7. Hypothesis Testing and Inference – Fisher, Neyman-Pearson, Frequentist Hybrid, Confidence Intervals and Bayes

Readings: Rosner: 7.1-7.4
Goodman – p-value fallacy (1999) and Bayes factor (1999) papers
Benjamin et al. (2017) – Redefining statistical significance
Chihara and Hesterberg, Ch. 3
OpenIntro Statistics: 4.2-4.3

Homework: Homework 3 due by noon on September 24
Homework 4 due by noon on October 1

Overview

A) Hypothesis tests
B) Approaches to hypothesis testing
C) One-sided vs. two-sided tests
D) General notes/summary

A) Hypothesis Tests

Hypothesis testing is a method of making inferences about a population quantity from a data sample. We begin with a statement or "hypothesis" about the population and use data to determine if the hypothesis is supportable or not.

A *hypothesis* is a claim or statement about a population parameter (or parameters). A *hypothesis test* is a statistical method of quantifying evidence (using sample information) to reach a decision about a hypothesis.

e.g. Recommended daily allowance of zinc for males over 50 is 15 mg/day. A study found a sample of 115 men aged 65-74 had an average intake of 11.3 mg/day and the s.d. of intake was 6.4 mg/day. Does the study indicate too little zinc for these men?

e.g. Does a full moon affect mental health? A Virginia psychiatric clinic collected data on number of admissions on the 12 full moons in August 1971-July 1972: 5, 13, 13, 16, 14, 25, 12, 13, 6, 14, 9, 20. For the rest of the year, admissions averaged 11.2 per day.

e.g. Serum cholesterol for 20-74 year-old US males has a population mean of 211 mg/dL and a s.d. of 46 mg/dl. A sample of 12 smokers with hypertension gives $\overline{X}$=217 mg/dL. Is there evidence that hypertensive smokers have different cholesterol levels than the general male population?

B) Approaches to Hypothesis Testing and Statistical Inference

1. Fisher's p-value – generate a null distribution from the data

2. Neyman-Pearson's tradeoffs – H_0, H_1, Type I and Type II errors, critical regions, likelihood ratios

3. Confidence intervals – frequentist improvement on p-values

4. Bayesian methods – prior probabilities, Bayes factors (likelihood ratios)

B1. Basics of Fisher permutation tests

Example from Chihara and Hesterberg (Section 3.3). Time in seconds it takes for a mouse to complete a maze where 3 mice have been given an experimental drug and 3 mice have been given placebo.

Experimental Group				*Control Group*		
30	25	20		18	21	22

The average time for mice on the drug is 25 seconds whereas the average time for the control group is 20.33 seconds.

Is this difference of 4.67 seconds meaningful?

Fisher's idea: There are $\binom{6}{3}$ = 20 random permutations of this data. These random shufflings of the data represent the possible differences if the two groups are really no different from each other.

TABLE 3.1 All Possible Distributions of {30, 25, 20, 18, 21, 22} into Two Sets

Drug			Control			$\bar{X}_D$	$\bar{X}_C$	Difference in means
18	20	21	22	25	30	19.67	25.67	−6.00
18	20	22	21	25	30	20	25.33	−5.33
18	20	25	21	22	30	21	24.33	−3.33
18	20	30	21	22	25	22.67	22.67	0.00
18	21	22	20	25	30	20.33	25	−4.67
18	21	25	20	22	30	21.33	24	−2.67
18	21	30	20	22	25	23	22.33	0.67
18	22	25	20	21	30	21.67	23.67	−2.00
18	22	30	20	21	25	23.33	22	1.33
18	25	30	20	21	22	24.33	21	3.33
20	21	22	18	25	30	21	24.33	−3.33
20	21	25	18	22	30	22	23.33	−1.33
20	21	30	18	22	25	23.67	21.67	2.00
20	22	25	18	21	30	22.33	23	−0.67
20	22	30	18	21	25	24	21.33	2.67
20	25	30	18	21	22	**25**	**20.33**	**4.67** *
21	22	25	18	20	30	22.67	22.67	0.00
21	22	30	18	20	25	24.33	21	3.33
21	25	30	18	20	22	**25.33**	**20**	**5.33** *
22	25	30	18	20	21	**25.67**	**19.67**	**6.00** *

Rows where the difference in means exceeds the original value are highlighted.

Our observed results

Among these 20 “permutations” of the data, 3 result in differences as large or larger than 4.67.

What does this tell us?

Specifically, it tells us that the observed difference of 4.67 seconds between drug and control groups is not that unlikely assuming that there's truly no difference between the groups (the "null" hypothesis).

In fact, 3/20 times, or 15% of the time, we would see a difference as or more extreme than what was observed simply by chance alone. The distribution of differences over the 20 "permutations" is called a *null* or *permutation* distribution, and the probability of seeing the result from the original data or anything more extreme is called the *p-value* for the test or comparison.

This is the basic idea around permutation tests. We'll revisit them in greater detail – assumptions, steps involved, interpretation, etc. – in a few weeks: Nonparametric tests.

B2a. Neyman and Pearson, H_0, H_1, Likelihood ratios and Type I and II errors

We start by stating a "null hypothesis": H_0 - read "H naught". This is a claim that is initially assumed to be true. The wording usually states something along the lines of:

- "There is *no change* between …"
- "… no difference…"
- "… no effect of …"
- "…no association …"

H_0 is where we place the burden of proof for data—what we would actually like to *disprove*.

Form of H_0: population characteristic = hypothesized value; e.g., H_0: $\mu = 90$ mmHg

Then we state an opposing, or alternative hypothesis: H_1 (or H_A). This statement contradicts H_0 so that the null and alternative hypotheses cannot both be true. H_1 is what we would like to prove to be true.

Form of H_1: can have the same form as H_0 with > or < or $\neq$ in place of =; e.g. H_0: $\mu \neq 90$ mmHg

We collect data assuming H_0 is true. Then we test that assumption. We make a decision about the truth of H_0. We recognize the role of chance in our decision-making.

When we make a decision about H_0 and H_1 from the data, there are 4 possibilities:

1) H_0 is true and we fail to reject H_0 (i.e., we say it is "true")
2) H_0 is true and we reject H_0 (i.e., we say it is "false")
3) H_0 is false and we fail to reject H_0
4) H_0 is false and we reject H_0

Type I error: Probability of rejecting the null hypothesis (H_0) when it is true (possibility #2). "Reject $H_0 \mid H_0$ is true" - usually considered the more serious error

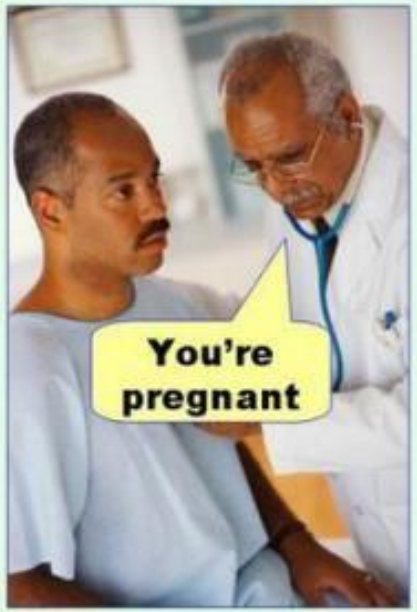
Type II error: Probability of failing to reject the null hypothesis when H_0 is false (possibility #3) – "Fail to Reject $H_0 \mid H_0$ is false"

Note that the probabilities for each of these errors are at odds with each other; i.e. as we increase the probability of making a Type I error we reduce the probability of making a Type II error, and vice versa.

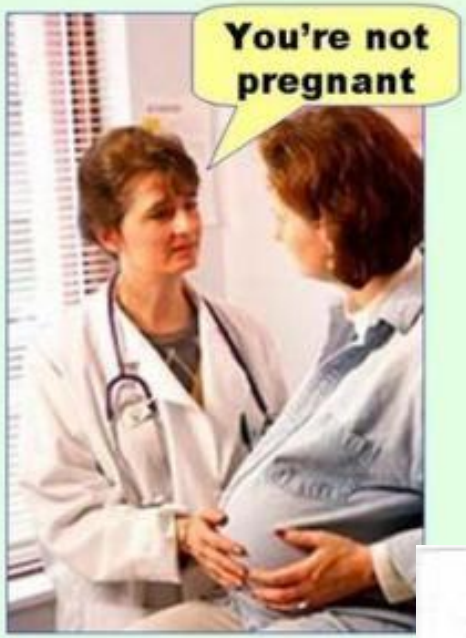
Based on the data we collect to address H_0, we make a decision to reject H_0 or not to reject H_0. Note that we don't "accept H_0" or say "H_0 is true" - all we can say is that we have evidence to reject it or we don't: we "reject H_0" or we "fail to reject H_0".

Reality $\Rightarrow$ What we decide $\Downarrow$	**H_0 True**	**H_0 False/H_1 True**
Fail to reject H_0	*Correct* Probability of correct decision = $1-\alpha$ = Level of confidence	*Type II Error* P(Type II Error) = β
Reject H_0	*Type I error* P(Type I error) = α Level of significance	*Correct* Probability of correct decision = $1-\beta$ = Power

Type I error
(false positive)



Type II error
(false negative)



Never confuse Type I and II errors again:

Just remember that the Boy Who Cried Wolf caused both Type I & II errors, in that order.

First everyone believed there was a wolf, when there wasn't. Next they believed there was no wolf, when there was.

Substitute "effect" for "wolf" and you're done.

Kudos to @danolner for the thought. Illustration by Francis Barlow "De pastoris puero et agricolis" (1687). Public Domain. Via wikimedia.org

For a simple null hypothesis, e.g. H_0: μ = 211 mg/dL (cholesterol) vs. H_1: μ = 225 mg/dL. N-P proposed using a likelihood ratio test to guide the decision about whether the data support H_0 or H_1 while limiting the probability of a Type I error.

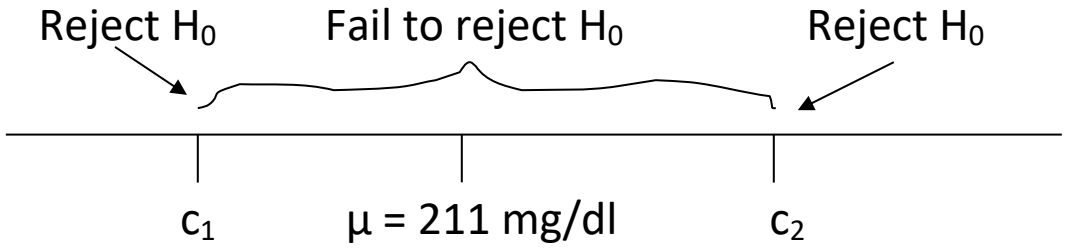
Recall from the lecture on Random Variables and Distributions (Lecture 3):

> If $X_1, X_2, \ldots, X_n$ follow the same distribution $f_X(x; \theta)$, then the likelihood function *L* for a sample of n independent and identically distributed observations, $x_1, x_2, \ldots, x_n$: $L \propto \prod_{i=1}^{n} f_X(x_i; \theta)$, where θ is a population parameter(s) that define the distribution.

The likelihood ratio is $LR = \frac{\prod_{i=1}^{n} f_X(x_i;\mu_0)}{\prod_{i=1}^{n} f_X(x_i;\mu_1)}$.

The null hypothesis would be rejected if LR is *too small* (i.e. the data are more likely under the alternative than the null hypothesis).

Defining a rejection region or critical value: the rejection region is the range of potentially observable values for which we would reject H_0. Where do we draw the boundary between rejecting and not rejecting the hypothesis?



First, we must specify the significance level of the test:
α = P(Type I error) = P(rejecting H_0 | H_0 is true).

This is frequently set at α = 0.05, but other reasonable values are 0.01, 0.10, etc. – depending on the severity of committing a Type I error.

Once α is set, then we find a cutoff for observed values of the sample mean $\bar{X}$ so that this holds. For this example, let's assume that the individual cholesterol values X_i are iid N(μ, σ^2).

Calculation: For α = 0.05 we want to identify constants c_1 and c_2 so that we will reject H_0 if: $\bar{X} < c_1$ or $\bar{X} > c_2$, and fail to reject H_0 if $c_1 < \bar{X} < c_2$.

To find c_1 and c_2 we use the following:

$$P(c_1 < \bar{X} < c_2 \mid H_0 \text{ true}) = 1 - \alpha.$$

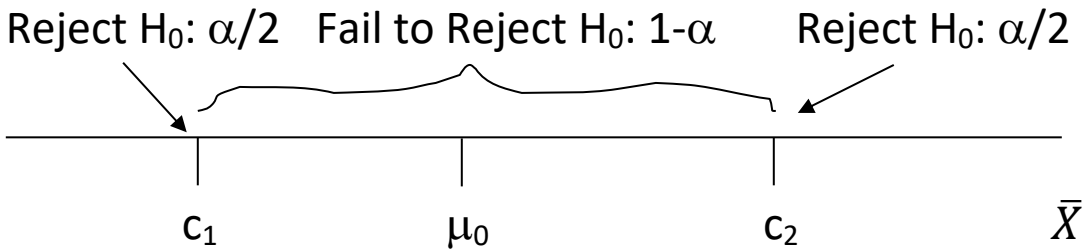
When H_0 is true, we would expect to see values of $\bar{X}$ outside the region α x 100% of the time. Since we assumed our data comes from a normal distribution, we can derive boundaries as:

$$P\left(-Z_{1-\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < Z_{1-\alpha/2} \mid H_0 \text{ true}\right) = 1 - \alpha$$

$$P\left(-Z_{1-\alpha/2} \times \sigma/\sqrt{n} < \bar{X} - \mu_0 < Z_{1-\alpha/2} \times \sigma/\sqrt{n} \mid H_0 \text{ true}\right) = 1 - \alpha$$

$$P\left(\mu_0 - Z_{1-\alpha/2} \times \sigma/\sqrt{n} < \bar{X} < \mu_0 + Z_{1-\alpha/2} \times \sigma/\sqrt{n} \mid H_0 \text{ true}\right) = 1 - \alpha$$

$$P(c_1 < \bar{X} < c_2 \mid H_0 \text{ true}) = 1 - \alpha$$



For the cholesterol example, let n = 12, $\bar{X}$ = 217 mg/dL, $\sigma^2 = 46^2$ (mg/dL)^2, μ_0 = 211 mg/dL, and α=0.05. Then, starting with our standard normal "Z" notation we have the probability of observing an extreme value beyond *c* of:

$$P(|Z| > c \mid H_0) = \alpha = 0.05$$

$$P\left(\left|\frac{\bar{X} - 211}{46/\sqrt{12}}\right| > c \mid H_0: \mu_0 = 211\right) = 0.05$$

Using $Z_{.975}$ = 1.96 and the work on the previous slide we can determine our rejection regions:

$$P\left(\mu_0 - Z_{1-\alpha/2} \times \sigma/\sqrt{n} < \bar{X} < \mu_0 + Z_{1-\alpha/2} \times \sigma/\sqrt{n} \mid \mathrm{H}_0\right) = 1 - \alpha$$

$$P\left(211 - 1.96 \times \left(46/\sqrt{12}\right) < \bar{X} < 211 + 1.96 \times \left(46/\sqrt{12}\right) \mid \mathrm{H}_0\right) = 1 - 0.05$$

$$P(\,185\ \mathrm{mg/dL} < \bar{X} < 237\ \mathrm{mg/dL}) = 0.95$$

In other words, we would *fail to reject H_0* if our sample mean is between 185 and 237 mg/dL. Therefore, since $\bar{X}$ = 217 mg/dL, we would *fail to reject* H_0: μ = 211 mg/dL. There is not enough evidence to reject ("disprove") the null hypothesis that the sample came from a population with a mean cholesterol of 211 mg/dL.

B2b. The Fisher-Neyman-Pearson "hybrid" method commonly used:

p-value interpretations:

1) The probability of obtaining a result as extreme or more extreme than the one observed in the sample under the null distribution (Fisher). This is usually obtained from a *parametric sampling distribution* that the *test statistic* for the data is assumed to follow.
2) The alpha level that would have had to have been specified to just (barely) reject H_0 based on the observed data (Neyman-Pearson).

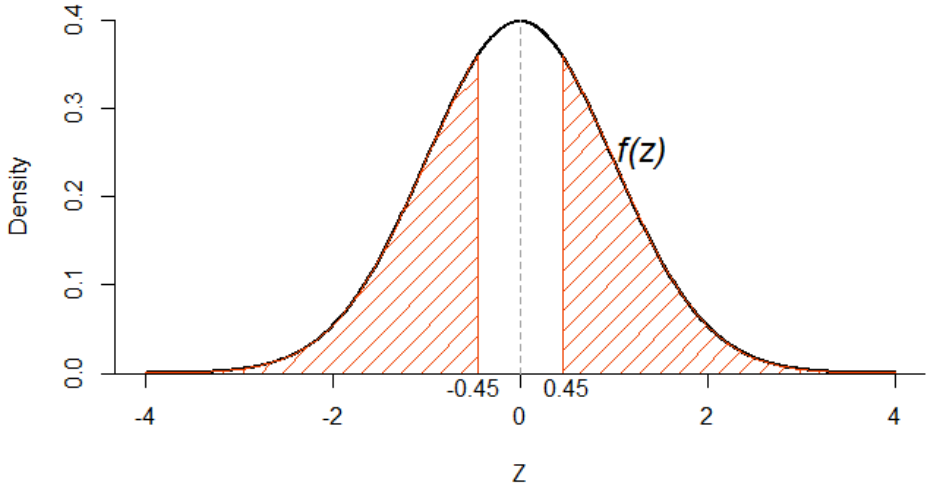
What's the *probability* of observing an $\bar{X}$ value as or more extreme than 217 mg/dL if the true value for the mean μ is 211 mg/dL? (Recall: $\sigma^2 = 46^2$ (mg/dL)^2 and n=12.)

$$P(|\bar{X} - 211| > |217 - 211| \mid H_0: \mu = 211 \text{ mg/dL})$$

$$P\left(\left|\frac{\bar{X}-211}{46/\sqrt{12}}\right| > \frac{6}{46/\sqrt{12}} \mid H_0: \mu = 211 \text{ mg/dL}\right)$$

$$P(|Z| > 0.45) = (1 - 0.6736) \times 2 = 0.6528$$

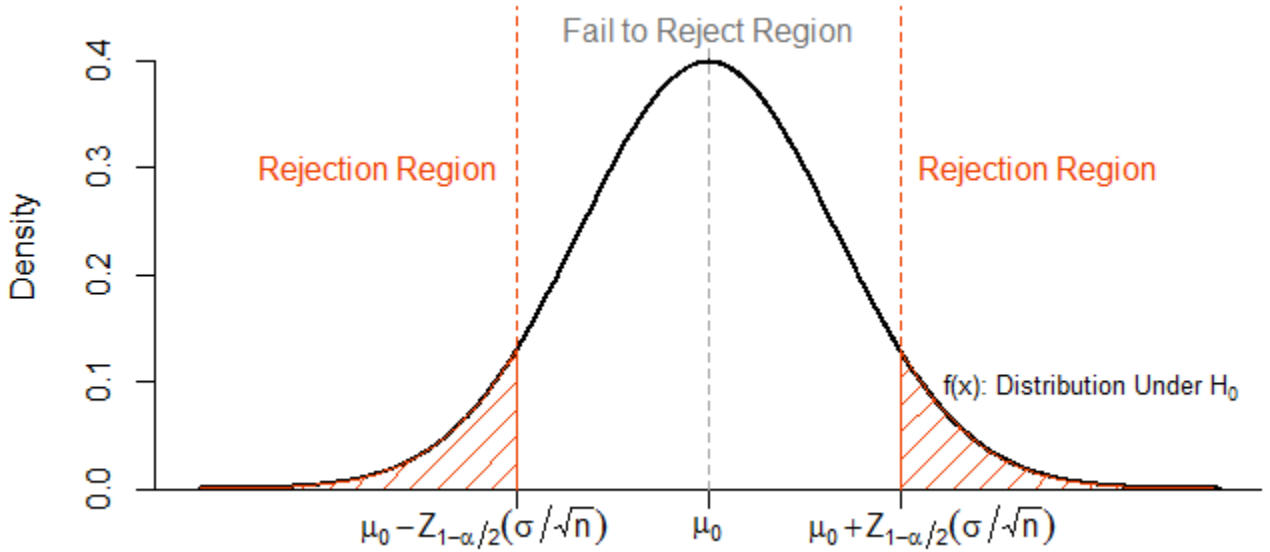
Conclusion: p > 0.05 so there is not enough evidence to reject H_0: μ = 211 mg/dL.



B3. Confidence intervals and hypothesis tests – for more detail on CI *per se*, see Lecture SA5 Confidence Intervals

For the hypothesis test H_0: $\mu = \mu_0$ vs. H_1: $\mu \neq \mu_0$, we define the fail to reject region for our observed $\bar{X}$ with the boundaries:

$$\left(\mu_0 - Z_{1-\alpha/2}\left(\sigma/\sqrt{n}\right), \mu_0 + Z_{1-\alpha/2}\left(\sigma/\sqrt{n}\right)\right)$$



Recall for the population mean μ that the form of a (1-α)x100% CI, when σ^2 is known, is:

$$\bar{X} \pm Z_{1-\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

If for H_0: $\mu = \mu_0$ we choose a value of μ_0 *inside* the CI around $\bar{X}$, then we see that:

$$\bar{X} - Z_{1-\alpha/2} \times \left(\sigma/\sqrt{n}\right) < \mu_0 < \bar{X} + Z_{1-\alpha/2} \times \left(\sigma/\sqrt{n}\right)$$

$$-Z_{1-\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < Z_{1-\alpha/2}$$

$$\mu_0 - Z_{1-\alpha/2} \times \sigma/\sqrt{n} < \bar{X} < \mu_0 + Z_{1-\alpha/2} \times \sigma/\sqrt{n}$$

Thus, $\bar{X}$ is in the fail to reject region for our null hypothesis H_0: $\mu = \mu_0$.

Recall for the population mean μ that the form of a (1-α)x100% CI, when σ^2 is known, is:

$$\bar{X} \pm Z_{1-\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

If, instead, we choose for H_0:$\mu = \mu_0$ a value of μ_0 *outside* of the CI, then:

$$\mu_0 < \bar{X} - Z_{1-\alpha/2} \times \left(\sigma/\sqrt{n}\right) \quad or \quad \mu_0 > \bar{X} + Z_{1-\alpha/2} \times \left(\sigma/\sqrt{n}\right)$$

$$\frac{\mu_0 - \bar{X}}{\sigma/\sqrt{n}} < -Z_{1-\alpha/2} \qquad or \qquad \frac{\mu_0 - \bar{X}}{\sigma/\sqrt{n}} > Z_{1-\alpha/2}$$

$$\bar{X} > \mu_0 + Z_{1-\alpha/2} \times \left(\sigma/\sqrt{n}\right) \quad or \quad \bar{X} < \mu_0 - Z_{1-\alpha/2} \times \left(\sigma/\sqrt{n}\right)$$

Thus, since $\bar{X}$ is in the rejection region, we are led *to reject* the null hypothesis H_0: $\mu = \mu_0$.

Therefore, a CI contains all of the values of μ_0 for H_0: $\mu = \mu_0$ for which we would *fail to reject* H_0, and all values outside the CI are values of μ_0 for which we would *reject* H_0.

e.g. Mean change in cholesterol before vs. after vegetarian diet. $\bar{X}$ change = 12 mg/dL; n = 15; σ^2 = 100 $(mg/dL)^2$

95% CI: $\bar{X} \pm Z_{1-\alpha/2}\left(\sigma/\sqrt{n}\right) =$

If H_0: μ change = 0 mg/dL vs. H_1: μ change $\neq$ 0 mg/dL, then based on the CI ….

If H_0: μ change = 10 mg/dL vs. H_1: μ change $\neq$ 10 mg/dL, then based on the CI …

B4. Bayesian inference – briefly; more in a couple of weeks …

Bayes Theorem:

$$P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)}$$

Which is a re-statement of conditional probability: $P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$.

If θ is an unknown parameter of a distribution (mean, variance, etc.) then the implication of the above is that θ is a random variable, just as X is a random variable, and that it follows a probability distribution, P(θ), known as a *prior* distribution.

The prior distribution can be based on previous studies, past experience or subjective belief. It can be non-informative (not weighted toward specific values of θ) or informative (heavily weighted toward specific values for θ).

Connecting Bayes Theorem to Bayesian terminology we have:

$$P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)} \rightarrow \text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Data}}$$

With regard to hypothesis testing, this can be framed as

$$\text{Posterior odds of } H_0 \text{ (i.e., after observing the data)} = \text{Prior odds of } H_0 \text{ (before observing the data)} \times \frac{P(\text{Data}|H_0)}{P(\text{Data}|H_1)}$$

which we can think of as "belief altered by data" (Goodman, 1999).

$\frac{P(\text{Data}|H_0)}{P(\text{Data}|H_1)}$ is a likelihood ratio (the same quantity used earlier in the lecture by Neyman-Pearson for decision-making!), also called **the Bayes factor**. It represents how much the prior odds of the null hypothesis being true are altered by the data. In practice, the data must be very strongly in favor of the alternative hypothesis in order to overturn the prior impression about H_0 being true.

Let's take a look at the table in the Bayes Factor paper by Goodman (1999)…

Table 1. Final (Posterior) Probability of the Null Hypothesis after Observing Various Bayes Factors, as a Function of the Prior Probability of the Null Hypothesis

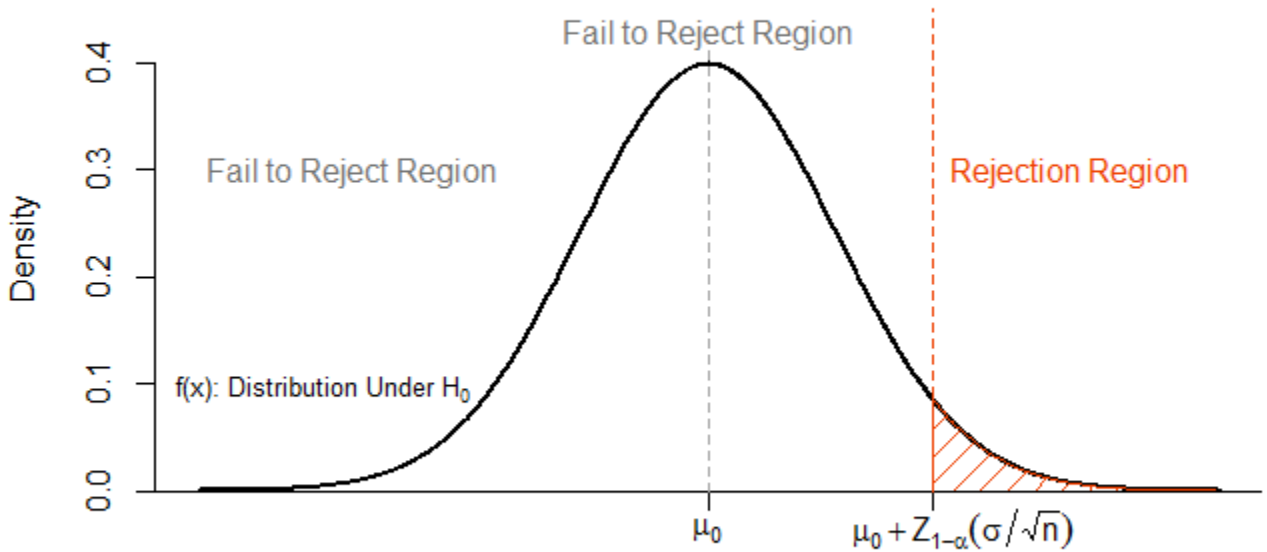
Strength of Evidence	Bayes Factor	Decrease in Probability of the Null Hypothesis	
		From	To No Less Than
		%	
Weak	1/5	90	64*
		50	17
		25	6
Moderate	1/10	90	47
		50	9
		25	3
Moderate to strong	1/20	90	31
		50	5
		25	2
Strong to very strong	1/100	90	8
		50	1
		25	0.3

* Calculations were performed as follows:
A probability (Prob) of 90% is equivalent to an odds of 9, calculated as Prob/(1 − Prob).
Posterior odds = Bayes factor × prior odds; thus, (1/5) × 9 = 1.8.
Probability = odds/(1 + odds); thus, 1.8/2.8 = 0.64.

We'll revisit Bayesian methods when we delve into conditional probability (for things such as the Positive Predictive Value, Negative Predictive Value, Likelihood Ratio +, Likelihood Ratio -) in a few weeks and also take a look at the results of a clinical trial (GUSTO II) evaluated using both frequentist and Bayesian methods.

C. One-sided vs. two-sided hypothesis tests

Testing null hypothesis against a directional alternative. The probability of (incorrect) rejection (α) is contained in one extreme (tail) of the sampling distribution (i.e., it does not have to be divided between both tails like with the two-sided tests).



e.g. Cholesterol study with H_0: $\mu = 211$ mg/dL *versus* H_1: $\mu > 211$ mg/dL
We reject null if $\bar{X}$ > constant c:

$$P(\bar{X} > c) = 1 - \alpha$$

$$P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha}\right) = 1 - \alpha$$

$$P\left(\bar{X} > \mu_0 + Z_{1-\alpha}\left(\sigma/\sqrt{n}\right)\right) = 1 - \alpha$$

$$P(\bar{X} > c) = 1 - \alpha$$

Random sample n = 12, $\bar{X}$ = 217 mg/dL, $\sigma^2 = 46^2$ (mg/dL)^2, α = .05

We reject H_0: $\mu = 211$ mg/dL if

$$\bar{X} > \mu_0 + Z_{1-\alpha}\left(\sigma/\sqrt{n}\right) \rightarrow 217 > 211 + 1.645\left(46/\sqrt{12}\right) = 232.8 \text{ mg/dL}$$

(Compare critical value of 232.8 mg/dl with critical value of 237 mg/dl for two-sided critical region.)

Since 217 mg/dL is smaller than 232.8 mg/dL, we (still) do not reject H_0: $\mu = 211$ mg/dL.

D) General Notes/Summary:

- p-value < α whenever $\bar{X}$ is in the rejection region

- When p-value < α we say that the result is "significant at the α x 100% level". For definitive studies, we might want very small α ("conservative"). For a pilot study, we might use a larger value such as $\alpha = 0.10$ ("liberal").

- If p-value > α, we do not "accept H_0", we just "don't reject H_0". Often a large p-value and non-significant result mean we did not have enough data to disprove H_0, i.e. a Type II error has occurred. We will touch on issues of power and sample in the next lecture.

- If the null hypothesis is actually false, the p-value will tend to decrease as n increases.

- Carrying out many significance tests in a single study is problematic and requires some thought. For instance, if we were to continue drawing samples of 12 participants (as in the cholesterol example) and computed p-values, about 5% of the tests would have $p < 0.05$ ***even if H_0 is true***. We will wrongly reject H_0 about 5% of the time. (This is what $\alpha = 0.05$ really means.)

- There is an important difference between *statistical significance* and *clinical or practical significance*. With increasing sample size, smaller differences will appear statistically significant ($p < \alpha$). It is very important to ask if the result is also important clinically, practically, from a public health perspective, etc. A statistical test gives us information about how likely the observed result (or anything more extreme) is if H_0 is true, i.e. if the result is statistically significant or not. The actual departure of the observed result from what is expected under H_0 relates to the practical (or clinical) significance. For example, with enough subjects we could find statistical significance in the cholesterol example above with μ = 211 mg/dL and $\bar{X}$ = 217 mg/dL, but the difference has no practical or clinical significance. On this issue, be sure to revisit Jessica Utts' paper: *What Educated Citizens Should Know About Statistics and Probability*. (Paper repository in Canvas)

- One-sided hypothesis tests are anti-conservative, i.e., they more often lead to rejection of the null hypothesis than the corresponding two-sided test at the same level of significance.

- Significance tests and CI's depend on assumptions (random sample, normality, etc.). If the assumptions are not met, the results may not give a true indication of what is in the data.

Limitations of Fisherian, Neyman-Pearson and Bayesian inference

Fisher's p-values – size of effect not taken into account; often misinterpreted as probability of truth of H_0

N-P decision making - α, β have taken on fixed values without a true meaning to back up their choice; no measure of evidence from the data used as a summary; expectation that, in the long run, correct decisions will mostly be made, but no focus on truth for an individual hypothesis test

Bayesian inference – prior probabilities are often subjective and can greatly influence inferences drawn from data; different prior distributions can lead to different inferences