Null Hypothesis Significance Testing I
Class 17, 18.05, Spring 2014
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definitions of the significance testing terms: NHST, null hypothesis, alternative hypothesis, simple hypothesis, composite hypothesis, significance level, power.

2. Be able to design and run a significance test for Bernoulli or binomial data.

3. Be able to compute a p-value for a normal hypothesis and use it in a significance test.

2 Introduction

Frequentist statistics is often applied in the framework of null hypothesis significance testing (NHST). We will look at the *Neyman-Pearson* paradigm which focuses on one hypothesis called the *null hypothesis*. There are other paradigms for hypothesis testing, but Neyman-Pearson is the most common. Stated simply, this method asks if the data is well outside the region where we would expect to see it under the null hypothesis. If so, then we reject the null hypothesis in favor of a second hypothesis called the alternative hypothesis.

The computations done here all involve the likelihood function. There are two main differences between what we'll do here and what we did in Bayesian updating.

1. The evidence of the data will be considered purely through the likelihood function it will not be weighted by our prior beliefs.

2. We will need a notion of extreme data, e.g. 95 out of 100 heads in a coin toss or a Mayfly that lives for a month.

many similarities to the computations we did for confidence intervals. In fact, confidence intervals can be used in one type of significance testing.

2.1 Motivating examples

Example 1. Suppose you want to decide whether a coin is fair. If you toss it 100 times and get 85 heads, would you think the coin is likely to be unfair? What about 60 heads? Or 52 heads? Most people would guess that 85 heads is strong evidence that the coin is unfair, whereas 52 heads is no evidence at all. Sixty heads is less clear. NHST is a frequentist approach to thinking quantitatively about these questions.

Example 2. Suppose you want to compare a new medical treatment to a placebo or the current standard of care. What sort of evidence would convince you that the new treatment is better than the placebo or the current standard? Again NHST is a quantitative framework for answering these questions.

3 Significance testing

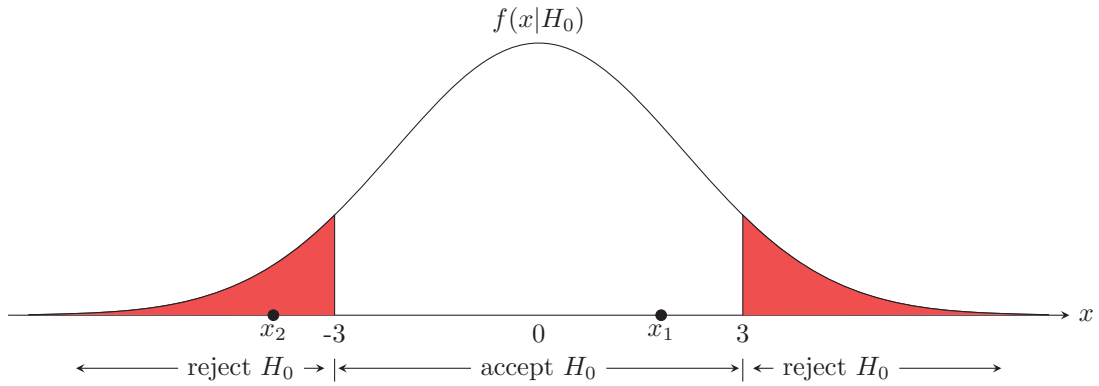
We'll start by listing the ingredients for NHST. Formally they are pretty simple. There is an art to choosing good ingredients. We will explore the art in examples.

3.1 Ingredients

- H_0: the *null hypothesis*. This is the default assumption for the model generating the data.
- H_A: the *alternative hypothesis*. If we reject the null hypothesis we accept this alternative as the best explanation for the data.
- X: the *test statistic*. We compute this from the data.
- *Null distribution*: the probability distribution of X assuming H_0.
- *Rejection region*: if X is in the rejection region we reject H_0 in favor of H_A.
- *Acceptance region*: the complement to the rejection region. If X is in this region we do not reject H_0. Note that we say 'do not reject' rather than 'accept' because usually the best we can say is that the data does not support rejecting H_0. We'll still use the term 'acceptance region', because it is simpler than 'non-rejection region'.

The null hypothesis H_0 and the alternative hypothesis H_A play different roles. Typically we choose H_0 to be either a simple hypothesis or the default which we'll only reject if we have enough evidence against it. The examples below will clarify this.

Example 3. The diagram below illustrates a null distribution with acceptance and rejection regions.



The test statistic x_1 is in the acceptance (technically, non-rejection) region. So, if our data produces the test statistic x_1 then we will not reject the null hypothesis H_0. On the other hand the test statistic x_2 is in the rejection region, so if our data produces this test statistic we will reject the null hypothesis in favor of the alternative hypothesis.

There are several things to note in this picture.
1. The rejection region consists of values far from the center of the null distribution.

2. The rejection region is two-sided. We will also see examples of one-sided rejection regions as well.
3. The alternative hypothesis is not mentioned. We accept or reject H_0 based only on $f(x|H_0)$, the likelihood of the test statistic conditioned on H_0. As we will see, the alternative hypothesis H_A should be considered when choosing a rejection region, but formally it only comes in when deciding how much weight to give the conclusion of the test.

4 NHST Terminology

In this section we will use one extended example to introduce and explore the terminology used in NHST.

Example 4. To test whether a coin is fair we flip it 10 times. If we get an unexpectedly large or small number of heads we'll suspect the coin is unfair. To make this precise in the language of NHST we set up the ingredients as follows. Let θ be the probability that the coin lands heads when flipped.

1. Null hypothesis: H_0 = 'the coin is fair', i.e. $\theta = .5$.
2. Alternative hypothesis: H_A = 'the coin is not fair', i.e. $\theta \neq .5$
3. Test statistic: X = number of heads in 10 flips
4. Null distribution: This is the probability function based on the null hypothesis
$$p(x \,|\, \theta = .5) \sim \text{binomial}(10, .5).$$

Here is the probability table for the null distribution.

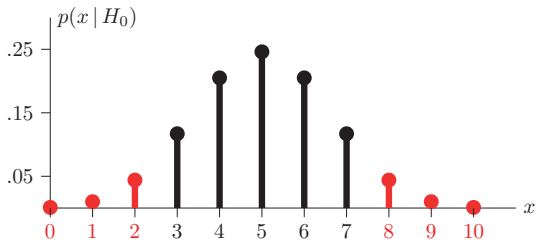
x	0	1	2	3	4	5	6	7	8	9	10
$p(x \,\|\, H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

5. Rejection region: under the null hypothesis we expect to get about 5 heads in 10 tosses. We'll reject H_0 if the number of heads is much fewer or greater than 5. Let's set the rejection region as $\{0, 1, 2, 8, 9, 10\}$. That is, if the number of heads in 10 tosses is in this region we will reject the hypothesis that the coin is fair in favor of the hypothesis that it is not.

We can summarize all this in the graph and probability table below. Both we show the null distribution in a probability table. The rejection region consists of those values of x in red. The probabilities corresponding to it are shaded in red. We also show the null distribution as a stem plot with the rejection values of x in red.

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

Rejection region and null probabilities as a table for example 4.



Rejection region and null probabilities as a stemp plot for example 4.

Notes for example 4:
1. The null hypothesis is the cautious default: we won't claim the coin is unfair unless we have good evidence.
2. The rejection region consists of data that is extreme under the null hypothesis. That is, it consists of the outcomes that are in the tail of the null distribution away from the high probability center. As we'll discuss soon, how far away depends on the significance level α of the test.
3. If we get 3 heads in 10 tosses, then the test statistic is in the acceptance region. The usual scientific language would be to say that the data 'does not support rejecting the null hypothesis'. Even if we got 5 heads, we would *not* claim that the data proves the null hypothesis is true.

Question: If we have a fair coin what is the probability that we will decide incorrectly it is unfair?

answer: The null hypothesis is that the coin is fair. The probability that for such a coin the data will land in the rejection region is the sum of the probabilities in red. That is, $P(\text{rejecting } H_0 \,|\, H_0 \text{ is true}) = .11$

Below we will continue with Example 4, define more terms used in NHST and see how to quantify properties of the significance test.

4.1 Simple and composite hypotheses

Definition: simple hypothesis: A *simple hypothesis* is one for which we can specify its distribution completely. A typical simple hypothesis is that a parameter of interest takes a specific value.

Definition: composite hypotheses: If its distribution cannot be fully specified, we say that the hypothesis is *composite*. A typical composite hypothesis is that a parameter of interest lies in a range of values.

In Example 4 the null hypothesis is that $\theta = .5$, so the null distribution is binomial$(10, .5)$. Since the null distribution is fully specified, H_0 is simple. The alternative hypothesis is that $\theta = .5$. This is really many hypotheses in one: θ could be .51, .7, .99, etc. Since the alternative distribution binomial$(10, \theta)$ is not fully specified, H_A is composite.

Example 5. Suppose we have data $x_1, \ldots, x_n$. Suppose also that our hypotheses are
H_0: the data is drawn from $N(0, 1)$
H_A: the data is drawn from $N(1, 1)$.
These are both *simple hypotheses* – each hypothesis completely specifies a distribution.

Example 6. (**Composite hypotheses.**) Now suppose that our hypotheses are
H_0: the data is drawn from a Poisson distribution of unknown parameter.
H_A: the data is not drawn from a Poisson distribution.
These are both composite hypotheses, as they don't fully specify the distribution.

Example 7. In an ESP experiment a subject is asked to identify the suits of 100 cards drawn (with replacement) from a deck of cards. Let T be the number of successes. The (simple) null hypothesis that the subject does not have ESP is given by

$$H_0\text{: } T \sim \text{binomial}(100, .25)$$

The (composite) alternative hypothesis that the subject has ESP is given by

$$H_A\text{: } T \sim \text{binomial}(100, p) \text{ with } p > .25$$

Another (composite) alternative hypothesis that something besides pure chance is going on is given by

$$H_A\text{: } T \sim \text{binomial}(100, p), \text{ with } p = .25$$

Values of $p < .25$ represent hypotheses that the subject has a kind of anti-esp.

4.2 Types of error

There are two types of errors we can make. We can incorrectly reject the null hypothesis when it is true or we can incorrectly fail to reject it when it is false. These are unimaginatively labeled *type I* and *type II errors*. We summarize this in the following table.

		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Accept' H_0	correct decision	Type II error

Type I: false rejection of H_0
Type II: false 'acceptance' of H_0

4.3 Significance level and power

Significance level and power are used to quantify the quality of the significance test. Ideally a significance test would not make errors. That is, it would not reject H_0 when H_0 was true and would reject H_0 in favor of H_A when H_A was true. Altogether there are 4 important probabilities corresponding to the 2×2 table just above.

$$\begin{array}{ll} P(\text{reject } H_0|H_0) & P(\text{reject } H_0|H_A) \\ P(\text{do not reject } H_0|H_0) & P(\text{do not reject } H_0|H_A) \end{array}$$

The two probabilities we focus on are:

$$\begin{aligned} \textit{Significance level} &= P(\text{reject } H_0|H_0) \\ &= \text{probability we incorrectly reject } H_0 \\ &= P(\text{type I error}). \end{aligned}$$

$$\begin{aligned} \textit{Power} &= \text{probability we correctly reject } H_0 \\ &= P(\text{reject } H_0|H_A) \\ &= 1 - P(\text{type II error}). \end{aligned}$$

Ideally, a hypothesis test should have a small significance level (near 0) and a large power (near 1). Here are two analogies to help you remember the meanings of significance and power.

Some analogies
1. Think of H_0 as the hypothesis 'nothing noteworthy is going on', i.e. 'the coin is fair', 'the treatment is no better than placebo' etc. And think of H_A as the opposite: 'something interesting is happening'. Then power is the probability of detecting something interesting

when it's present and significance level is the probability of mistakenly claiming something interesting has occured.

2. In the U.S. criminal defendents are presumed innocent until proven guilty beyond a reasonable doubt. We can phrase this in NHST terms as

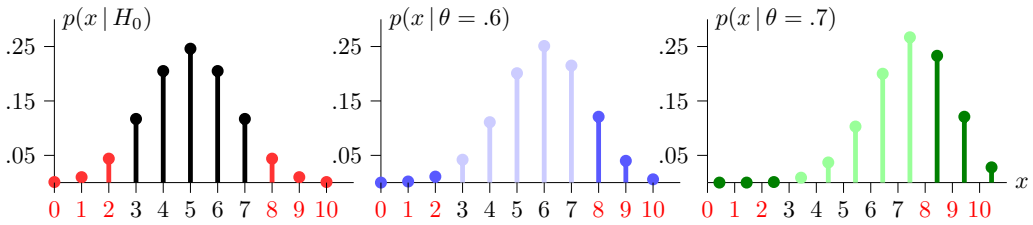
H_0: the defendent is innocent (the default)
H_A: the defendent is guilty.

Significance level is the probability of finding and innocent person guilty. Power is the probability of correctly finding a guilty party guilty. 'Beyond a reasonable doubt' means we should demand the significance level be very small.

Composite hypotheses

H_A is composite in Example 4, so the power is different for different values of θ. We expand the previous probability table to include some alternate values of θ. We do the same with the stem plots.

x	0	1	2	3	4	5	6	7	8	9	10
$H_0: p(x\|\theta = .5)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001
$H_A: p(x\|\theta = .6)$	.000	.002	.011	.042	.111	.201	.251	.215	.121	.040	.006
$H_A: p(x\|\theta = .7)$	.000	.0001	.001	.009	.037	.103	.200	.267	.233	.121	.028



Rejection region and null and alternative probabilities for example 4

We use the probability table to compute the significance level and power of this test.

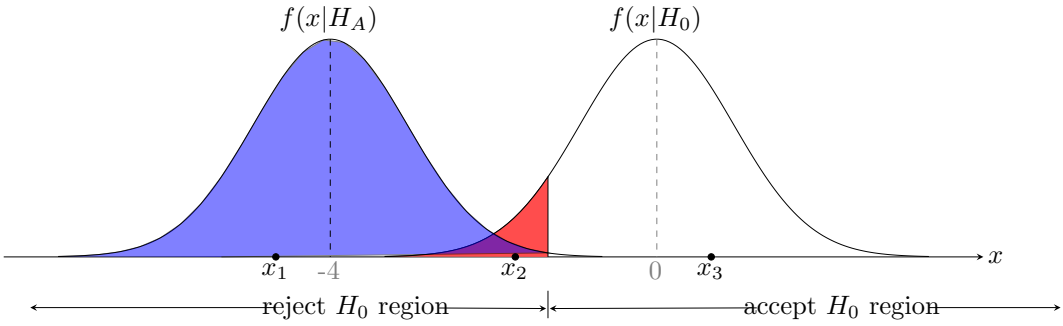
Significance level = probability we reject H_0 when it is true
= probability the test statistic is in the rejection region when H_0 is true
= probability in the rejection region in the H_0 row of the table
= sum of red boxes in the $\theta = .5$ row
= .11

Power when $\theta = .6$ = probability we reject H_0 when $\theta = .6$
= probability the test statistic is in the rejection region when $\theta = .6$
= probability in the rejection region in the $\theta = .6$ row of the table
= sum of dark blue boxes in the $\theta = .6$ row
= .180

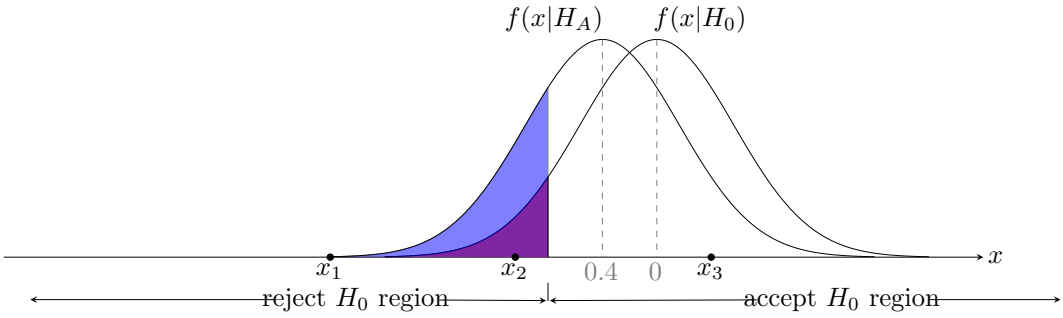
Power when $\theta = .7$ = probability we reject H_0 when $\theta = .7$
= probability the test statistic is in the rejection region when $\theta = .7$
= shaded probability in the $\theta = .7$ row of the table
= sum of dark green boxes in the $\theta = .7$ row
= .384

We see that the power is greater for $\theta = .7$ than for $\theta = .6$. This isn't surprising since we expect it to be easier to recognize that a .7 coin is unfair than is is to recognize .6 coin is unfair. Typically, we get higher power when the alternate hypothesis is farther from the null hypothesis. In Example 4, it would be quite hard to distinguish a fair coin from one with $\theta = .51$.

We illustrate this with the following two figures. The shaded area under $f(x|H_0)$ represents the significance level, i.e., the probability that the test statistic falls in the rejection region even though H_0 is true. Likewise, the shaded area under $f(x|H_A)$ represents the power, i.e. the probability that the test statistic is in the rejection (of H_0) region when H_A is true. Both tests have the same significance level, but if $f(x|H_A)$ has considerable overlap with $f(x|H_0)$ the power is much lower. It is well worth your while to thoroughly understand these graphical representations of significance testing.



High power test



Low power test

In both tests both distributions are standard normal. The null distribution, rejection region and significance level are all the same. (The significance level is red/purple area under $f(x \mid H_0$ and above the rejection region.) In the top figure we see the means of the two distributions are 4 standard deviations apart. Since the areas under the densities have very little overlap the test has high power. That is if the data x is drawn from H_A it will almost certainly be in the rejection region. For example x_3 would be a very surprising outcome for the H_A distribution.

In the bottom figure we see the means of the two distributions are just 0.4 standard deviations apart. Since the areas under the densities have a lot of overlap the test has low power. That is if the data x is drawn from H_A it is highly likely to be in the acceptance region. For example x_3 would be not be a very surprising outcome for the H_A distribution.

Typically we can increase the power of a test by increasing the amount of data and thereby

decreasing the variance of the null and alternative distributions. In experimental design it is important to determine ahead of time the number of trials or subjects needed to achieve a desired power.

Example 8. Suppose a drug for a disease is being compared to a placebo. We choose our null and alternative hypotheses as

H_0 = the drug does not work better than the placebo

H_A = the drug works better than the placebo

The power of the hypothesis test is the probability that the test will conclude that the drug is better, if it is indeed truly better. The significance level is the probability that the test will conclude that the drug works better, when in fact it does not.

5 Designing a hypothesis test

Formally all a hypothesis test requires is H_0, H_A, a test statistic and a rejection region. In practice the design is often done using the following steps.

1. **Pick the null hypothesis H_0.**
The choice of H_0 and H_A is not mathematics. It's art and custom. We often choose H_0 to be simple. Or we often choose H_0 to be the simplest or most cautious explanation, i.e. no effect of drug, no ESP, no bias in the coin.

2. **Decide if H_A is one-sided or two-sided.**
In the example 4 we wanted to know if the coin was unfair. An unfair coin could be biased for or against heads, so $H_A : \theta \neq .5$ is a two-sided hypothesis. If we only care whether or not the coin is biased for heads we could use the one-sided hypothesis $H_A : \theta > .5$.

3. **Pick a test statistic.**
For example, the sample mean, sample total, or sample variance. Often the choice is obvious. Some standard statistics that we will encounter are z, t, and χ^2. We will learn to use these statistics as we work examples over the next few classes. One thing we will say repeatedly is that the distributions that go with these statistics are always conditioned on the null hypothesis. That is, we will compute likelihoods such as $f(z \mid H_0)$.

4. **Pick a significance level and determine the rejection region.**
We will usually use α to denote the significance level. The Neyman-Pearson paradigm is to pick α in advance. Typical values are .1, .05, .01. Recall that the significance level is the probability of a type I error, i.e. of incorrectly rejecting the null hypothesis when it is true. The value we choose will depend on the consequences of a type I error.

Example 9. If $\alpha = .1$ then we'd expect to make a type I error in 10% of those experiments where the null hypothesis was true. If you're running an experiminent to determine if your chocolate is more than 72% cocoa then a 10% error type I error rate, i.e. falsely believing some 72% chocalate is greater that 72%, is probably acceptable. If your forensic lab is identifying fingerprints for a murder trial then a 10% type I error rate, i.e. mistakenly claiming that fingerprints found at the crime scene belonged to someone who was truly innocent, is definitely not acceptable.

If H_0 is composite then P(type I error) depends on which member of H_0 is true and significance level is defined as the maximum of these probabilities.

Once the significance level is chosen we can determine the rejection region in the tail(s) of the null distribution. In Example 4, H_A is two sided so the rejection region is split between the two tails of the null distribution. This distribution is given in the following table:

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

If we set $\alpha = .05$ then the rejection region must contain at most .05 probability. For a two-sided rejection region we get

$$\{0, 1, 9, 10\}.$$

If we set $\alpha = .01$ the rejection region is

$$\{0, 10\}.$$

Suppose we change H_A to 'the coin is biased in favor of heads'. We now have a one-sided hypothesis $\theta > .5$. Our rejection region will now be in the right-hand tail since we don't want to reject H_0 in favor of H_A if we get a small number of heads. Now if $\alpha = .05$ the rejection region is the one-sided range

$$\{9, 10\}.$$

If we set $\alpha = .01$ then the rejection region is

$$\{10\}.$$

5. **Determine the power(s).**
As we saw in Example 4, once the rejection region is set we can determine the power of the test at various values of the alternate hypothesis.

6 p-values

In practice people often specify the significance level and do the significance test using p-values. If the p-value is less than the significance level α they reject H_0. Otherwise they do not reject H_0.

The p-value is the probability, assuming the null hypothesis, of seeing data at least as extreme as the experimental data. What 'at least as extreme' means depends on the experimental design. We illustrate the definition and use of p-values with an example.

Example 10. The z-test for normal hypotheses
IQ is normally distributed in the population according to a $N(100, 15^2)$ distribution. We suspect that most MIT students have above average IQ so we frame the following hypotheses.

H_0 = MIT student IQs are distributed identically to the general population
= MIT IQ's follow a $N(100, 15^2)$ distribution.

H_A = MIT student IQs tend to be higher than those of the general population
= the average MIT student IQ is greater than 100.

Notice that H_A is one-sided.

Suppose we test 9 students and find they have an average IQ of $\bar{x} = 112$. Can we reject H_0 at a significance level $\alpha = .05$?

answer: The average $\bar{x} = 112$ is the result of one experiment. If we ran the experiment again we could get a different value for $\bar{x}$. For a one-sided alternative hypothesis the phrase 'data at least as extreme' is a one-sided tail. The p-value is then

$$p = P(\bar{x} \geq 112 \,|\, H_0).$$

That is, it is the probability, assuming H_0, that the experiment would produce data as extreme as 112.

To compute p we standardize the data to get a z-statistic

$$z = \frac{\bar{x} - 100}{15/\sqrt{9}} = \frac{36}{15} = 2.4.$$

Under the null hypothesis $\bar{x} \sim \mathrm{N}(100, 15^2/9)$ and therefore $z \sim \mathrm{N}(0, 1)$. That is, the null distribution for z is standard normal.

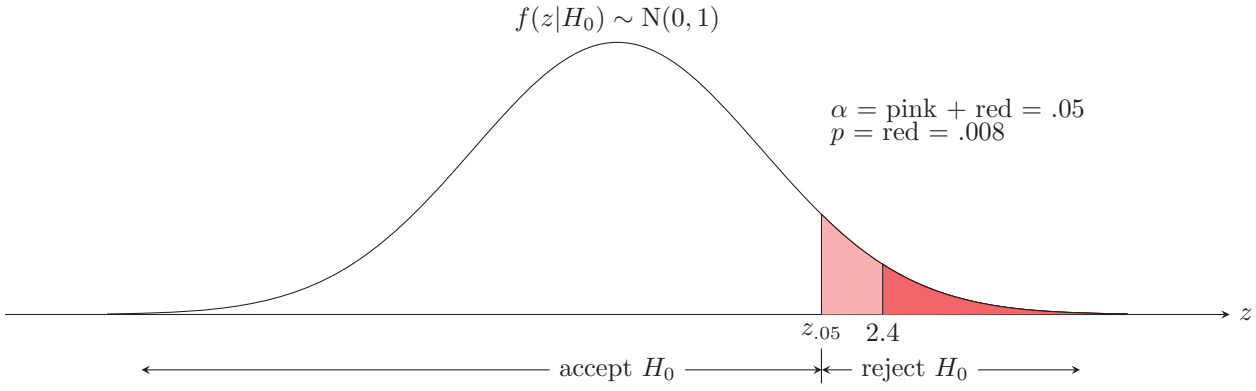
$$p = P(\bar{x} \geq 112 \,|\, H_0) = P(z \geq 2.4) = \texttt{1- pnorm(2.4,0,1) = 0.0081975}.$$

Since $p \leq \alpha$ we reject the null hypothesis in favor of the alternative hypothesis that MIT students have higher IQs on average. We have done this at significance level .05 with a p-value of .008. (The computation was done in R using the function pnorm. Below we use the function qnorm.)

We can rephrase this directly in terms of rejection regions. In the figure below, the rejection region is the shaded tail to the right of

$$z_{.05} = \texttt{qnorm(.95,0,1)} = 1.65.$$

The p-value is the area in the tail to the right of $z = 2.4$. Since z is in the rejection region, we reject H_0.



Note that we can use the language of rejection regions or p-values because:

$$z \text{ is in the rejection region if and only if } p \leq \alpha.$$

7 More examples

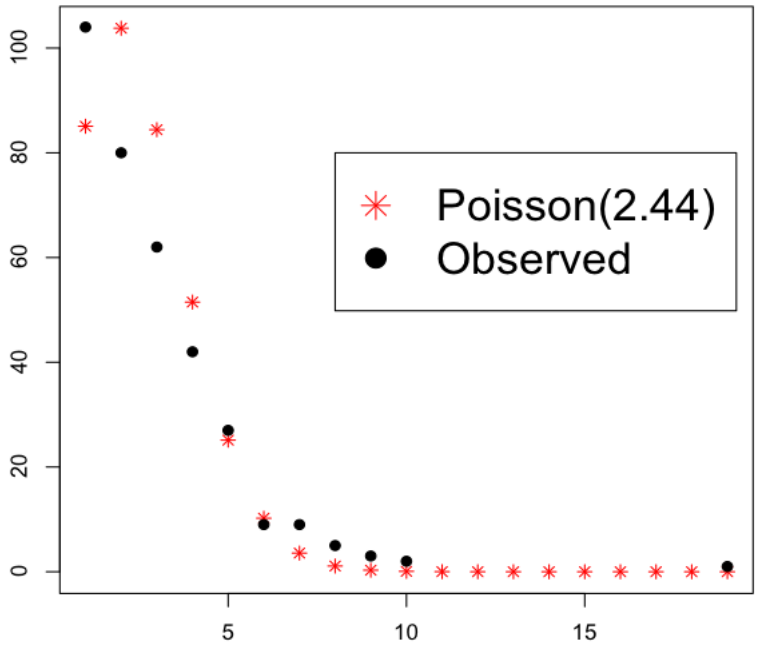
Hypothesis testing is widely used in inferential statistics. Read these examples quickly to get a sense of how it's used. We will explore the details of these examples in class.

Example 11. The chi square statistic and goodness of fit. (Rice, example B, p.313)

To test the level of bacterial contamination, milk was spread over a grid with 400 squares. The amount of bacteria in each square was counted. We summarize in the table below. The bottom row of the table is the number of different squares that had a given amount of bacteria.

Amount of bacteria	0	1	2	3	4	5	6	7	8	9	10	19
Number of squares	56	104	80	62	42	27	9	9	5	3	2	1

So the average amount of bacteria per square is 2.44. To see if these counts could come from a Poisson distribution we graphically compare the observed frequencies with those expected from Poiss(2.44).



The picture is suggestive, so we do a hypothesis test with

H_0 : the samples come from a Poiss(2.44) distribution.

H_A : the samples come from a different distribution.

We use a chi square statistic, so called because it (approximately) follows a chi square distribution. To compute X^2 we first combine the last few cells in the table so that the minimum expected count is around 5 (a general rule-of-thumb in this game.)

The expected number of squares with a certain amount of bacteria comes from considering 400 trials from a Poiss(2.44) distribution, e.g., with $l = 2.44$ the expected number of squares with 3 bacteria is $400 \times e^{-l}\dfrac{l^3}{3!} = 84.4$.

The chi square statistic is $\sum \dfrac{(O_i - E_i)^2}{E_i}$, where O_i is the observed number and E_i is the expected number.

Number per square	0	1	2	3	4	5	6	> 6
Observed	56	104	80	62	42	27	9	20
Expected	34.9	85.1	103.8	84.4	51.5	25.1	10.2	5.0
Component of X^2	12.8	4.2	5.5	6.0	1.7	0.14	0.15	44.5

Summing up we get $X^2 = 74.9$.

Since the mean (2.44) and the total number of trials (400) are fixed, the 8 cells only have 6 degrees of freedom. So, assuming H_0, our chi square statistic follows (approximately) a χ^2_6 distribution. Using this distribution, $P(X^2 > 74.59) = 0$ (to at least 6 decimal places). Thus we decisively reject the null hpothesis in favor of the alternate hypothesis that the distribution is not Poiss(2.44).

To analyze further, look at the individual components of X^2. There are large contributions in the tail of the distribution, so that is where the fit goes awry.

Example 12. Student's t test.

Suppose we want to compare a medical treatment for increasing life expectancy with a placebo. We give n people the treatment and m people the placebo. Let $X_1, \dots, X_n$ be the number of years people live after receiving the treatment. Likewise, let $Y_1, \dots, Y_m$ be the number of years people live after receiving the placebo. Let $\bar{X}$ and $\bar{Y}$ be the sample means. We want to know if the difference between $\bar{X}$ and $\bar{Y}$ is statistically significant. We frame this as a hypothesis test. Let μ_X and μ_Y be the (unknown) means.

$$H_0 : \mu_X = \mu_Y, \quad H_A : \mu_X \neq \mu_Y.$$

With certain assumptions and a proper formula for the pooled standard error s_p the test statistic $t = \dfrac{\bar{X} - \bar{Y}}{s_p}$ follow a t distribution with $n + m - 2$ degrees of freedom. So our rejection region is determined by a threshold t_0 with $P(t > t_0) = \alpha$.

18.05 Introduction to Probability and Statistics

Spring 2014