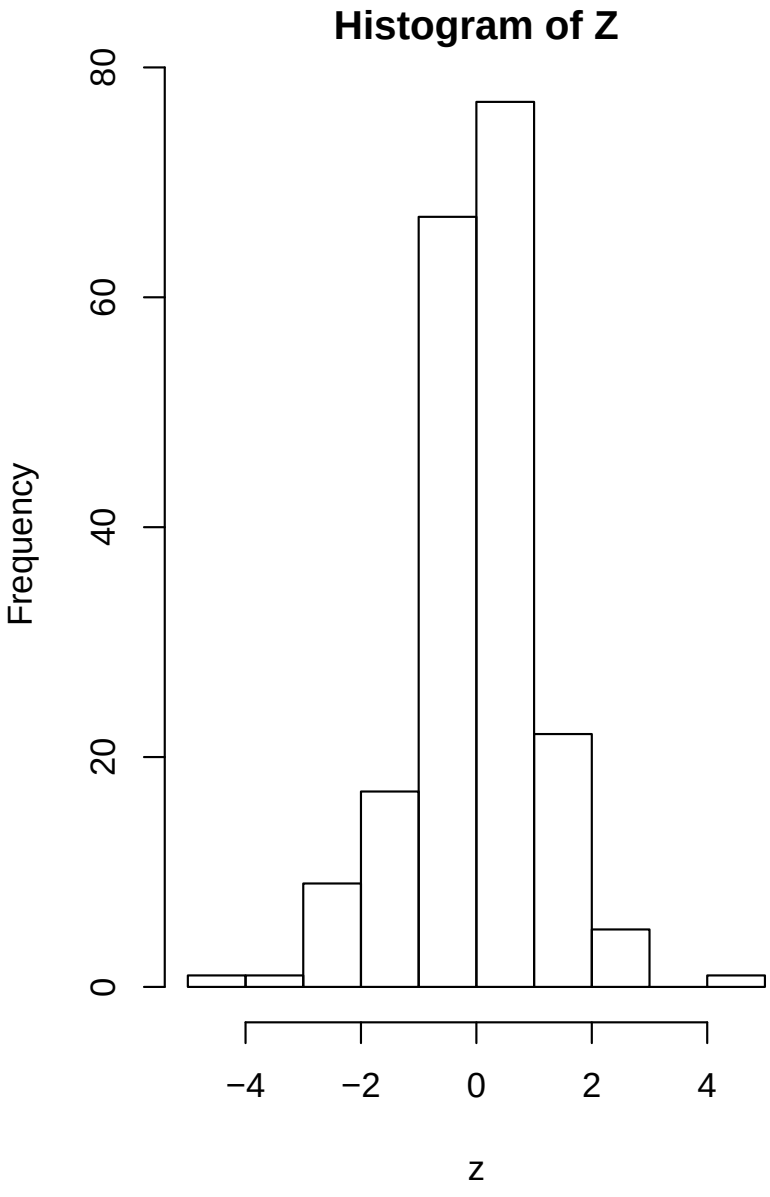
Sampling from a Normal Population with $n = 5$

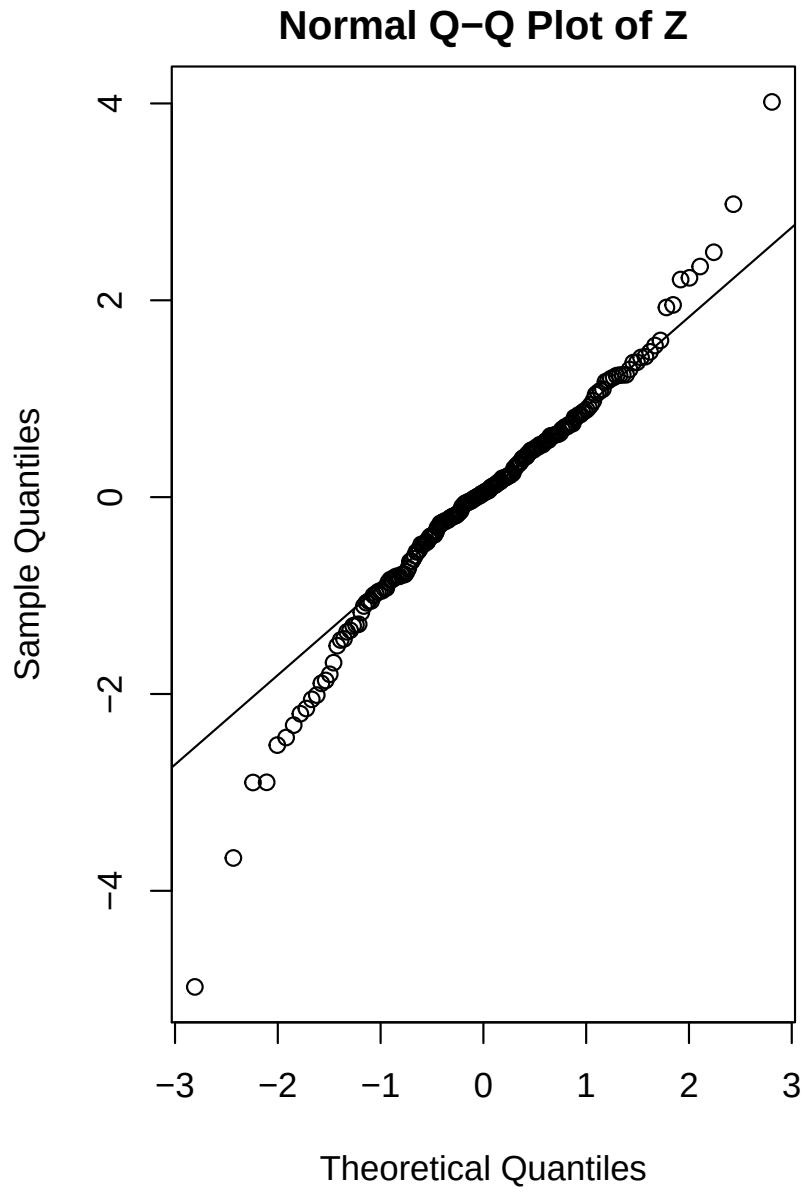
The distribution of the **sample mean** $\bar{X}$ is approximately normal.





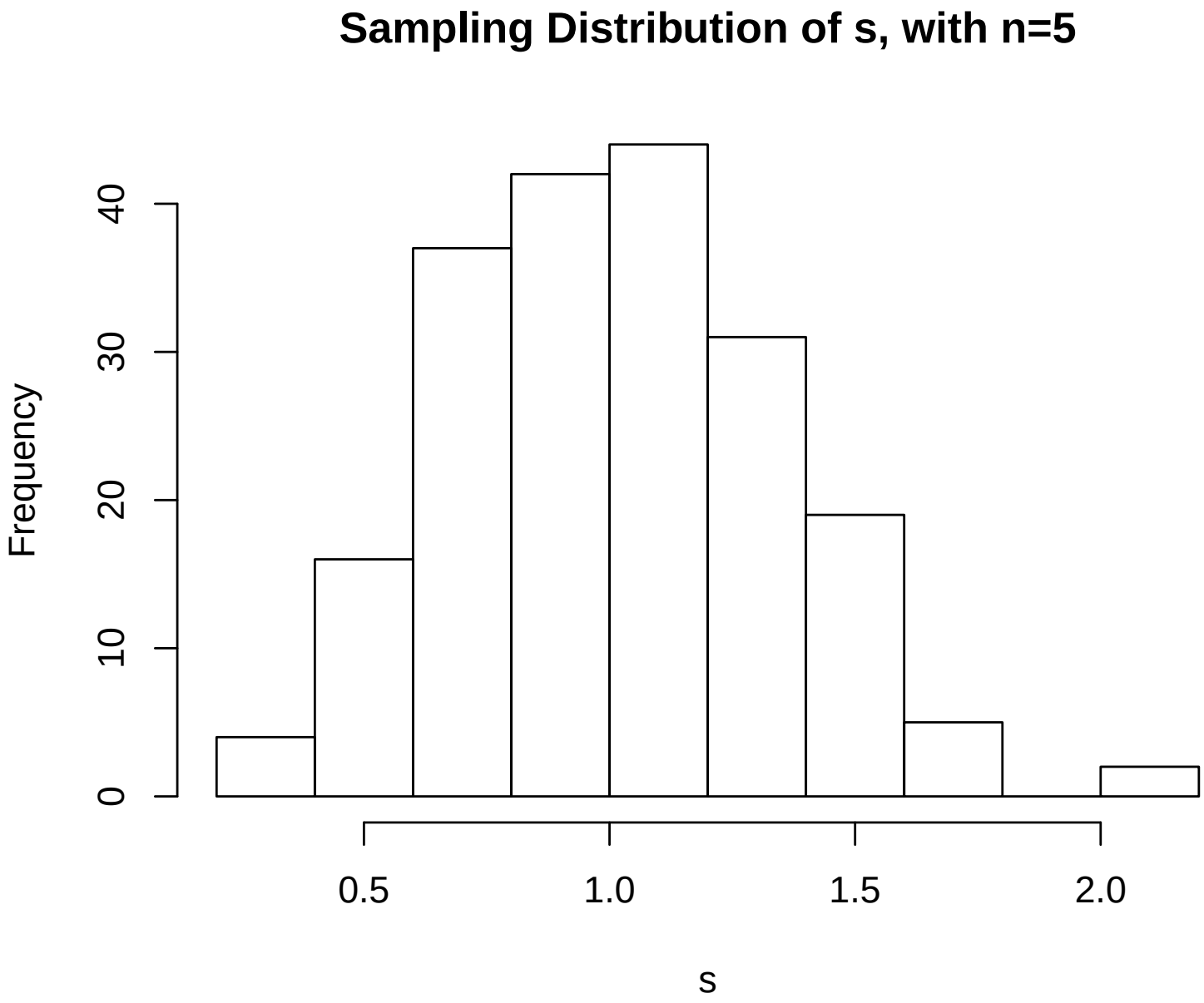
$\cdots\cdots$ **but** the distribution of Z is **not** approximately normal:





The Problem: s is not a good estimate of σ with $n = 5$

The true value is $\sigma = 1$. There is too much variability in s.



The T-Test

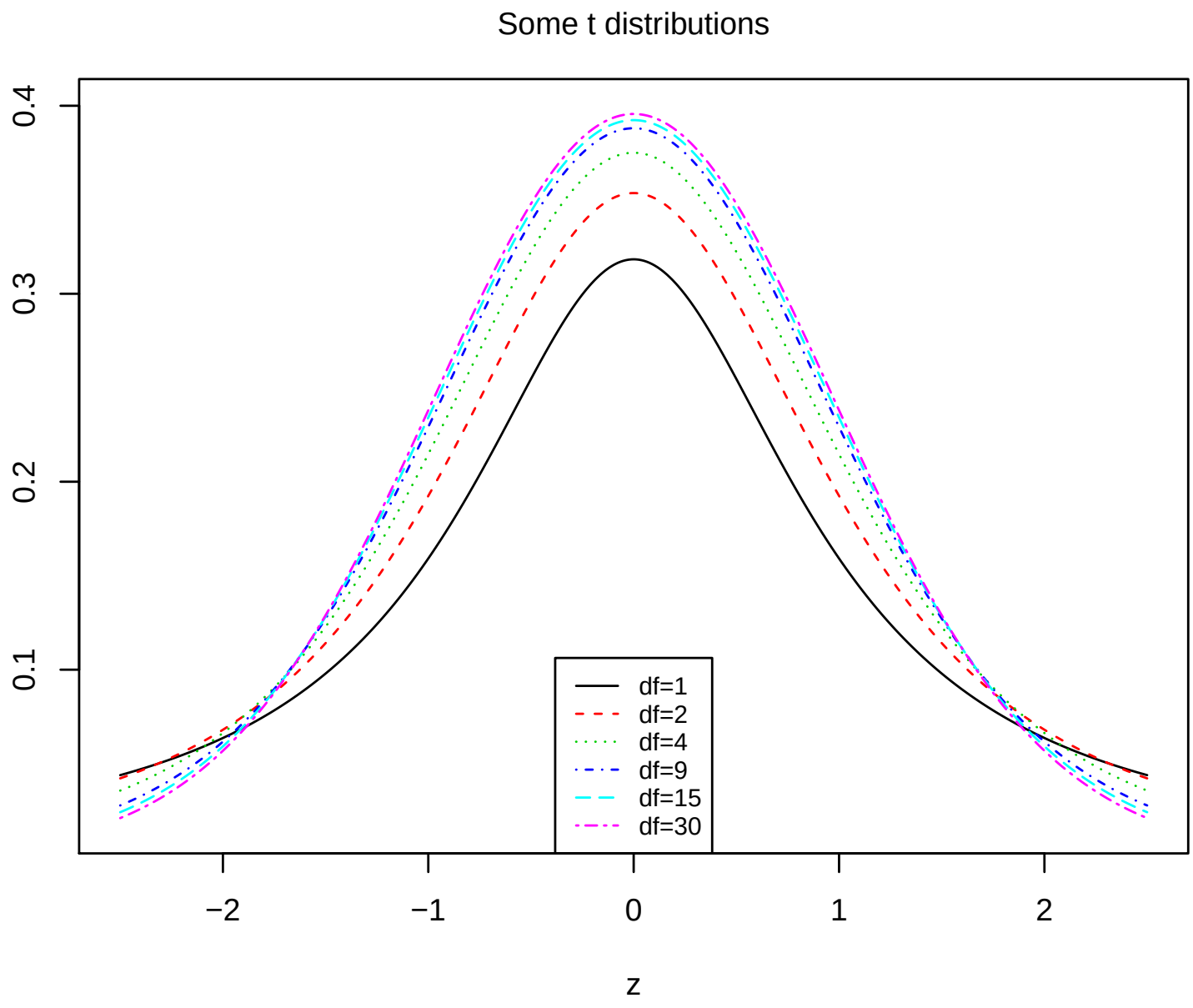
With small n, we take advantage of result from probability theory:

For *sampling from a normal population*, under the null hypothesis $H_0 : \mu = 0$, the sampling distribution of the test statistic Z is a t-distribution with $n - 1$ degrees of freedom, which is denoted t_{n-1}.

Therefore, in this case, we should use a critical value chosen based on the t_{n-1} distribution instead of the value from the normal distribution. For a test with significance level 0.05, for large n we would use the critical value 1.96. But for small n we would use a corresponding value based on the t_{n-1} distribution.

For small n (i.e., less than about 20), there can be an important difference between the normal value and the t value. But if n is large, the t_{n-1} distribution is almost identical to the standard normal distribution.

t-distributions have heavier tails than the normal



Critical Values for the t-test ($\alpha = 0.05$)

```
n=c(2:5,10,20,30,50,100,200)
data.frame(n,df=n-1,t.critical.value=round(qt(.975,df=n-1),2))
```

```
     n  df t.critical.value
1    2   1              12.71
2    3   2               4.30
3    4   3               3.18
4    5   4               2.78
5   10   9               2.26
6   20  19               2.09
7   30  29               2.05
8   50  49               2.01
9  100  99               1.98
10 200 199               1.97
```

As $n \to \infty$ the critical value approaches the normal distribution value of 1.96. In practice, the distinction between the normal and t values can be ignored for n of approximately 50 or greater.

Critical Values for the t-test for other significance levels

For significance level $\alpha = 0.05$ the t critical value is denoted $t_{n-1,0.05}$ ("t.05" in table below). For smaller α values the t-distribution approaches the normal a little more slowly (requires larger n to use the normal value).

	n	df	t.05	t.01	t.001
1	10	9	2.26	3.25	4.78
2	20	19	2.09	2.86	3.88
3	30	29	2.05	2.76	3.66
4	50	49	2.01	2.68	3.50
5	100	99	1.98	2.63	3.39
6	200	199	1.97	2.60	3.34

The normal critical values are 1.96, 2.58, and 3.29, for $\alpha = 0.05, 0.01, 0.001$, respectively.

Simulation under the Uniform Distribution with $n = 100$

```
set.seed(123456)
n=100
reps=200
z=rep(NA,reps)
for(i in 1:reps){
  x=round(runif(n,-10,10),0)
  z[i]=mean(x)/(sd(x)/sqrt(n))
}
summary(z)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.19800 -0.61970 -0.05312  0.04987  0.69740  2.88400
```

```
critical.value = 1.96
mean(abs(z)>critical.value)
```

```
[1] 0.045
```

Good match to theory.

Simulation under the Uniform Distribution with $n = 20$

```
set.seed(123456)
n=20
reps=200
z=rep(NA,reps)
for(i in 1:reps){
  x=round(runif(n,-10,10),0)
  z[i]=mean(x)/(sd(x)/sqrt(n))
}
summary(z)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.43800 -0.71080 -0.09481 -0.03724  0.54950  2.96400
```

```
critical.value = 1.96
mean(abs(z)>critical.value)
```

```
[1] 0.065
```

A little off, but not too bad!

Simulation under the Dichotomous Distribution with $n = 20$

```
set.seed(123456)
n=20
reps=200
z=rep(NA,reps)
for(i in 1:reps){
  x=sample(c(-5,5),size=n,replace=T)
  z[i]=mean(x)/(sd(x)/sqrt(n))
}
summary(z)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-2.517000 -0.438100  0.000000  0.008888  0.889800  3.269000
```

```
critical.value = 1.96
mean(abs(z)>critical.value)
```

```
[1] 0.02
```

As before, the test does not perform so well under this distributional assumption with $n = 20$.

Simulation under a Normal Distribution with $n = 20$

```
set.seed(123456)
n=20
reps=200
z=rep(NA,reps)
for(i in 1:reps){
  x=rnorm(n,mean=0,sd=5)
  z[i]=mean(x)/(sd(x)/sqrt(n))
}
summary(z)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.87300 -0.76890 -0.06437  0.04882  0.84190  3.06100
```

```
critical.value = 1.96
mean(abs(z)>critical.value)
```

```
[1] 0.085
```

This is surprising! Since the population distribution is normal the sampling distribution of $\bar{X}$ has to be normal. So why does Z not have a normal distribution?

The problem is with the substitution of s for σ. With a sample size of only 20, s is not a very good estimator of σ. The extra variability introduced by using s in place of σ leads to a type I error probability that is too high. Using a critical value from the t-distribution solves this problem.

The P-Value for a 1-sample test for a mean

The *p*-value is defined in the same way and calculated in a similar way as for the test for a proportion. The *p*-value is the probability of getting a result as or more extreme than the one observed.

For the data set of simulated weight changes, we have a test statistic of $Z = 0.73$ (to 2 decimal places). Thus, the *p*-value is the 2-sided tail probability for the standard normal distribution.

```
2*(1-pnorm(0.73))
```

```
[1] 0.4653902
```

The interpretation is that we would have about a 47% chance of getting a result as or more extreme than this (meaning indicating as or greater departures from a mean of 0), if the null hypothesis is true.

Testing a Non-Zero Value for a Population Mean

Sometimes we are interested in a non-zero value as the null hypothesis value. If the hypothesize value of μ is μ_0, the hypotheses become

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

The rejection rule becomes

$$\text{Reject } H_0 \text{ if } |Z| = \frac{|\bar{X} - \mu_0|}{s/\sqrt{n}} > 1.96.$$

So, the only difference is that we subtract μ_0 from $\bar{X}$ in the test statistic.

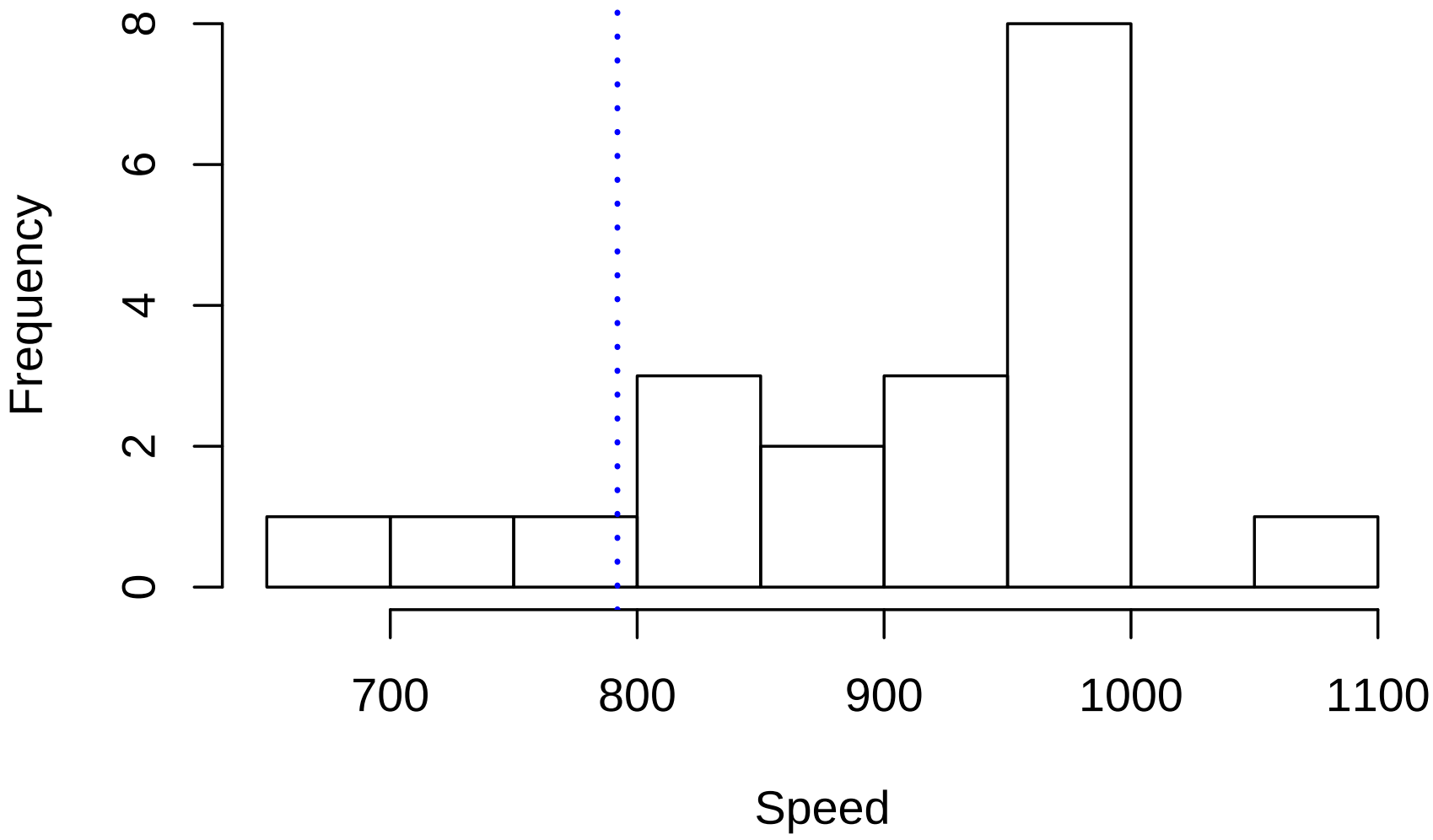
Example: Was the Morley experiment biased?

Let μ be the population mean of the Morley measurements of the speed of light (with 299,000 subtracted). If the experiment is unbiased then μ should exactly equal the true value, i.e., $\mu = 792$. Therefore our null hypothesis value for the mean is $\mu_0 = 792$ and we wish to test $H_0 : \mu = 792$ versus $H_1 : \mu \neq 792$.

Example: Was the Morley Experiment Biased?

We know the true value is 792 $(299,792 - 299,000)$.

```
Speed = morley$Speed[morley$Expt==1]
hist(Speed,main="")
abline(v=792,lty=3,col=4,lwd=2)
```



Testing the Null Hypothesis

We wish to test $H_0 : \mu = 792$ versus $H_1 : \mu \neq 792$, where μ is the population mean of the Morley measurements (i.e., the value we would get if we averaged an infinite number of measurements).

```
n=length(Speed)
z=(mean(Speed)-792)/(sd(Speed)/sqrt(n))
z
```

```
## [1] 4.98675
```

This is much larger than the critical value of the t_{19} distribution:

```
qt(0.975,df=n-1)
```

```
## [1] 2.093024
```

Thus, we would reject the null hypothesis that $\mu = 792$ at significance level 0.05, and conclude that we have evidence that the Morley Experiment 1 was biased. The p-value for the test is < 0.001.

```
2*(1-pt(z,df=n-1))
```

Summary: 1-Sample T-Test for a Population Mean with Unknown Variance

Based on a random sample of size n we wish to test the hypothesis that the mean is equal to 0 against the alternative hypothesis that the mean is not 0, i.e.,

$$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0.$$

Assuming that the population variance is not known we use the following rejection rule:

$$\text{Reject } H_0 \text{ if } |Z| = \frac{|\bar{X}|}{s/\sqrt{n}} > C,$$

where the critical value C is chosen based on the t_{n-1} distribution. This test is valid if **either of the following holds**:

1. The population distribution is normal.
2. sample size is sufficiently large.

Summary of Hypothesis Testing Terminology

Null Hypothesis (H_0): an hypothesis about the true state of nature.

Alternative Hypothesis (H_1): a set of alternatives to the null hypothesis

Rejection Rule: a rule for deciding to reject H_0 based on the observed value of a random variable X, written as $X \in R$ where R is the rejection region for X

Type I Error probability (α): the probability of rejecting H_0 when it is true, i.e., $P(X \in \mathrm{R}|H_0)$, also called the significance level

Power: the probability of rejecting H_0 when it is false, i.e., $P(X \in \mathrm{R}|H_1)$. This requires specifying a *specific* alternative H_1.

Type II Error probability (β): the probability of *not* rejecting H_0 when it is false, i.e., $\beta = 1 -$ Power

P-Value: the observed level of significance, i.e., the probability of the test statistic being as extreme or more extreme than the observed value, under the assumption that H_0 is true.

Summary of Interpretational Issues in Hypothesis Testing

1. Confidence intervals and hypothesis tests complement each other. We can think of the confidence interval as containing all the values that would *not* have been rejected by a corresponding hypothesis test. This relationship is exact in some situations but only approximate in other situations.
2. The p-value complements the decision rule of the hypothesis test. It quantifies the evidence in the data against the null hypothesis. A test rejects H_0 when the p-value is less than α.
3. p-values and hypothesis tests can be abused. In particular, if we perform too many hypothesis tests we are likely to get many type I errors.
4. We should never "accept" the null hypothesis – we can only "not reject".

Summary of 1-Sample Tests for a Population Mean

We have a random sample from a population with mean μ and unknown* variance σ^2. We want to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu 0$.

1. Calculate the test statistic: $Z = (\bar{X} - \mu_0)/(s/\sqrt{n})$
2. Calculate the critical value for the desired significance level (α).

If n is large, use the standard normal distribution, (e.g., 1.96) for $\alpha = 0.05$.

If n is small *and the population distribution is normal*, then use the t_{n-1} distribution.

If n is small and *we are not sure if the population distribution is normal* then we are stuck for now! (see Exact methods later in course)

*Note: If the variance is known then the only difference is that we calculate the test statistic as $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$