Independence

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Independent events

- Two events A and B are independent if

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

- Two random variables, X and Y are independent if for any two sets A and B

$$P([X \in A] \cap [Y \in B]) = P(X \in A)P(Y \in B)$$

- If A is independent of B then
  - A^c is independent of B
  - A is independent of B^c
  - A^c is independent of B^c

Example

- What is the probability of getting two consecutive heads?
- $A = \{\text{Head on flip } 1\} \sim P(A) = .5$
- $B = \{\text{Head on flip } 2\} \sim P(B) = .5$
- $A \cap B = \{\text{Head on flips } 1 \text{ and } 2\}$
- $P(A \cap B) = P(A)P(B) = .5 \times .5 = .25$

Example

- Volume 309 of Science reports on a physician who was on trial for expert testimony in a criminal trial
- Based on an estimated prevalence of sudden infant death syndrome of 1 out of $8,543$, Dr Meadow testified that that the probability of a mother having two children with SIDS was $\left(\frac{1}{8,543}\right)^2$
- The mother on trial was convicted of murder

Example: continued

- For the purposes of this class, the principal mistake was to assume that the probabilities of having SIDs within a family are independent
- That is, $P(A_1 \cap A_2)$ is not necessarily equal to $P(A_1)P(A_2)$
- Biological processes that have a believed genetic or familiar environmental component, of course, tend to be dependent within families
- (There are many other statistical points of discussion for this case.)

Useful fact

We will use the following fact extensively in this class:

If a collection of random variables $X_1, X_2, \ldots, X_n$ are independent, then their joint distribution is the product of their individual densities or mass functions

That is, if f_i is the density for random variable X_i we have that

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)$$

IID random variables

- Random variables are said to be iid if they are independent and identically distributed
- iid random variables are the default model for random samples
- Many of the important theories of statistics are founded on assuming that variables are iid

Example

- Suppose that we flip a biased coin with success probability p n times, what is the join density of the collection of outcomes?
- These random variables are iid with densities $p^{x_i}(1-p)^{1-x_i}$
- Therefore

$$f(x_1,\ldots,x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Correlation

- The covariance between two random variables X and Y is defined as

$$Cov(X,Y) = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - E[X]E[Y]$$

- The following are useful facts about covariance
  1. $Cov(X,Y) = Cov(Y,X)$
  2. $Cov(X,Y)$ can be negative or positive
  3. $|Cov(X,Y)| \leq \sqrt{Var(X)Var(y)}$

Correlation

- The correlation between X and Y is

$$Cor(X,Y) = Cov(X,Y)/\sqrt{Var(X)Var(y)}$$

1. $-1 \leq Cor(X,Y) \leq 1$
2. $Cor(X,Y) = \pm 1$ if and only if $X = a + bY$ for some constants a and b
3. $Cor(X,Y)$ is unitless
4. X and Y are uncorrelated if $Cor(X,Y) = 0$
5. X and Y are more positively correlated, the closer $Cor(X,Y)$ is to 1
6. X and Y are more negatively correlated, the closer $Cor(X,Y)$ is to -1

Some useful results

- Let $\{X_i\}_{i=1}^n$ be a collection of random variables
  - When the $\{X_i\}$ are uncorrelated

$$Var\left(\sum_{i=1}^{n} a_i X_i + b\right) = \sum_{i=1}^{n} a_i^2 Var(X_i)$$

- A commonly used subcase from these properties is that if a collection of random variables $\{X_i\}$ are uncorrelated, then the variance of the sum is the sum of the variances

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)$$

- Therefore, it is sums of variances that tend to be useful, not sums of standard deviations; that is, the standard deviation of the sum of bunch of independent random variables is the square root of the sum of the variances, not the sum of the standard deviations

The sample mean

Suppose X_i are iid with variance σ^2

$$
\begin{aligned}
Var(\bar{X}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) \\
&= \frac{1}{n^2} \times n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

Some comments

- When X_i are independent with a common variance $Var(\bar{X}) = \frac{\sigma^2}{n}$
- $\sigma/\sqrt{n}$ is called the standard error of the sample mean
- The standard error of the sample mean is the standard deviation of the distribution of the sample mean
- σ is the standard deviation of the distribution of a single observation
- Easy way to remember, the sample mean has to be less variable than a single observation, therefore its standard deviation is divided by a $\sqrt{n}$

The sample variance

- The sample variance is defined as

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

- The sample variance is an estimator of σ^2
- The numerator has a version that's quicker for calculation

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

- The sample variance is (nearly) the mean of the squared deviations from the mean

The sample variance is unbiased

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] &= \sum_{i=1}^{n} E\left[X_i^2\right] - nE\left[\bar{X}^2\right] \\
&= \sum_{i=1}^{n}\left\{Var(X_i) + \mu^2\right\} - n\left\{Var(\bar{X}) + \mu^2\right\} \\
&= \sum_{i=1}^{n}\left\{\sigma^2 + \mu^2\right\} - n\left\{\sigma^2/n + \mu^2\right\} \\
&= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\
&= (n-1)\sigma^2
\end{aligned}
$$

Hoping to avoid some confusion

- Suppose X_i are iid with mean μ and variance σ^2
- S^2 estimates σ^2
- The calculation of S^2 involves dividing by $n-1$
- $S/\sqrt{n}$ estimates $\sigma/\sqrt{n}$ the standard error of the mean
- $S/\sqrt{n}$ is called the sample standard error (of the mean)

Example

```
data(father.son)
x <- father.son$sheight
n <- length(x)
```

```
round(c(sum((x - mean(x))^2)/(n - 1), var(x), var(x)/n, sd(x), sd(x)/sqrt(n)),
    2)
```

```
## [1] 7.92 7.92 0.01 2.81 0.09
```