L17: Recap of Part II

Rules of probability

Review of probability rules

Probabilities are numbers between 0 and 1.

$0 \leq P(A) \leq 1$

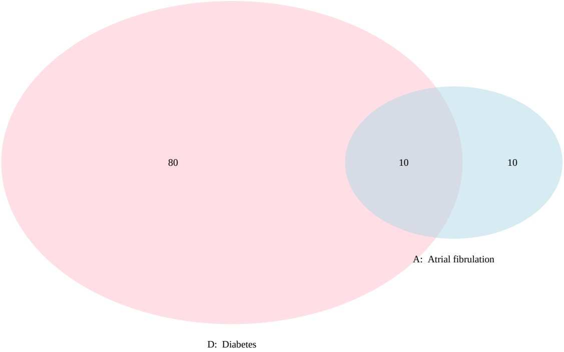
The probabilities in the probability space must sum to 1.

The probabilities of an event and it's complement must sum to 1

$P(A) + P(\bar{A}) = 1$

Ven diagrams

If there are 180 total people in this study, what is the number missing from our parameter space?



Adding and decomposing probability

For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

So what is the union of Atrial fibrillation and Diabetes in this example $P(A \cup D)$?

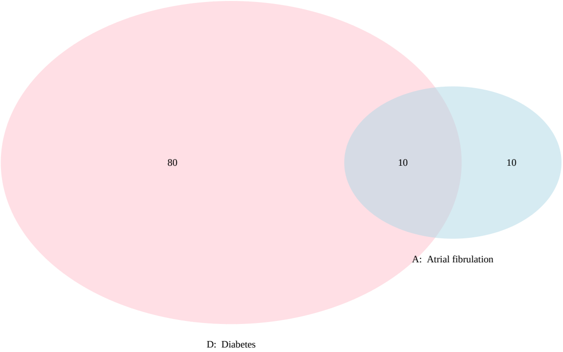
Adding and decomposing probability

For any two events A and B, $P(A) = P(A \cap B) + P(A \cap \bar{B})$

What would this look like in our example?

Ven diagram

There are 180 total people in this study, the intersect here is not included in the other pieces of the diagram.



Rules for independence

Written out in probability notation, for any two events A and B, the events are independent if:

$$P(A|B) = P(A)$$

or

$$P(B|A) = P(B)$$

or

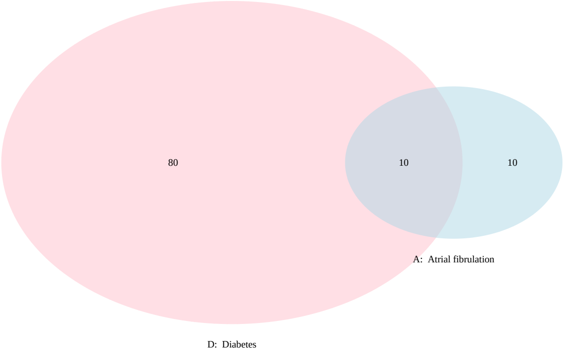
$$P(A \cap B) = P(A) * P(B)$$

Rules for independence

In our example is Atrial fibrillation independent of Diabetes?

Ven diagram

There are 180 total people in this study, the intersect here is not included in the other pieces of the diagram.



Multiplication rule and conditional probability

For any two events, the probability that both events occur is given by:

$$P(A \cap B) = P(B|A) \times P(A)$$

When $P(A) > 0$, the conditional probability of B, given A is:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Bayes' theorem (simple version)

Suppose that A and A^c are disjoint events whose probabilities are not 0 and add exactly to 1. That is, any outcome has to be exactly in one of these events. Then if B is any other event whose probability is not 0 or 1,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Example calculations

Let's say a group of 10 friends are deciding between Majors at Berkeley.

If I choose a science major, the probability of early morning lectures is .6

In this group 6 students choose science majors and 4 students choose language/literature related majors

I randomly choose one student to interview

The probability that the student I talk to has chosen a language/literature related major and has early morning lectures is 0.2.

What is the probability that students in language/literature related majors have early morning lectures?

Example calculations

From the information given I can know the following:

P(Science major)= .6 P(Language/literature major)=.4

P(Early lectures $\mid$ Science major)=.6

P(Early lectures $\cap$ Language/literature major)=.2

Example calculations

Putting the known information into a table we have this:

Early lectures	Science Majors	Language/Literature majors	Totals
Yes	3.6	2	
No			
Totals	6	4	10

and we can figure out that of the 4 Language/Literature majors, if 2 have early lectures, then 2 must not have early lectures, so 50% of those who are Langugage/Literature majors have early lectures.

Example calculations - tree

Conditional probabilities of Screening tests

Test result	Samples with known Disease	Samples without Disease	Totals
Positive	90	8	98
Negative	14	96	110
Totals	104	104	208

Two characteristics that are conditional on true disease status

- Sensitivity = P(Test positive | Disease)
- Specificity = P(Test negative | No Disease)

Two characteristics that are conditional on test result

- Predictive value positive = P(Disease | Test positive)
- Predictive value negative = P(Not disease | Test negative)

Conditional probabilities of Screening tests

What happens to sensitivity if we are in a context where the disease is more prevalent?

What happens to predictive value positive?

Distributions

Distribution	Defined by:	Type of outcome	R notation
Normal	Mean and SD	Continuous	norm
Binomial	number and p	Binary (success or failure in n trials)	binom
Poisson	mean (λ)	Discrete count of events in an interval	pois

You should be familiar with what these distributions look like and what changes in the shape of the distribution as the key parameters change.

Which distribution?

What distribution would you think of for the following studies?

- Tracking the incidence of influenza during the weeks of winter
- Estimating the proportion of male and female children in a school who missed at least one day of school due to flu
- Estimating the number of minutes of exercise among students before and after new years day

Calculations with the normal distribution

What proportion of adult women in the United States are taller than Beyonce?

In the US the mean height is 5’5” with a sd of 3.5”

Beyonce is 5’7” tall.

In R?

```
#code to calculate - fill in during class

#_norm(____, _____, ____, option)
```

calculations with the normal distribution

What is the Z- value for Beyonce's height?

$$Z = \frac{x - \mu}{sd}$$

calculations with the normal distribution

How would we use the Z value to calculate the probability of being shorter than Beyonce?

```
#code to calculate taller than Beyonce using measured height
pnorm(q=67, mean=65, sd=3.5, lower.tail=F)
```

```
## [1] 0.2838546
```

```
# code to calculate shorter using Z value
# to do during class
```

calculations with the normal distribution

How many women are taller than Ariana Grande (5' 0") and shorter than Beyonce?

```
#code to calculate taller than Beyonce using measured height
pnorm(q=67, mean=65, sd=3.5, lower.tail=F)
```

```
## [1] 0.2838546
```

```
# code to calculate proportion in range
# to do during class
```

Calculations with binomial

Imagine you are working at an aquarium shop. You have a tank with 600 guppies, 30% of which have black spots on their tail.

You have a client who wants to take home 4 guppies, 2 with black spots and 2 without black spots.

You can net 4 fish at a time. What is the probability of netting the fish your client wants in any attempt?

Calculations with binomial

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$\binom{n}{x} = \frac{n!}{x!(n - x)!}$$

Normal approximation for binomial distributions

Suppose that a count X has the binomial distribution with n observations and success probability p. When n is large, the distribution of X is approximately Normal. That is,

$$X \dot{\sim} N(\mu = np, \sigma = \sqrt{np(1-p)})$$

As a general rule, we will use the Normal approximation when n is so large that $np \geq 10$ and $n(1-p) \geq 10$.

It is most accurate for p close to 0.5, and least accurate for p closer to 0 or 1.

Calculations with poisson

If X has the Poisson distribution with mean number of occurrences per interval μ, the possible values of X are 0, 1, 2, ...If k is any one of these values, then

$$P(X = k) = \frac{e^{-\mu}\mu^k}{k!}$$

Calculations with poisson

The rate of measles in California is roughly 1.75 cases per month, usually from travelers exposed while outside of the country. Between December 2014 and April 2015 the rate was roughly 26.2 cases per month.

What is the probability of observing exactly 2 cases in a normal month? (worked out by hand, then confirm with r)

```
##fill in during class
```

```
##fill in during class
```

Calculations with poisson

What is the probability of observing 0,1 or 2 cases in a normal month? (there is more than one way to do this)

Calculations with poisson

```
dpois(0,1.75)+dpois(1,1.75)+dpois(2,1.75)
```

```
## [1] 0.7439697
```

```
ppois(2,1.75)
```

```
## [1] 0.7439697
```

```
1-ppois(25,1.75)
```

```
## [1] 0
```

Calculations with poisson

What is the probability of observing 26 cases or more in a normal month? Would you feel comfortable calling this an outbreak?

```
## fill in during class
```

Parameter and statistic

μ and p are population parameters for the mean and proportion. There is one unique value for μ and p in the underlying population.

$\bar{x}$ and $\hat{p}$ are statistics computed using samples. We refer to them as the sample mean and sample proportion, respectively. If we change the sample our statistics will likely also change. Statistics vary across samples.

Sampling distribution of a sample mean for a Normal population

- If individual observations have a $N(\mu, \sigma)$ distribution, then the sample mean $\bar{x}$ of a simple random sample of size n has a $N(\mu, \frac{\sigma}{\sqrt{n}})$

You should be able to think through what happens when we adjust parts of this equation. (ie what happens to variability of the sample mean when we increase n?)

We can use the Central Limit Theorem to treat the distribution of a sample mean as normally distributed under conditions when the underlying population values are not normally distributed.

The Central Limit Theorem (CLT)

Draw a simple random sample of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \dot{\sim} N(\mu, \frac{\sigma}{\sqrt{n}})$$

The CLT allows us to use Normal probability calculations to answer questions about sample means from many observations (questions relying on the sampling distribution of the sample mean) even when the population distribution is not Normal.

Sampling distribution of the proportion $\hat{p}$

- The mean of the sampling distribution is p, the population parameter
- The standard deviation of the sampling distribution is $\sqrt{\frac{p(1-p)}{n}}$
- As the sample size increases, the sampling distribution of $\hat{p}$ becomes approximately Normal. This is the Central Limit Theorem for a proportion!
- For this to apply, we require:
    - the population is at least 20 times as large as the sample
    - both np and n(1-p) are larger than 10.

Developing inference

We can use information about the variability of sample means to generate confidence intervals and p values for our estimates and begin to use this information to draw inference from our data.

Confidence intervals for the mean μ

Confidence level C	90%	95%	99%
Critical value z*	1.645	1.960 ($\approx$ 2)	2.576

- These numbers correspond to the value on the x-axis corresponding to having 90%, 95%, or 99% of the area under the Normal density between -z and z.

The generic format of a confidence interval is then:

$$\bar{x} \pm z * \frac{\sigma}{\sqrt{n}}$$

You should know how to create a confidence interval and what changes to your study/data would cause the confidence interval to be larger or smaller.

Define the Hypothesis

A Null Hypothesis (H_0) is the hypothesis that is assumed to be true and the start of a test. This is often expressed as a statement of equality (ie. mean equal to a certain value or no difference between groups)

An Alternative Hypothesis (H_A) is usually the inverse of the null hypothesis and is expressed as a statement of difference.

- H_A: The mean is greater than the Null (one tailed)
- H_A: The mean is less than the null (one tailed)
- H_A: The mean is not equal to (greater or less than) the null (two tailed)

When we test a hypothesis, we are not trying to prove H_A, we are trying to disprove H_0

Decide on a threshold for rejecting the null

We choose a probability that we decide is small enough that we are unlikely to have observed it by chance if H_0 is true.

This threshold is our α.

We must decide if our hypothesis is one-tailed or two-tailed

You should be able to read a description of a study or hypothesis test and know whether they hypothesis test would be one tailed/one sided or two tailed/two sided.

P-value

P-value: The probability, assuming that H_0 is true, that the test statistic would take a value at least as extreme (in the direction of H_a) as that actually observed. The smaller the p-value, the stronger the evidence against H_0 provided by the data.

Type I error, and Type II error in hypothesis tests

You should know the difference between type I and type II error and what would cause error or power to increase or decrease

	H_a is true	H_0 is true
Reject H_0	Correct decision	Type I error (α)
Fail to reject H_0	Type II error (β)	Correct decision

power calculation example

Example

Here we will go through another example from the Baldi and Moore textbook. This example assumes you are planning a quality control study to look at whether storage impacts the perceived sweetness of a beverage. Ten professional tasters will rate the sweetness on a 10 point scale before and after storage. We know that the standard deviation of sweetness ratings is $= 1$. We also know that a mean sweetness change of 0.8 units on this scale is noticed by consumers. We want 90% power and an alpha of 0.05 for our study. We have a set of 10 values representing the difference in sweetness caused by storage. Let's presume that we only care about a perceptible *loss* in sweetness.

What is the null hypothesis here?

What is our alternative?

Is our hypothesis one or two sided?

Example

What is the null hypothesis here? no difference in sweetness or μ=0

What is our alternative? μ decreases by .8 or more

Is our hypothesis one or two sided? one sided (we only care about decreases)

Example

We will start by finding the Z alpha:

$$Z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

```
qnorm(.05)
```

```
## [1] -1.644854
```

$$-1.645 = \frac{\overline{x} - 0}{\frac{1}{\sqrt{10}}}$$

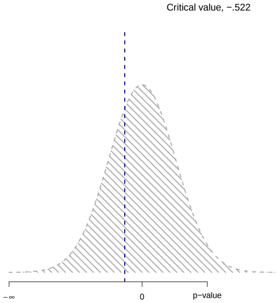
Solve this for $\overline{x}$

$$\bar{x} = -1.645 \times \frac{1}{\sqrt{10}} = -0.520$$

### alternate calculation Note that you could also calculate the cutpoint value

null distribution

So here we have our null distribution with the value at which we reject the null



What is our β ?

Example

We must choose a value at which to evaluate power. Here we will choose an alternate hypothesis that the mean sweetness difference is -0.8. Since we know a sample mean greater than -0.520 causes us to fail to reject H_0 we need to calculate the proportion of a distribution centered at -0.8 that would be below this value.

Example:

Using R to calculate the probability, relative to a sampling distribution centered at the alternative value -.8

```
pnorm(-.522, -.8, (1/sqrt(10)))
```

```
## [1] 0.81033
```

Thus power or P(reject null(0)|Null is false (true sweetness change is -0.8)) is $\sim$ 0.81

Remember that Power is 1-β = P(reject null | null is false)

In this example, β is 1-0.81 or $\sim$ 0.19

Calculating sample size

To think about sample size for a z-test (or more generally for any test), four things matter:

- Significance level α : How much protection do we want against getting a statistical significant results from our sample when there really is no effect in the population?
- Effect size: How large an effect in the population is important in practice?
- Power $(1 - \beta)$: How confident do we want to be that our study will detect an effect of the size we think is important? I.e., what is the probability of rejecting H_0 when the alternative hypothesis is true?
- variability in the population: Remember that the underlying variability in our population affects the variability of our sample mean

sample size summary

When we are calculating sample size the steps we follow are: - Find the Z alpha and use this to calculate the value of our variable at which we would reject the null. - Find Z beta and use that to calculate what value this would be on the curve of the alternative hypothesis - Set these values equal to each other and solve for n