14.310x Lecture 13

Statistics

Where were we anyhow?

Summary to date

Probability basics

Introduced concept and talked about simple sample spaces, independent events, conditional probabilities, Bayes Rule

Random variables

Defined a random variable, discussed ways to represent distributions (PF, PDF, CDF), covered random variable versions of concepts above

Functions of random variables

Saw some basic strategies and several important examples

Summary to date

Moments

Defined moments of distributions and learned many techniques and properties to help compute moments of functions of random variables

Special distributions

Binomial, hypergeometric, geometric, negative binomial, Poisson, exponential, uniform, normal

Estimation

CLT, had general discussion and discussion about sample mean, criteria for assessing, frameworks for deriving

Statistics---quantifying reliability

We have seen various estimators, made observations about their distributions, discussed how we might derive them, and also discussed criteria that we might use to choose among them.

That's all well and good, but when we actually have to report estimates, people will want to have some objective measure of how good, or reliable, or precise, our estimates are. One way we quantify this is by reporting the variance (or estimated variance) of the estimator along with the estimate.

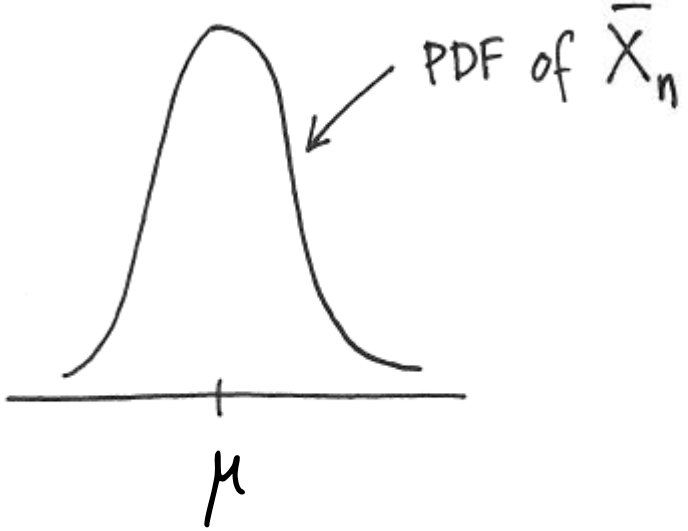
Statistics---quantifying reliability

The standard error of an estimate is the standard deviation (or estimated standard deviation) of the estimator.

Statistics---quantifying reliability

The <u>standard error</u> of an estimate is the standard deviation (or estimated standard deviation) of the estimator.

For example, $\bar{X}_n$ has mean μ and variance σ^2/n, so $SE(\bar{X}_n) = \sigma/\sqrt{n}$ (or $\hat{\sigma}/\sqrt{n}$ if you don't know σ^2 and need to estimate it).



Statistics---quantifying reliability

The <u>standard error</u> of an estimate is the standard deviation (or estimated standard deviation) of the estimator.

For example, $\bar{X}_n$ has mean μ and variance σ^2/n, so $SE(\bar{X}_n) = \sigma/\sqrt{n}$ (or $\hat{\sigma}/\sqrt{n}$ if you don't know σ^2 and need to estimate it).

So we often report an estimate along with its standard error. Sometimes we put the standard error in parentheses right after the estimate.

Statistics---quantifying reliability

The standard error certainly gives us some idea of how tightly concentrated around the unknown parameter the distribution of the estimator is. That's useful, but sometimes it might be useful to report essentially equivalent information in a different form, an interval.

Statistics---quantifying reliability

The standard error certainly gives us some idea of how tightly concentrated around the unknown parameter the distribution of the estimator is. That's useful, but sometimes it might be useful to report essentially equivalent information in a different form, an interval.

In other words, we could construct an interval using information about the distribution of the estimator. That interval will be narrow when the estimator has a tight distribution and wide when it has a dispersed distribution. We'll call it a <u>confidence interval</u>.

Statistics---confidence intervals

We want to find functions of the random sample $A(X_1, \ldots, X_n)$ and $B(X_1, \ldots, X_n)$ such that

$$P(A(X_1, \ldots, X_n) < \theta < B(X_1, \ldots, X_n)) = 1 - \alpha$$

Statistics---confidence intervals

We want to find functions of the random sample $A(X_1, \ldots, X_n)$ and $B(X_1, \ldots, X_n)$ such that

$$P(A(X_1, \ldots, X_n) < \theta < B(X_1, \ldots, X_n)) = 1 - \alpha$$

random functions

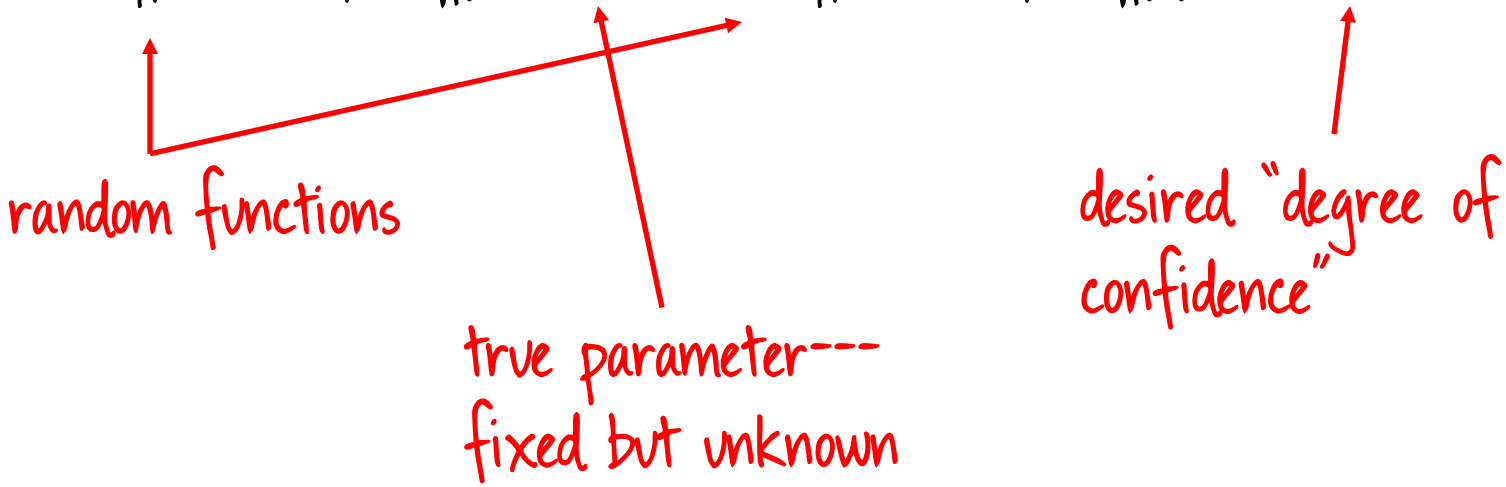
true parameter---
fixed but unknown

desired "degree of confidence"

Statistics---confidence intervals

We want to find functions of the random sample $A(X_1, \ldots, X_n)$ and $B(X_1, \ldots, X_n)$ such that

$$P(A(X_1, \ldots, X_n) < \theta < B(X_1, \ldots, X_n)) = 1 - \alpha$$



(Compare this with point estimation: want a function such that $E(\hat{\theta}(X_1, \ldots, X_n)) = \theta$, for instance.)

Statistics---confidence intervals

If you can find such functions A&B, then

$$[A(x_1, \ldots, x_n), B(x_1, \ldots, x_n)]$$

is said to be a $1-\alpha$ confidence interval for θ.

Statistics---confidence intervals

If you can find such functions A&B, then

$$[A(x_1, \ldots, x_n), B(x_1, \ldots, x_n)]$$

is said to be a $1-\alpha$ confidence interval for θ.

After we plug in the realizations, these are just numbers now.

Statistics---confidence intervals

If you can find such functions A&B, then

$[A(x_1, \ldots, x_n), B(x_1, \ldots, x_n)]$

is said to be a $1-\alpha$ confidence interval for θ.

Notes: ---These functions are not unique. So how do we choose them? Typically, we choose A&B such that $\alpha/2$ of the probability falls on each side of the interval.

---Keep in mind that $A(x_1, \ldots, x_n)$ and $B(x_1, \ldots, x_n)$ are just numbers, so probability statements involving those quantities and θ don't make sense.

Statistics---confidence intervals

Where do we find those functions? You can find them "from scratch." In most cases that you will encounter in everyday data analysis, though, others have found those functions and a resulting formula for a "95% confidence interval for the mean of an unknown distribution with sample size greater than 30" (or whatever).

Statistics---confidence intervals

Where do we find those functions? You can find them "from scratch." In most cases that you will encounter in everyday data analysis, though, others have found those functions and a resulting formula for a "95% confidence interval for the mean of an unknown distribution with sample size greater than 30" (or whatever).

Important to note: in order to find those functions and derive the formulae, one needs to know how an estimator for the unknown parameter is distributed (typically not just the mean and variance).

Statistics---χ^2, t, and F distributions

This is precisely where our friends, the χ^2, t, and F distributions, come in. These are all distributions that, unlike other special distributions we've encountered, don't really appear in nature, don't really describe stochastic phenomena we observe. Rather, they were "invented" because estimators or functions of estimators had distributions that needed to be described and tabulated.

Statistics---χ^2, t, and F distributions

χ^2

Recall that we briefly mentioned an estimator called the sample variance, s^2.

$$S^2 = \frac{1}{n-1} \sum (X_i - \bar{X}_n)^2$$

We said that it was an unbiased estimator for the variance of a distribution.

Well, $(n-1)s^2/\sigma^2$ has one of these distributions, in particular,

$$(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$$

Statistics---χ^2, t, and F distributions

χ^2

Recall that we briefly mentioned an estimator called the sample variance, s^2.

$$S^2 = \frac{1}{n-1} \sum (X_i - \bar{X}_n)^2$$

We said that it was an unbiased estimator for the variance of a distribution.

Well, $(n-1)s^2/\sigma^2$ has one of these distributions, in particular,

$$(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$$

This is a parameter of the χ^2 distribution---it's called "degrees of freedom"

Statistics---χ^2, t, and F distributions

t

If $X \sim N(0,1)$ and $Z \sim \chi^2_n$ and they're independent, then

$X/(Z/n)^{1/2} \sim t_n$

Why is that a useful fact? Suppose we are sampling from a $N(\mu,\sigma^2)$ distribution. We know that $(\bar{X}-\mu)/(\sigma^2/n)^{1/2} \sim N(0,1)$ and that $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$. (We do not know that they're independent but, in fact, they are.)

We form $\dfrac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\sqrt{\dfrac{\Sigma((X_i-\bar{X})/\sigma)^2}{(n-1)}}}$ cancel a few things, and get $\dfrac{\sqrt{n}(\bar{X}-\mu)}{s}$

Statistics---χ^2, t, and F distributions

t

So we have that $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$ for X_i i.i.d. $N(\mu,\sigma^2)$.

Statistics---χ^2, t, and F distributions

t

So we have that $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$ for X_i i.i.d. $N(\mu,\sigma^2)$.

again, "degrees of freedom"

Statistics---χ^2, t, and F distributions

The t distribution was formulated by William Sealy Gosset in his job as Chief Brewer in the Guinness Brewery in Dublin. He derived and tabulated this distribution to aid in his analysis of data for quality control across batches of beer. He published it under the pseudonym "Student" in 1908.

Statistics---χ^2, t, and F distributions

The t distribution was formulated by William Sealy Gosset in his job as Chief Brewer in the Guinness Brewery in Dublin. He derived and tabulated this distribution to aid in his analysis of data for quality control across batches of beer. He published it under the pseudonym "Student" in 1908.

took photo during my recent visit to the brewery

Statistics---χ^2, t, and F distributions

F

If $X \sim \chi^2_n$ and $Z \sim \chi^2_m$ and they're independent, then $(X/n)/(Z/m) \sim F_{n,m}$

Why is that a useful fact? Suppose we have samples from two different populations. We might want to know whether the distributions in the two populations were, in fact, the same. If they are, we can form the ratio of the sample variances divided by their degrees of freedom (true variances canceling because they're the same) and the ratio then has the above distribution.

Statistics---confidence intervals

So let's construct some confidence intervals. We will focus initially on two cases. These are not the only cases you will ever encounter, but they are, by far, the most important.

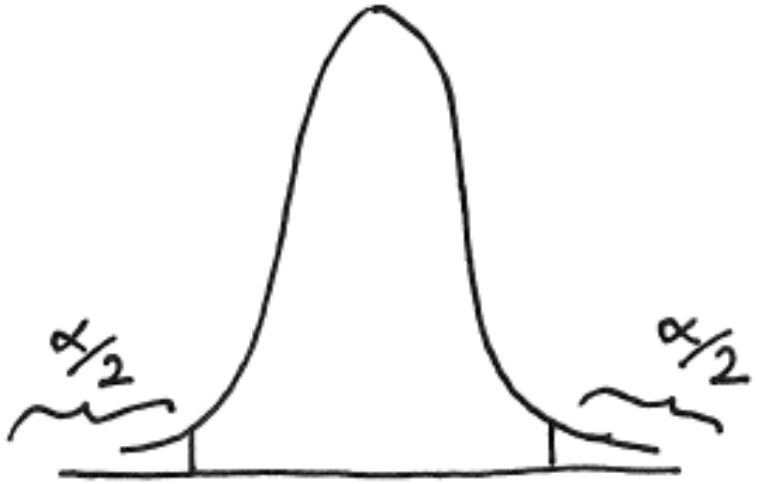
Case 1: We are sampling from a normal distribution with a known variance and we want a confidence interval for the mean.

Case 2: We are sampling from a normal distribution with an unknown variance and we want a confidence interval for the mean.

Statistics---confidence intervals, case 1

We have an estimator for the mean, $\bar{X}$, which has a normal distribution with mean μ and variance σ^2/n.

So $P\{\Phi^{-1}(\alpha/2) < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < -\Phi^{-1}(\alpha/2)\} = 1-\alpha$



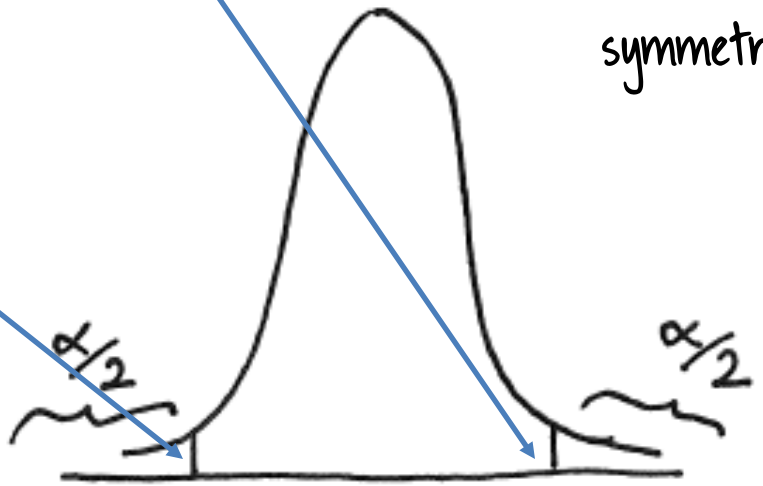
Statistics---confidence intervals, case 1

We have an estimator for the mean, $\bar{X}$, which has a normal distribution with mean μ and variance σ^2/n.

So $P\{\Phi^{-1}(\alpha/2) < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < -\Phi^{-1}(\alpha/2)\} = 1-\alpha$

We just put a negative sign in front of it to get the value for the other tail due to symmetry.

This is the inverse CDF of the standard normal evaluated at $\alpha/2$.



Statistics---confidence intervals, case 1

We have an estimator for the mean, $\bar{X}$, which has a normal distribution with mean μ and variance σ^2/n.

So $P\{\Phi^{-1}(\alpha/2) < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < -\Phi^{-1}(\alpha/2)\} = 1-\alpha$

So, rearranging we have, $P\{\bar{X} + \Phi^{-1}(\alpha/2)\,\sigma/\sqrt{n} < \mu < \bar{X} - \Phi^{-1}(\alpha/2)\,\sigma/\sqrt{n}\} = 1-\alpha$.

Statistics---confidence intervals, case 1

We have an estimator for the mean, $\bar{X}$, which has a normal distribution with mean μ and variance σ^2/n.

So $P\{\Phi^{-1}(\alpha/2) < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < -\Phi^{-1}(\alpha/2)\} = 1-\alpha$

So, rearranging we have, $P\{\bar{X} + \Phi^{-1}(\alpha/2)\,\sigma/\sqrt{n} < \mu < \bar{X} - \Phi^{-1}(\alpha/2)\,\sigma/\sqrt{n}\} = 1-\alpha$.

Our $1-\alpha$ CI is $[\bar{X} + \Phi^{-1}(\alpha/2)\sigma/\sqrt{n}, \bar{X} - \Phi^{-1}(\alpha/2)\sigma/\sqrt{n}]$

Statistics---confidence intervals, case 1

We have an estimator for the mean, $\bar{X}$, which has a normal distribution with mean μ and variance σ^2/n.

So $P\{\Phi^{-1}(\alpha/2) < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < -\Phi^{-1}(\alpha/2)\} = 1-\alpha$

So, rearranging we have, $P\{\bar{X} + \Phi^{-1}(\alpha/2)\,\sigma/\sqrt{n} < \mu < \bar{X} - \Phi^{-1}(\alpha/2)\,\sigma/\sqrt{n}\} = 1-\alpha$.

We plug realizations in and our CI is just two numbers.

Our $1-\alpha$ CI is $[\bar{X} + \Phi^{-1}(\alpha/2)\,\sigma/\sqrt{n},\ \bar{X} - \Phi^{-1}(\alpha/2)\,\sigma/\sqrt{n}]$

Statistics---confidence intervals, case 2

We have an estimator for the mean, $\bar{X}$, which has a normal distribution with mean μ and variance σ^2/n, but we do not know σ^2. We do know, though, that $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$.

So $P\{t_{n-1}^{-1}(\alpha/2) < \frac{\sqrt{n}(\bar{X}-\mu)}{S} < -t_{n-1}^{-1}(\alpha/2)\} = 1-\alpha$

This is the inverse of the t_{n-1} CDF evaluated at $\alpha/2$.

Statistics---confidence intervals, case 2

We have an estimator for the mean, $\bar{X}$, which has a normal distribution with mean μ and variance σ^2/n, but we do not know σ^2. We do know, though, that $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$.

So $P\{t_{n-1}^{-1}(\alpha/2) < \frac{\sqrt{n}(\bar{X}-\mu)}{S} < -t_{n-1}^{-1}(\alpha/2)\} = 1-\alpha$

So, rearranging we have, $P\{\bar{X} + t_{n-1}^{-1}(\alpha/2)\, S/\sqrt{n} < \mu < \bar{X} - t_{n-1}^{-1}(\alpha/2)\, S/\sqrt{n}\} = 1-\alpha$

Statistics---confidence intervals, case 2

We have an estimator for the mean, $\bar{X}$, which has a normal distribution with mean μ and variance σ^2/n, but we do not know σ^2. We do know, though, that $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$.

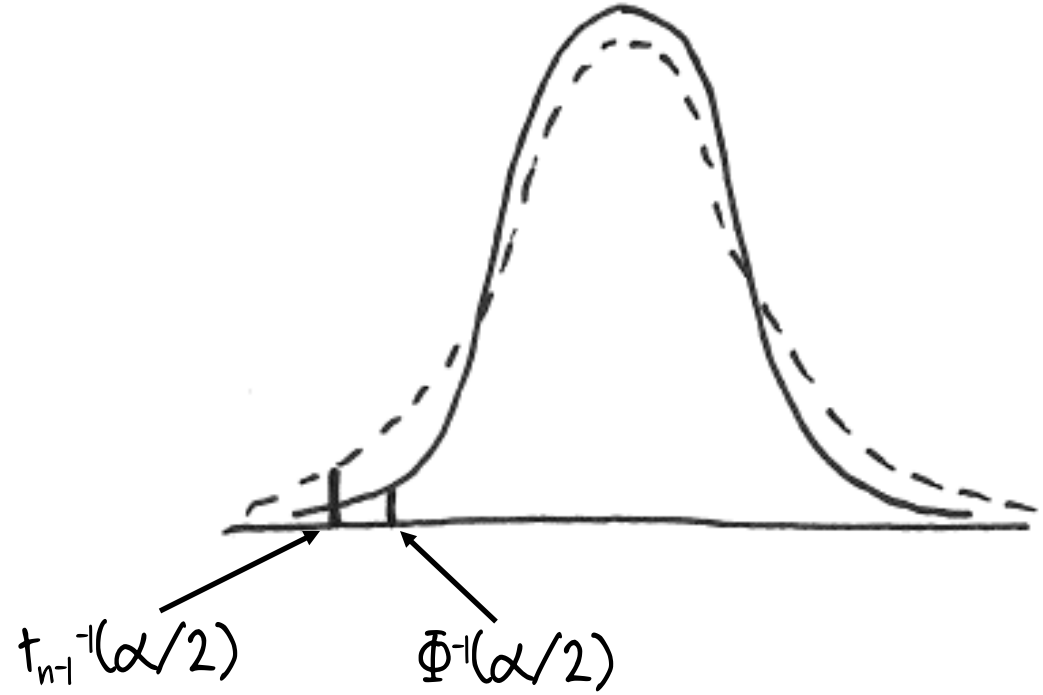
So $P\{t_{n-1}^{-1}(\alpha/2) < \frac{\sqrt{n}(\bar{X}-\mu)}{S} < -t_{n-1}^{-1}(\alpha/2)\} = 1-\alpha$

So, rearranging we have, $P\{\bar{X} + t_{n-1}^{-1}(\alpha/2)\,S/\sqrt{n} < \mu < \bar{X} - t_{n-1}^{-1}(\alpha/2)\,S/\sqrt{n}\} = 1-\alpha$

Our $1-\alpha$ CI is $[\bar{X} + t_{n-1}^{-1}(\alpha/2)\,S/\sqrt{n},\ \bar{X} - t_{n-1}^{-1}(\alpha/2)\,S/\sqrt{n}]$

Statistics---confidence intervals

Comparison of cases 1 and 2:



The t distribution is similar to the normal but has "fatter tails." It converges to the normal as $n \to \infty$.

Statistics---confidence intervals

Comparison of cases 1 and 2:

$-t_{n-1}^{-1}(\alpha/2) \to \Phi^{-1}(\alpha/2)$ as $n \to \infty$

The t gives you a wider interval than the normal for finite n. The intuition is that you are less sure of the distribution of your estimator because you don't know $\mathrm{Var}(\overline{X})$ and must estimate it. The t "penalizes" your confidence interval by making it wider, reflecting your greater uncertainty. As n goes to infinity, your uncertainty becomes relatively less important.

Statistics---confidence intervals

We don't always fall into case 1 or case 2. What do we do then?

Using facts that you know about how functions of random variables are distributed, you can construct a confidence interval "from scratch" on your own.

In practice, we usually appeal to CLT-like results to argue that the estimator has an approximate normal distribution, and then just use the t confidence interval formula with an estimated variance. (For large n, the t and normal confidence intervals are the same.)

Statistics---hypothesis testing

Well, now we know what an estimator is, how to estimate unknown parameters, and a couple of different ways to express how confident we are in our estimates. That gets us a long way and gives us a very good foundation for studying all kinds of estimation going forward.

One more foundational bit is quite important: hypothesis testing.

Statistics---hypothesis testing

In social science (as well as other settings for data analysis), we often encounter questions that we want to answer. (Some of you have started formulating them for your empirical project.) Do the lifespans of popes follow a lognormal distribution? Does the income tax rate affect the number of hours employees are willing to work? Do used books cost more on the internet than they do in brick and mortar stores? Has NAFTA hurt US manufacturing workers?

Statistics---hypothesis testing

The tool that statisticians have invented to help answer such questions (and quantify how confident we are in the answers) is the hypothesis test.

Purpose: Given a random sample from a population, is there enough evidence to contradict some assertion about the population?

Statistics---hypothesis testing

The tool that statisticians have invented to help answer such questions (and quantify how confident we are in the answers) is the hypothesis test.

Purpose: Given a random sample from a population, is there enough evidence to contradict some assertion about the population?

Let's build the structure underlying the hypothesis test.

Statistics---hypothesis testing

First, we'll need a bunch of definitions:

An <u>hypothesis</u> is an assumption about the distribution of a random variable in a population.

A <u>maintained hypothesis</u> is one that cannot or will not be tested.

A <u>testable hypothesis</u> is one that can be tested using evidence from a random sample.

The <u>null hypothesis</u>, H_0, is the one that will be tested.

The <u>alternative hypothesis</u>, H_A, is a possibility (or series of possibilities) other than the null.

Statistics---hypothesis testing

For instance, we might want to perform a test concerning unknown parameter θ where $X_i \sim f(x|\theta)$.

H_0: θ in Θ_0

H_A: θ in Θ_A, where Θ_0 and Θ_A disjoint.

More definitions:

A <u>simple hypothesis</u> is one characterized by a single point, i.e., $\Theta_0 = \theta_0$.

A <u>composite hypothesis</u> is one characterized by multiple points, i.e., Θ_0 is multiple values or a range of values.

Statistics---hypothesis testing

For instance, we might want to perform a test concerning unknown parameter θ where $X_i \sim f(x|\theta)$.

H_0: θ in Θ_0

H_A: θ in Θ_A, where Θ_0 and Θ_A disjoint.

More definitions:

A <u>simple hypothesis</u> is one characterized by a single point, i.e., $\Theta_0 = \theta_0$.

A <u>composite hypothesis</u> is one characterized by multiple points, i.e., Θ_0 is multiple values or a range of values.

usual set-up: simple null and composite alternative

Statistics---example set-up

X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known.

Interested in testing whether $\mu = 0$.

H_0: $\mu = 0$

H_A: $\mu = 1$

Statistics---example set-up

X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known. maintained hypotheses

Interested in testing whether $\mu = 0$. testable hypothesis

H_0: $\mu = 0$ null hypothesis, simple

H_A: $\mu = 1$ alternative hypothesis, simple

Statistics---example set-up

X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known. maintained hypotheses

Interested in testing whether $\mu = 0$. testable hypothesis

H_0: $\mu = 0$ null hypothesis, simple

H_A: $\mu \neq 0$

Statistics---example set-up

X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known. maintained hypotheses

Interested in testing whether $\mu = 0$. testable hypothesis

H_0: $\mu = 0$ null hypothesis, simple

H_A: $\mu \neq 0$ alternative hypothesis, composite, two-sided

Statistics---example set-up

X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known. maintained hypotheses

Interested in testing whether $\mu = 0$. testable hypothesis

H_0: $\mu = 0$ null hypothesis, simple

H_A: $\mu > 0$

Statistics---example set-up

X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known. maintained hypotheses

Interested in testing whether $\mu = 0$. testable hypothesis

H_0: $\mu = 0$ null hypothesis, simple

H_A: $\mu > 0$ alternative hypothesis, composite, one-sided

Statistics---example set-up

X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known.

Interested in testing whether $\mu = 0$.

H_0: $\mu = 0$

H_A: $\mu = 1$

We then either "accept" or "reject" the null hypothesis based on evidence from our sample.

Statistics---hypothesis testing

Obviously, mistakes can be made--we can "reject" a null that is true or "accept" a null that is false. We want to set up our hypothesis test to analyze and control these errors. We first need a taxonomy.

	H_0 true	H_0 false
accept H_0	No error	Type II error
reject H_0	Type I error	No error

Statistics---hypothesis testing

	H_0 true	H_0 false
accept H_0	No error	Type II error
reject H_0	Type I error	No error

The <u>significance level</u> of the test, α, is the probability of type I error.

The <u>operating characteristic</u> of the test, β, is the probability of type II error.

We call $1-\alpha$ the <u>confidence level</u>. We call $1-\beta$ the <u>power</u>.

Statistics---hypothesis testing

Finally, we define the <u>critical region</u> of the test, C or C_X, as the region of the support of the random sample for which we reject the null.

Statistics---example

X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known.

H_0: $\mu = 0$

H_A: $\mu = 1$

Suppose, first, that $n = 2$.

Think about what kind of sample would lead you to believe the null or doubt the null in favor of the alternative.

Statistics---example

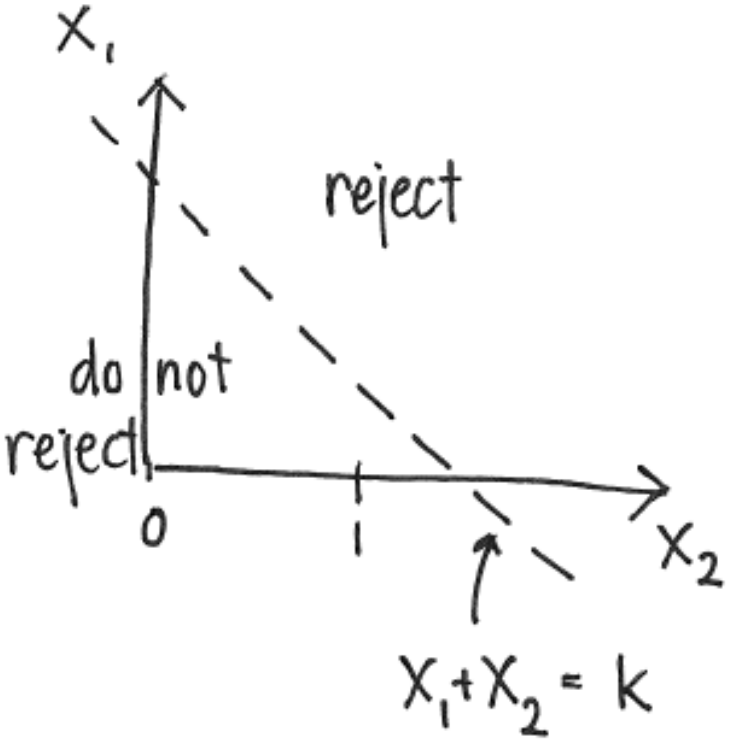
X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known.

H_0: $\mu = 0$

H_A: $\mu = 1$

Suppose, first, that $n = 2$.

Think about what kind of sample would lead you to believe the null or doubt the null in favor of the alternative.



Statistics---example

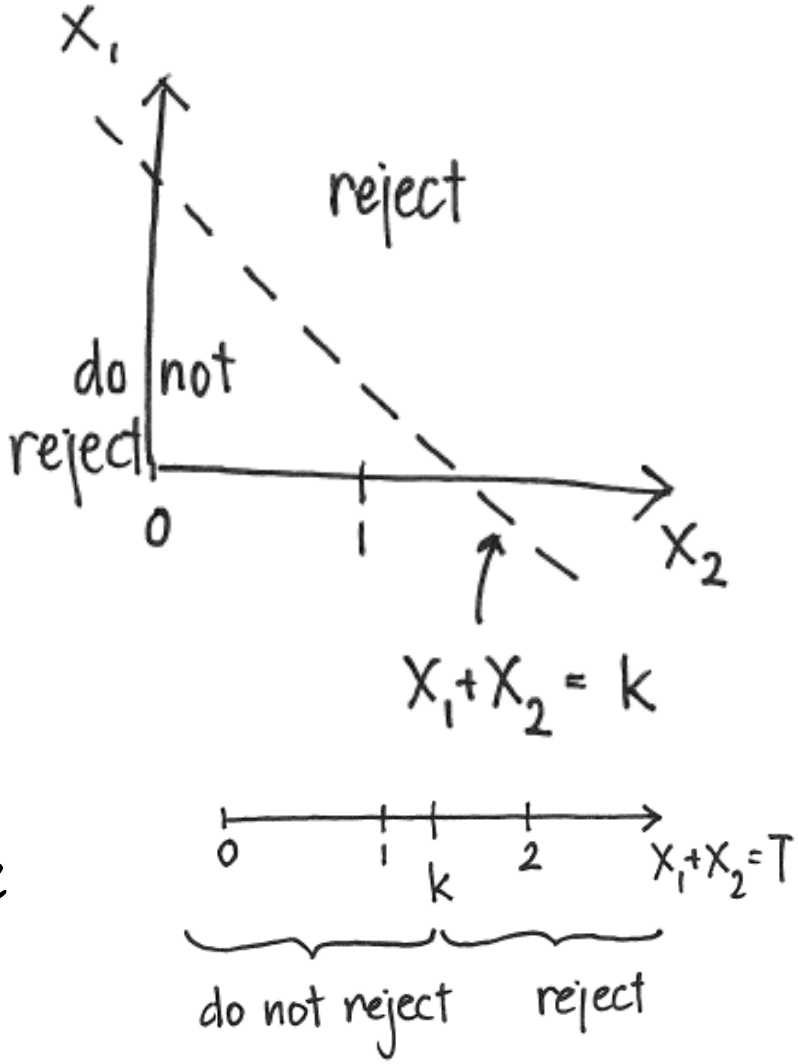
X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known.

H_0: $\mu = 0$

H_A: $\mu = 1$

Suppose, first, that $n = 2$.

Think about what kind of sample would lead you to believe the null or doubt the null in favor of the alternative.



Statistics---example

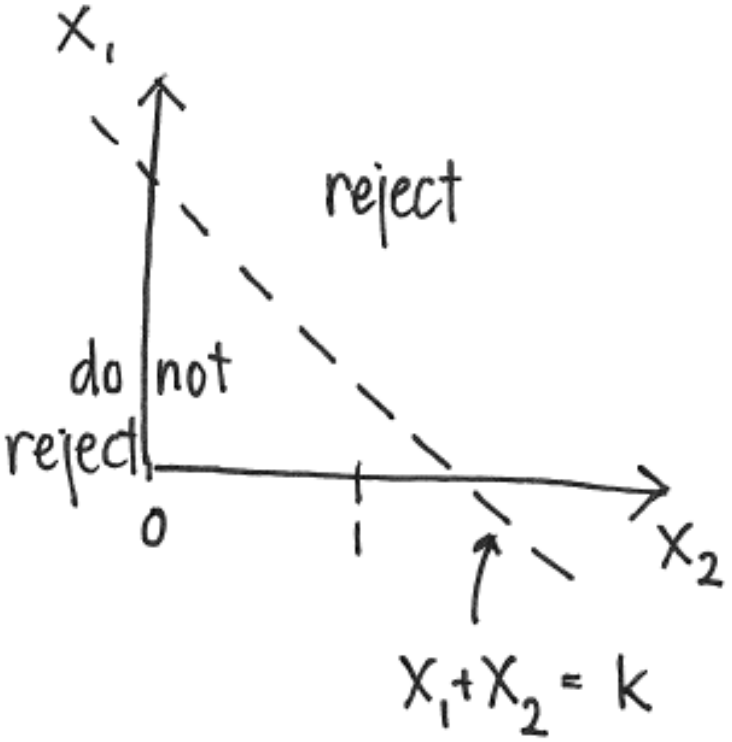
X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known.

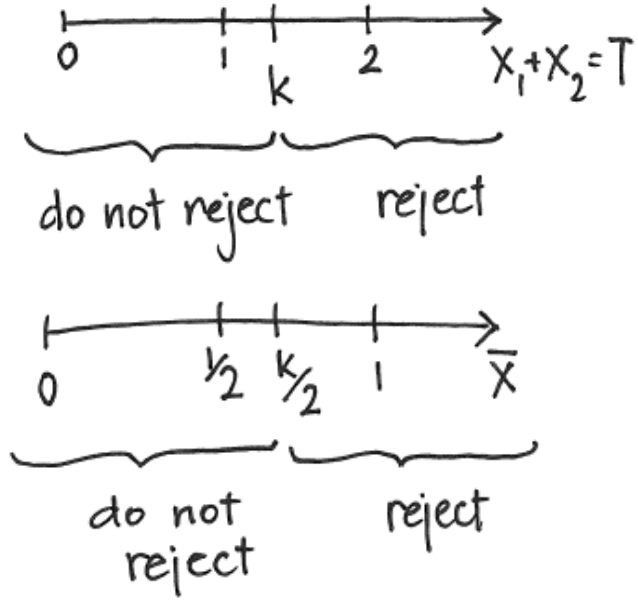
H_0: $\mu = 0$

H_A: $\mu = 1$

Suppose, first, that $n = 2$.

Think about what kind of sample would lead you to believe the null or doubt the null in favor of the alternative.





Statistics---example

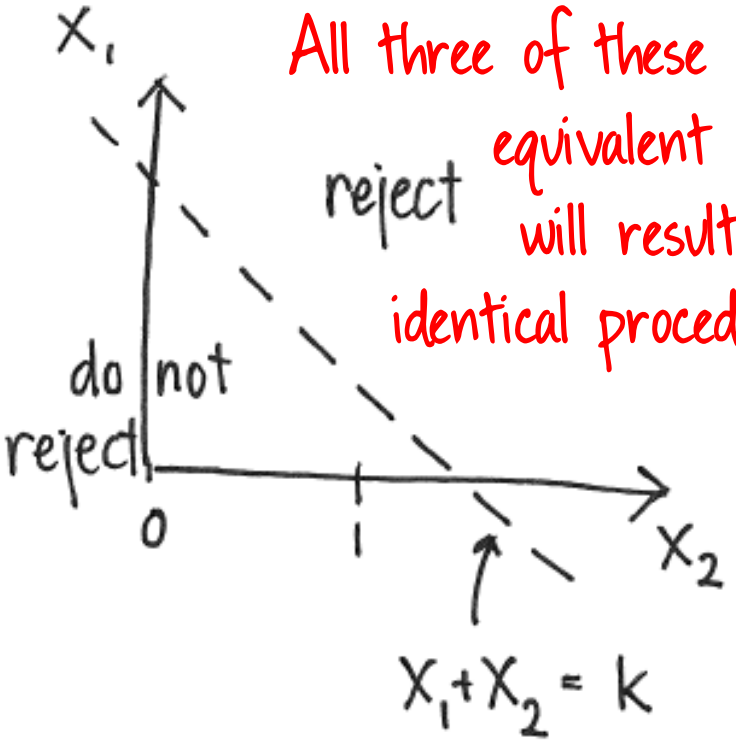
X_i i.i.d. $N(\mu, \sigma^2)$, where σ^2 known.

H_0: $\mu = 0$

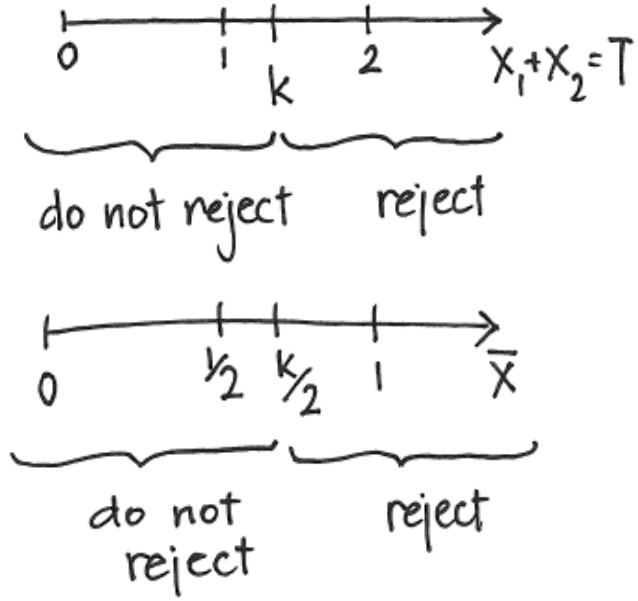
H_A: $\mu = 1$

Suppose, first, that $n = 2$.

Think about what kind of sample would lead you to believe the null or doubt the null in favor of the alternative.



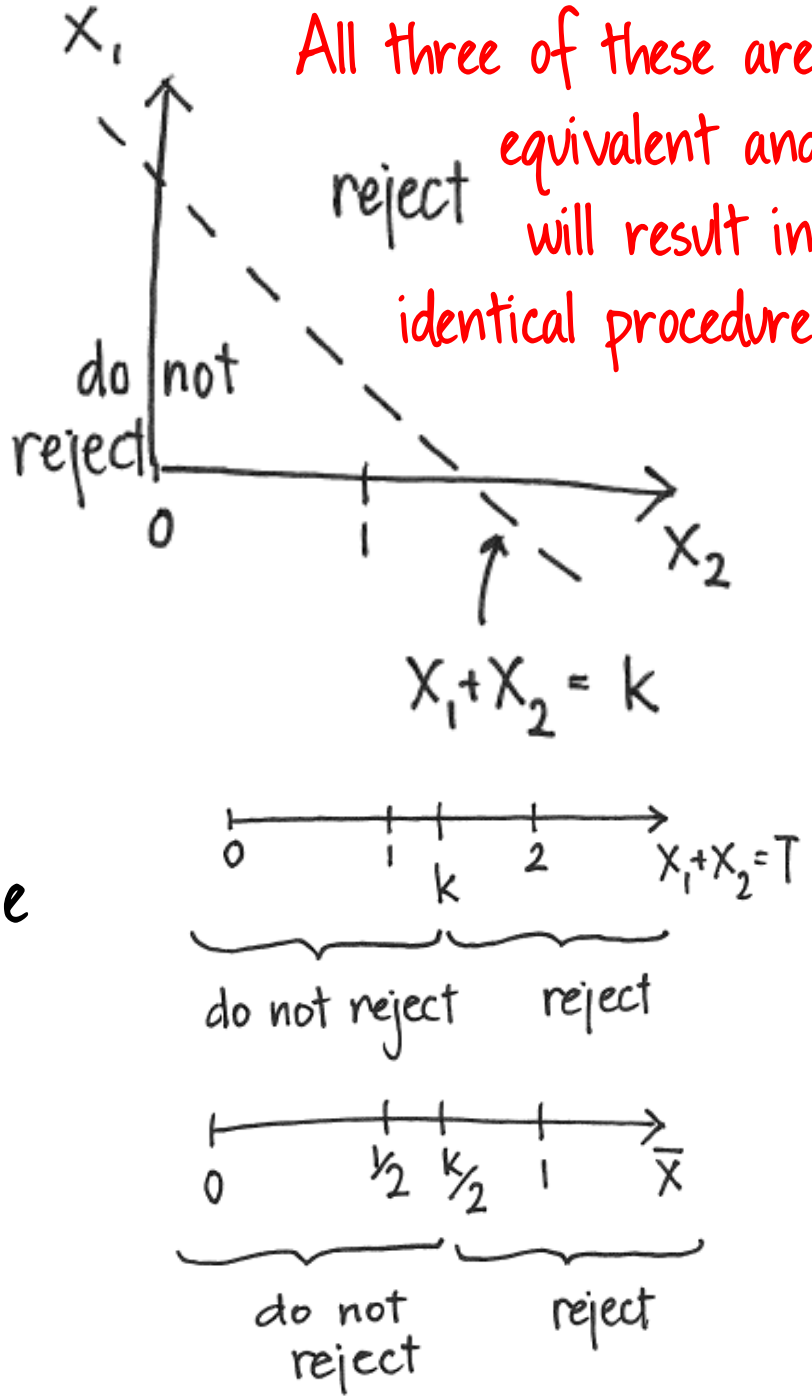
All three of these are equivalent and will result in identical procedures.



Statistics---example

So do we prefer one over the others? Not really, but when n gets big, we don't want to have to worry about n-dimensional spaces, so we would just as soon base the test on the sum or the sample mean.

All three of these are equivalent and will result in identical procedures.



Statistics---example

So we'll base our testing procedure on the <u>test statistic</u>

$T = \bar{X}$ and reject for "large" values.

In other words, critical region C will take the form $\bar{X} > k$

for some k yet to be determined.

Statistics---example

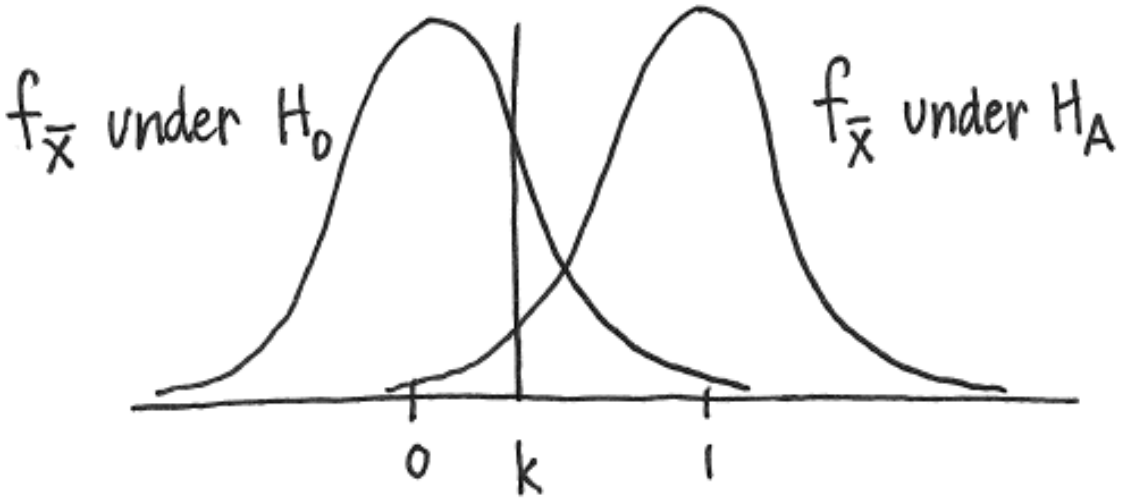
So we'll base our testing procedure on the <u>test statistic</u>

$T = \bar{X}$ and reject for "large" values.

In other words, critical region C will take the form $\bar{X} > k$

for some k yet to be determined.

How do we choose k? Trade off two types of error.



Statistics---example

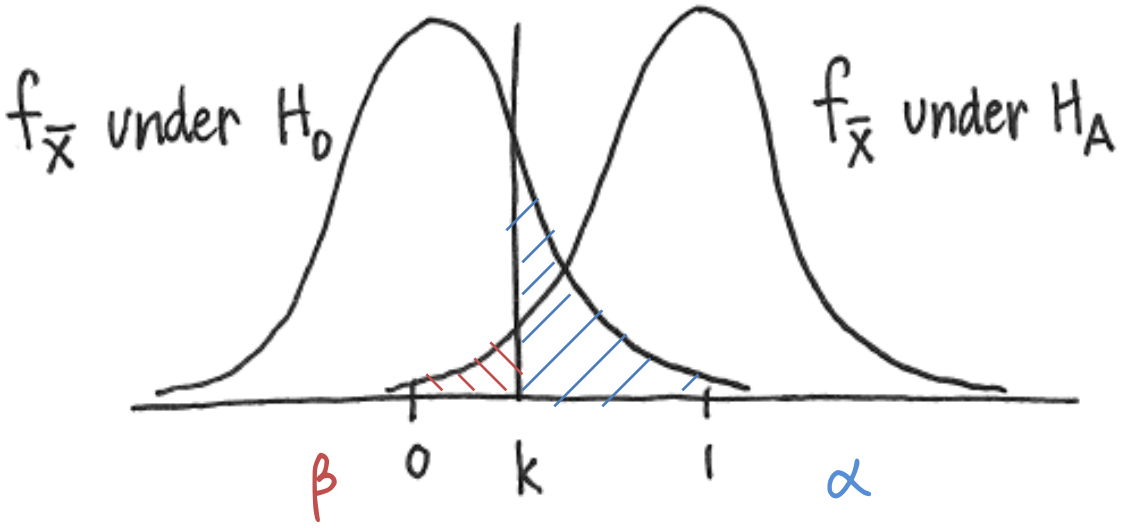
So we'll base our testing procedure on the <u>test statistic</u>

$T = \bar{X}$ and reject for "large" values.

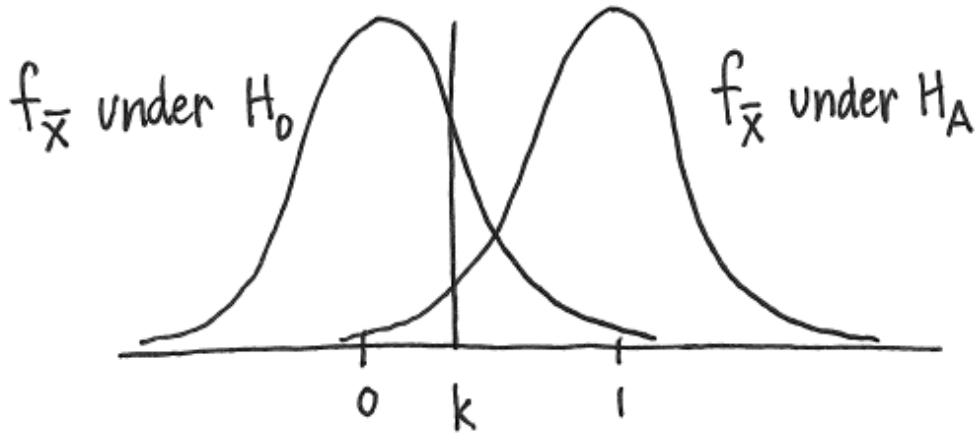
In other words, critical region C will take the form $\bar{X} > k$

for some k yet to be determined.

How do we choose k? Trade off two types of error.



Statistics---example

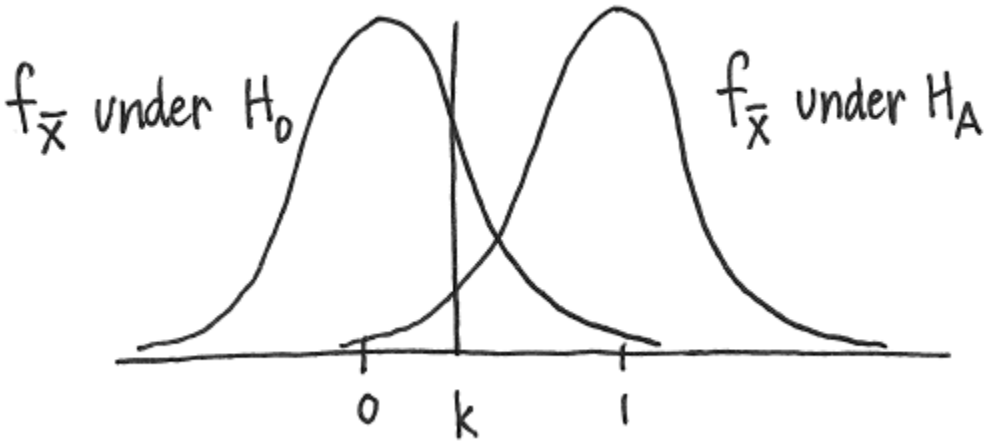


Choice of any one of α, β, or k determines the other two. Furthermore, choosing them involves an explicit trade-off between the probability of type I and type II errors.

increasing k means $\alpha \downarrow$ and $\beta \uparrow$

decreasing k means $\alpha \uparrow$ and $\beta \downarrow$

Statistics---example



Let's compute α and β using some specific numbers. If $\sigma^2 = 4$ and $n = 25$, then $T \sim N(0, 4/25)$ under H_0 and $N(1, 4/25)$ under H_A.

$\alpha = P(T > k | \mu = 0) = 1 - \Phi((k-0)/(2/5))$

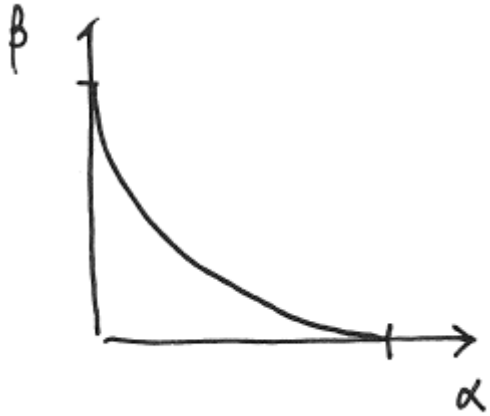
$\beta = P(T < k | \mu = 1) = \Phi((k-1)/(2/5))$

Statistics---example

$$\alpha = P(T>k|\mu=0) = 1 - \Phi((k-0)/(2/5))$$

$$\beta = P(T<k|\mu=1) = \Phi((k-1)/(2/5))$$

If you plugged in different values of k, you would get a graph in α-β space that looked like this:

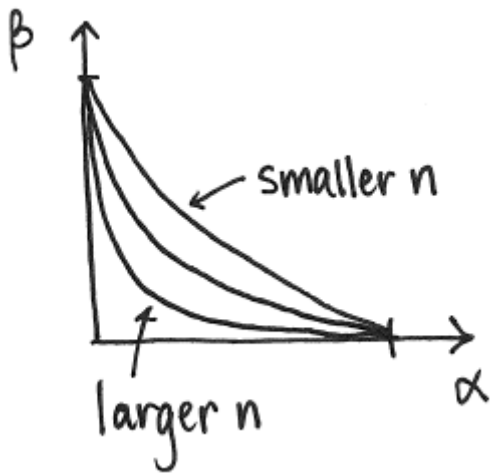


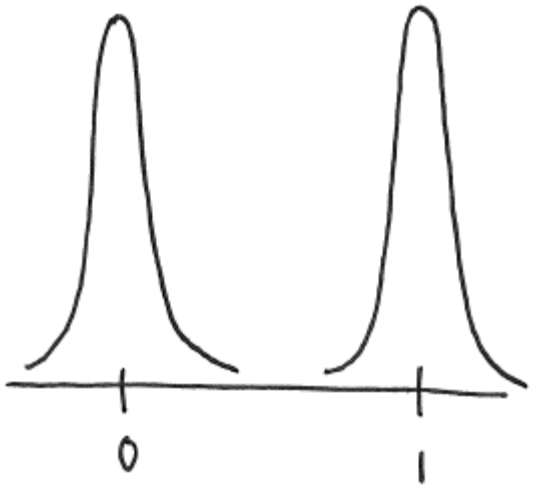
Statistics---example

$$\alpha = P(T>k|\mu=0) = 1 - \Phi((k-0)/(2/5))$$

$$\beta = P(T<k|\mu=1) = \Phi((k-1)/(2/5))$$

What happens as n increases or decreases?





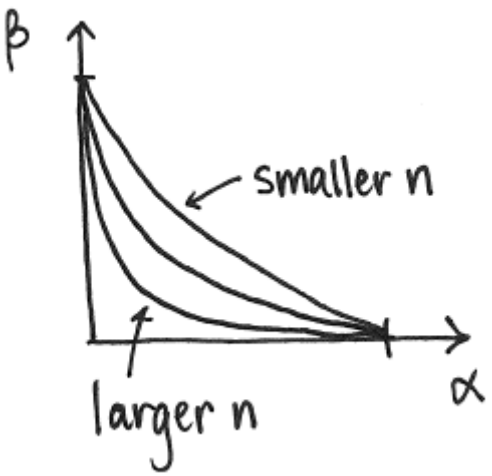
Statistics---example

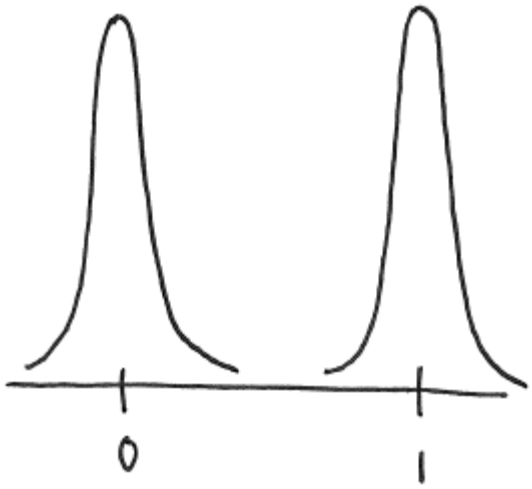
$$\alpha = P(T>k|\mu=0) = 1 - \Phi((k-0)/(2/5))$$

$$\beta = P(T<k|\mu=1) = \Phi((k-1)/(2/5))$$

this changes

What happens as n increases or decreases?





Statistics---example

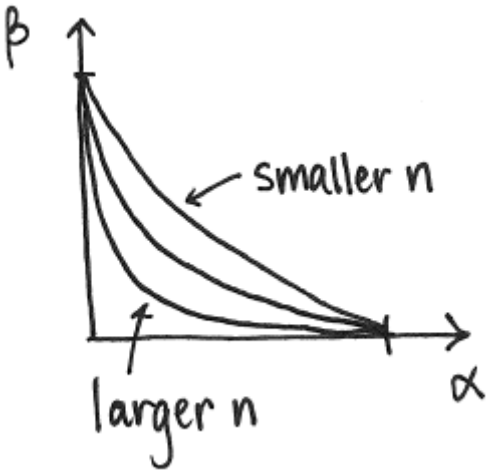
As n increases, the two distributions get tighter around their means---can do better on both α and β.

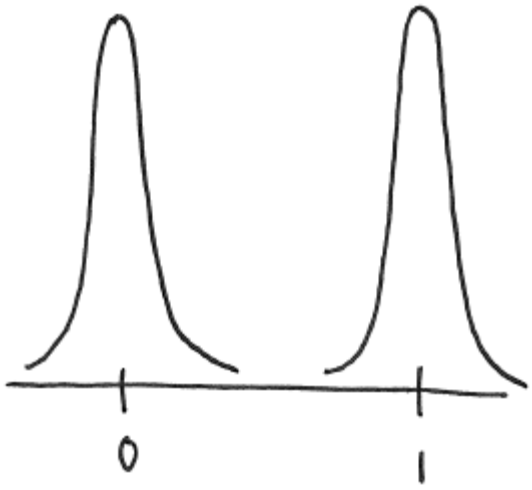
$$\alpha = P(T>k|\mu=0) = 1 - \Phi((k-0)/(2/5))$$

$$\beta = P(T<k|\mu=1) = \Phi((k-1)/(2/5))$$

this changes

What happens as n increases or decreases?





Statistics---hypothesis testing

Notes:

How do we know which hypothesis should be the null and which should be the alternative? Well, we are free to choose. We often choose so that the type I error is the more serious of the errors. Then we will choose k so that α is at an acceptably low level, such as .05 or .01. (Of course, if n is really large and we keep α fixed, then β might be very small, which might not be what we want.)

Statistics---hypothesis testing

Notes:

What if μ is not either 0 or 1? Well, much of the time we set up a hypothesis test so that $\Theta_0 \cup \Theta_A$ is the entire parameter space. So either the null or the alternative must be true. That means that one or both of the hypotheses are composite. When we have composite hypotheses, the test becomes more difficult to analyze. In particular, α and β may no longer be values but could be functions of the unknown parameter(s) θ.

Statistics---hypothesis testing

Notes:

What if our hypotheses were

H_0: $\mu = 0$

H_A: $\mu \neq 0$?

We could use the same test statistic $\bar{X}$, but what should the critical region look like?

Power calculations

- For a sample of size N, we will observe $W_1 \ldots W_N$, and $Y_1^{obs} \ldots Y_N^{obs}$
- Suppose we are interested in testing: $H_o = E[Y_i(1) - Y_i(0)] = 0$ against $H_a : E[Y_i(1) - Y_i(0)] \neq 0$
- A reminder:

	H_o true	H_o false
accept H_o	No error	Type II error
reject H_o	Type I error	No error

The significance level of the test, α, is the probability of type I error.

The operating characteristic of the test, β, is the probability of type II error.

We call $1-\alpha$ the confidence level. We call $1-\beta$ the power.

What ingredients goes into the power calculation?

- We tend to pick α low because society does not want to conclude that some treatment work when it fact it really does not.
- Following Fisher, it is often $\alpha = 0.05$
- We want to pick $N = N_c + N_t$ such that , if the average treatment effect is in fact some value τ, the power of the test will be at least $1 - \beta$ for some β, given that a fraction γ of the units are assigned to the treatment group.
- In addition we must assume (know) something about the variance of the outcome in each treatment arm: for simplicity we often assume it is the same, and some parameter σ^2.
- In summary we know, impose, or assume $\alpha, \beta, \tau, \sigma$, and γ, and we are looking for N.
- Alternatively, we could be interested in the power for a given sample size: we know $\alpha, \gamma, \tau, \sigma$, and N and look for —β.

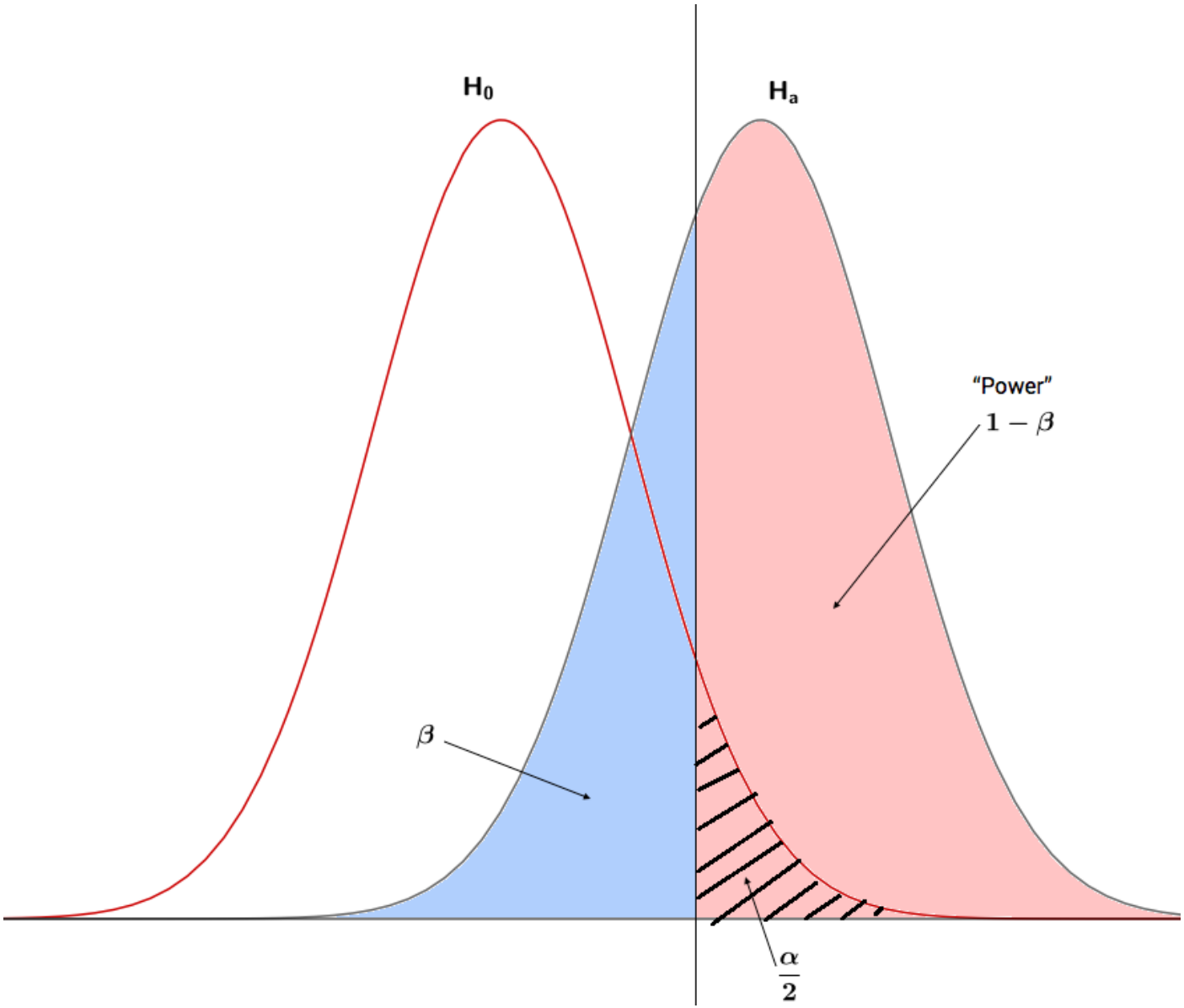
Guess work

- α and β are imposed and we can decide γ (if this was just power what would we pick?)
- Problem: how do we know/determine τ and σ ?
  - τ: could be known from a pilot, from a previous study, or could be picked as a value of interest.
    - For example: the lowest value such that, if we could reject zero when the effect is really τ, the program would be worth doing.
    - This is more about optics than about statistics... (rejecting zero is not "accepting" the point estimate...)
    - But it has the merit to remind us that we may be interesting in 'detecting' a small effect, we will work with large sample. If the program is very expensive such that it won't be adopted unless the effects are very large anyway, we can go with a smaller sample.
  - σ: Need to get that from prior data, with similar outcomes.
  - Some item it is wide guess work!

Now for the formulas

- This is of course in practice the easy part: many software will give you power curves as you tinker with the parameters and the sample size.
- But it is worth working through the logic.
- $T = \frac{\overline{Y_t^{obs}} - \overline{Y_c^{obs}}}{\sqrt{\hat{V}_{Neyman}}} \approx \frac{\overline{Y_t^{obs}} - \overline{Y_c^{obs}}}{\sqrt{\frac{\sigma^2}{N_t}} + \sqrt{\frac{\sigma^2}{N_c}}}$
- We reject this hypothesis if $|T| > \Phi(1 - \frac{\alpha}{2})$, e.g. if $\alpha = 0.05$, if $|T| > 1.96$
- What is the probability that this occurs?
- By the central limit theorem, the difference in means minus the true treatment effect, scaled by the true standard error of that difference, has distribution that is approximately $N(0, 1)$:
- $\frac{\overline{Y_t^{obs}} - \overline{Y_c^{obs}} - \tau}{\sqrt{\frac{\sigma^2}{N_t} + \frac{\sigma^2}{N_c}}} \approx \mathcal{N}(0, 1)$ and hence $T \approx \mathcal{N}(\frac{\tau}{\sqrt{\frac{\sigma^2}{N_t} + \frac{\sigma^2}{N_c}}}, 1)$

Statistical Power



So

$$P\big(|T| > \Phi(1-\frac{\alpha}{2})\big) \approx \Phi\big((-\Phi^{-1}(1-\frac{\alpha}{2}) + \frac{\tau}{\sqrt{\frac{\sigma^2}{N_t}+\frac{\sigma^2}{N_c}}}\big) +$$

$$\Phi\big(-\Phi^{-1}(1-\frac{\alpha}{2}) - \frac{\tau}{\sqrt{\frac{\sigma^2}{N_t}+\frac{\sigma^2}{N_c}}}\big)$$

The second term is very small, so we ignore it.
So we want the first term to be equal to $1-\beta$, which requires:

$$\Phi^{-1}(1-\beta) = -\Phi^{-1}(1-\frac{\alpha}{2}) + \frac{\tau\sqrt{N}\sqrt{\gamma(1-\gamma)}}{\sigma}$$

Which leads to the required sample size:

$$N = \frac{(\Phi^{-1}(1-\beta) + \Phi^{-1}(1-\frac{\alpha}{2}))^2}{\frac{\tau^2}{\sigma^2}.\gamma.(1-\gamma)}$$

Other considerations to take into account when you do power calculations

- If you have stratified or not: with stratified design, variance of estimated treatment effect is lower.
- If you have clustered or not: with clustered design, variance of estimated treatment effect is larger

Analysis of a stratified design

- Take the difference in means within each strata
- Take a weighted average of the treatment effect with the weight the size of the strata $\sum_g (\frac{N_g}{N}) \widehat{\tau_g}$
- This will be an unbiased estimate of the average treatment effect
- And the variance will be calculated as $\sum_g (\frac{N_g}{N})^2 \widehat{V_g}$
- Special case: probability of assignment to control group stays the same in each strata. Then this coefficient is equal to the simple difference between treatment and control, but the variance is always weakly lower.
- Stratification will lower the required sample size for a given power.

Analysis of a clustered design

- The opposite happens with a clustered design (all the unit within a same unit are either treated or control).
- We need to take into account the fact that the potential outcomes for units within a randomization clusters are not independent.
- Conservative way to do this: just average the outcome by unit, and treat each as an observation (like we did for classrooms in the Duflo-Hanna data.
- Then the number of observations is the number of clusters, and you can analyze this data exactly as a completely randomized experiment but with clusters as the unit of analysis.
- For example, this tells you that a randomization with two clusters is unlikely to go very far!!

References

- Imbens and Rubin *Causal Inference for Statistics Social and biomedical Sciences*
- Duflo, Hanna, Ryan "Incentives work". American Economic Review