Recap of Part II probability and distributions

Rules of probability

Review of probability rules

Probabilities are numbers between 0 and 1.

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

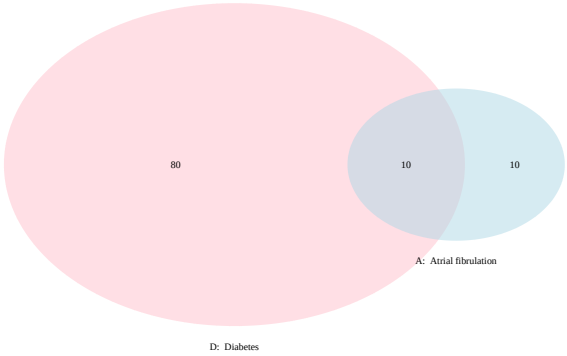
The probabilities in the probability space must sum to 1.

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 $P(A \cup D)$?

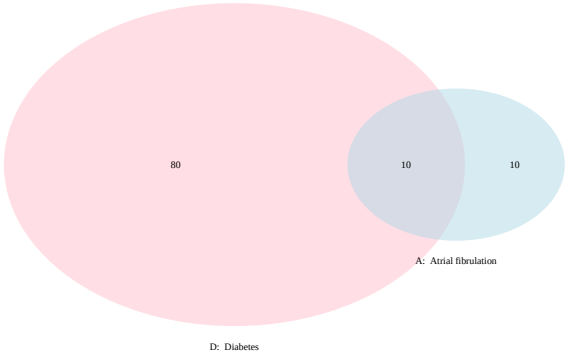
Adding and decomposing probability

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

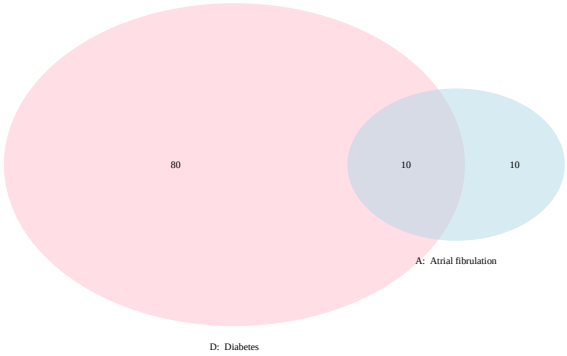
$$P(A \cap B) = P(A) * P(B)$$

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

The Acme Corporation manufactures sticks of dynamite. Unfortunately (fortunately?), these sticks often fail in hilarious but harmless fashion.

Each stick manufactured at Plant A fails with a probability of 0.5

Wile E Coyote has a bag of 10 sticks of dynamite: 4 are from Plant A and 6 are from Plant B. The Coyote pulls 1 stick from the bag at random and lights it.

The probability that Coyote picks a stick from Plant B and it fails is 0.2.

What is the probability that a stick from Plant B fails?

Example calculations

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 are 2 ways 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 (≈ 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 example

Data from the Framingham study allow us to compare the distribution of initial serum cholesterol levels for two populations of males: those who go on to develop coronary heart disease and those who do not. The mean serum cholesterol level in men has a standard deviation of σ=41 mg/100 ml.

The mean initial serum cholesterol level of men who eventually develop coronary heart disease is μ is 244 mg/100ml.

power example

Since it is believed that the mean serum cholesterol for those who do not develop heart disease cannot be higher than the mean level for men who do, a once sided test conducted at the α=0.05 level of significance is appropriate.

a) For this scenario, what is the probability of type I error?

power example

Since it is believed that the mean serum cholesterol for those who do not develop heart disease cannot be higher than the mean level for men who do, a once sided test conducted at the α=0.05 level of significance is appropriate.

a) For this scenario, what is the probability of type I error?

Type I error or α here is the probability of rejecting the null when in fact the null is true. Here we are setting alpha at 0.05 so the probability of making a type I error is 0.05 or 5%.

power example

b) We presume a mean serum cholesterol of 219 among those who do not develop heart disease. If a sample size of 25 is selected from the population of men who do not go on to develop coronary heart disease, what is the probability of making a type II error?

Remember that Beta (type II error) is evaluated under the condition that the alternative hypothesis is true. Here our alternative proposed mean is 219.
We first need to find the value in actual measured units at which we would reject the null.

We can do this two ways:

power example - find value to reject null:

Find the cuttoff in terms of Z score and then convert it to mg/100ml

Note that

$$Z = \frac{x - \mu}{\sigma/\sqrt{n)}}$$

Which we can re-arrange to solve for the x

```
Zalpha_cutpoint<-qnorm(0.05)

#convert
Zalpha_cutpoint_converted=Zalpha_cutpoint*(41/sqrt(25))+244

Zalpha_cutpoint_converted
```

```
## [1] 230.5122
```

power example - find value to reject null:

Or we can use R to give us the cutpoint in units of mg/100ml directly by modifying the qnorm statement

```
cutpoint_alpha<-qnorm(0.05, mean=244, sd=(41/sqrt(25)))

cutpoint_alpha
```

```
## [1] 230.5122
```

note also that we are looking at qnorm of 0.05 here because our hypothesized mean (219) is lower than the null mean (244) so we are interested in the lower tail (the default in R).

power example

Now that we have this cutpoint in terms of the value we are measuring, we can find where this is on the distribution under the scenario where the alternative hypothesis (219) is the truth.

```
pnorm(cutpoint_alpha, mean=219, sd=(41/sqrt(25)), lower.tail=FALSE)
```

```
## [1] 0.08017032
```

This represents the probability of failing to reject the null when we should reject the null.

Note that we are interested in the upper tail here, because the distribution of our alternative hypothesis (mu=219) is lower than (to the left of) the null hypothesis distribution so the cutpoint (230.5) at which we would reject the null is on the right side of our alternative hypothesis distribution.

power example

Note that here again we are using R to give us probability relative to the distribution of means in units of mg/100ml.

We could instead have converted the cutpoint to Z units (this time with respect to the alternative distribution):

$$Z = \frac{230.5122 - 219}{41/\sqrt{25}}$$

```
(230.5122-219)/(41/sqrt(25))
```

```
## [1] 1.403927
```

```
pnorm(1.403927, lower.tail=FALSE)
```

```
## [1] 0.08017029
```

power example

c) What is the power of the test?

Remember that Power is the probability that you reject the null when the hypothesized alternative is true, it is the complement of type II error.

Power is 1-β

so power here is 1-0.08 or 0.92

This means that if the alternative of 219 is true and we draw a 25 person sample, then when we test against a null hypothesis of 244, we would correctly reject the null 92% of the time at an α of 0.05.

How could you increase the power?

The easiest way to increase the power here is to increase the sample size.

Sample size?

You wish to test the null hypothesis mu $=$ 244 mg/100ml against the one sided alternative hypothesis that mu $<$ 244 mg /100ml at the alpha $=$ 0.05 level of significance. If the true population mean is as low as 219 mg/100ml, and you want to risk only a 5% chance of failing to reject the null when the null should be rejected. How large a sample would be required?

Sample size?

Here we are looking for the n, so we need to start with finding the cutpoint (in terms of Z) at which we would reject the null with an alpha of 0.05 and a one sided test.

```
qnorm(0.05)
```

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

Note that we keep the negative here because we are interested in the lower tail.

Sample size?

Now we re-arrange our Z equation to put x (the cutpoint) on one side of the equation:

$$-1.645 = \frac{x - 244}{41/\sqrt{n})}$$

$$x = -1.645 * (41/\sqrt{n}) + 244$$

Now we look for the cutpoint (in terms of Z) for Beta.

In our problem we are now setting the Beta to 0.05 (this is fairly stringent - many studies default to a 0.2 for Beta)

```
qnorm(0.05, lower.tail=FALSE)
```

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

Keep note of which side of the distribution we are working with.

Sample size?

We will re-arrange this Z equation to put x (the cutpoint) on one side of the equation:

$$1.645 = \frac{x - 219}{41/\sqrt{n})}$$

$$x = 1.645 * (41/\sqrt{n}) + 219$$

Since we know that the cutpoints must have the same value in real units, we can now set these two equations equal to each other:

$$-1.645 * (41/\sqrt{n}) + 244 = 1.645 * (41/\sqrt{n}) + 219$$

and now there is only one variable that is unknown (n) which we can solve for.

```
(41/(-25/(-1.645-1.645)))^2
```

```
## [1] 29.1125
```

So we would require a sample size of 30 (can't have .1125 of a person so we round up)

Sample size?

You can check this against the pwr package in R

```
library(pwr)
diff <- 219-244
sigma <-41
d <- diff/sigma
pwr.t.test(d=d, sig.level=.05, power = .95, type = 'one.sample',
           alternative="less")
```

```
##
##      One-sample t test power calculation
##
##               n = 30.51692
##               d = -0.6097561
##       sig.level = 0.05
##           power = 0.95
##     alternative = less
```

Sample size?

The answer you get from pwr.t.test() will be slightly larger because the pwr package is using a t distribution rather than a Z distribution as the basis for testing.

How would the sample size change if you were willing to risk a 10% chance of failing to reject a false null hypothesis?

If we were less stringent, we would need a smaller sample size.

You can check this by finding the Z for this other Beta:

```
qnorm(0.1, lower.tail=FALSE)
```

```
## [1] 1.281552
```

Sample size?

Substituting this in to our previous calculations we would get:

$$-1.645 * (41/\sqrt{n}) + 244 = 1.282 * (41/\sqrt{n}) + 219$$

which we can solve for n.

```
(41/(-25/(-1.645-1.282)))^2
```

```
## [1] 23.04269
```

So we would require a sample size of 24 under these criteria

Sample size?

```
library(pwr)
diff <- 219-244
sigma <-41
d <- diff/sigma
pwr.t.test(d=d, sig.level=.05, power = .9, type = 'one.sample',
           alternative="less")
```

```
##
##      One-sample t test power calculation
##
##               n = 24.45157
##               d = -0.6097561
##       sig.level = 0.05
##           power = 0.9
##     alternative = less
```