Stats Notes

Conditional probability - Bayes Rule

- $P(A)$ - the probability of A occuring
- $P(A \cap B)$ - the intersection - the probability of both A and B occuring
- $P(A \cup B)$ - the union - the probability of A or B (or both) occuring

Note that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The probability of both A and B occuring is not neccasrily just their sum, as this double counts the intersection. This is why avangers don't have double the crit rate of other classes, if you roll double 20s, it only counts as one crit. If A is a roll of 20 on one die, and B is a roll of 20 on the other, then the probabilty of a crit is given by $P(A \cap B)$, which is $1/20 + 1/20 - 1/20 \times 1/20 = 0.0975$

Bayes rule

Bayes rule allows us to invert the condition on a conditional probablility. That is, if we know the probability of B given A, then we can infer the probability of A when B is satisfied/given/present.

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

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

Consider a diagnostic test. We can assess the test for sensitivity and specificity (see below), which describe how the test behaves in the presence of the disease. After taking the test, however, we are probably more interested in inverting this relationship - what is the probability of having the disease given the test was positive?

In this case, A is a condition (or disease), whereas B is an observation (the state of a test - positive or negative)

- $P(A|B)$ is the probability of A given B (i.e., how likely are we to have the disease given a positive test?)
  - this is the **positive predictive value**
- $P(B|A)$ is the probability of B given A (i.e., how likely is the test to be positive when the disease is present)
  - this is the probability of a "true positive"

  - for disease tests this is called the **sensitivity**

- $P(B|A^c)$ is probability of B in the absence of A (i.e. how likely are we to have the disease when the test is false)
- $P(B^c|A^c)$ is probability of not B in the absence of A (i.e. how likely are we to not have the disease when the test is false)
  - this is the probability of a "true negative"
  - for disease tests this is called the **sensitivity**
- $P(A^c|B^c)$ is the probability of not having the disease given the test was negative
  - this is the **negative predictive value**
- $P(A)$ is the **prevalence** , or probability of the condition in the population (given no other information)

Bayes rule gives the probability of having the disease given a positive test:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$
$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + (1 - P(B^c|A^c))(1 - P(A))}$$

Distributions

Bernoulli

coin flip. Outcome is 0 (failure) or 1 (success), with probability of success p probability mass function is

$$P(x) = p^x(1-p)^{1-x}$$

Binomial

Gives the probability of getting some number of successes for k samples of a Bernouli distribution (e.g. probability of rolling a die 6 times and getting 3 4's, probability of having 7 girls from 8 children)

Normal

gaussian

Exponential

discrete and continuous forms

Variance and stuff

Bessel's correction

standard deviaion is defined variance is defined as $\sum((x_i - (\bar{x})^2)/\sqrt{n-1}$. This is to avoid bias. The population variance is given by $\sum_i (x_i - \mu)^2/N$. In reality, we do not know μ, only the sample mean $\bar{x}$, so we compute the sample variance using $\bar{x}$ instead of μ. While the sample mean is an unbiased estimator of the population mean, there is some uncertainty in it. The effect of using the sample mean instead of the population mean is to decrease the variance. What we want is the some of the squared distances from the population mean, but we end up using the sum of squared distances from the sample mean. The sample mean can be shown to be that value that minimises the sum of squared distances, and thus using it will always underestimate the population variance (unless the sample mean happens to be equal to the population mean)

standard error of the mean

The variance of the sample mean is given by

$$var(\bar{X}) = \frac{sigma^2}{n}$$

Even though we don't know the population variance (σ^2), the sample variance (S^2) is an unbiased estimate of this, so we can get a pretty good idea.

The standard deviation of the sample mean is then given by

$$\begin{aligned} sd(\bar{X}) &= \frac{sigma}{\sqrt{n}} \\ &\approx \frac{S}{\sqrt{n}} \end{aligned}$$

In general, the standard deviation of a statistic is called the standard error of that statistic. ## Confidence Intervals

normal confidence intervals

t statistics and confidence intervals

hypothesis testing

Hypothesis testing is used to determine whether or not a measurement is statistically significant. For example, we might measure a mean value to be $\bar{X} = 0.1$, and want to know if this mean is close enough to or appreciably different from a mean of zero.

In hypothesis testing, a null hypothesis is compared to an alternate hypothesis. The null hypothesis is usually something along the lines of "there is no difference between those two samples", "the population mean is equal to zero", etc. It supposes that there is no underlying feature in the data, and that any variation is only due to statistical fluctuations. The alternate hypothesis supposes that there **is** some underlying feature in the data, such as a nonzero difference or mean. The alternate hypothesis may either be one sided ("the population mean is greater than Y" or "the population mean is less than Y"), or two sided ("the population mean is not equal to Y"). The null hypothesis is assumed by default, we require a certain threshold of significance from our data in order to overturn the null hypothesis and instead accept the alternate hypothesis.

For example, we might want to be at least 95% certain that the population mean is greater than some threshold. To start, we assume the null hypothesis (that the population mean is equal to our threshold). We then calculate the probability that the underlying population distribution could produce a sample with mean that we observed. We compute an appropriate statistic (either the t-statistic or the z-statistic, or something else) of our sample mean under the null hypothesis.

Error types

A type I error is incorrectly rejecting the null hypothesis (a false positive). In cases where the null hypothesis is actually true, it will still be rejected with probability $(1 - \alpha)$, where α is the value of the confidence interval being used.

A type II error is where we incorrectly accept the null hypothesis (a false negative). ## Power

Say we have a sample, and two hypotheses: $H_0 : \mu = \mu_0$ and $H_A : mu > mu_0$. The statistical power is the probability of rejecting the null hypothesis, for a given scenario (e.g $\mu = \mu_\alpha, n = 1000$, etc.). To reject the null, the test statistic needs to lie outside our confidence interval. Power is the probability that this will occur if the alternate hypothesis is true, that is , it is the probability that the alternate hypothesis will generate a test statistic outside of the null hypothesis's confidence interval.

If the distributions of the test statistics are narrow and well seperated, then there should be quite a lot of power (the alternate scenario is unlkely to produce a statistic inside the null's confidence interval), something close to 100%. If the distributions are seperated but broad, then there is some probabilty of the alternate scenario yielding a statistic inside the null's interval.

Regression Models

least squares

residuals

inference

Generalised Linear Models

instead of transforming data and then applying a linear model, generalised linear models include any data transformation, so can work directly with observed data.

have three things

- An exponential family model for the variation in the response
- A systematic component (a linear predictor)
- A link function that connects the means of the response to the predictor

In the bernouli example, The response is assumed to follow a Bernoulli distribution $E[Y_i] = \mu_i$, where $0 \leq \mu_i \leq 1$. The predictor is η, which is treated as a linear variable. The link function $g(\mu) = \eta$ give η as a function of μ. For the Bernoulli case, the link function $\eta = g(\mu) = \log \frac{\mu}{1-\mu}$ maps μ (which runs from 0 to 1) to η (which runs from $-\infty$ to ∞)