Probability And Statistics

Michael Harmon

November 6, 2018

Contents

1 Probability & Random Variables

1.1 Probability

Sample Space ($\mathcal{S}$): The set of all possible outcomes.

Event: Any collection of possible outcomes, $E \in \mathcal{S}$ from the sample space.

Sigma Algebra: Σ is a collection of subsets of $\mathcal{S}$ such that,

1. $\emptyset \in \Sigma$
2. If $A \in \Sigma$, then $A^c \in \Sigma$.
3. If $A_1, A_2, \ldots \in \Sigma$ then $\cup_{i=1}^{\infty} A_i \in \Sigma$

Probability Function: Given a sample space $\mathcal{S}$ and a Sigma algebra Σ then a probability functon with domain Σ satisfies

1. $P(A) \geq 0, \qquad \forall A \in \Sigma$
2. $P(\mathcal{S}) = 1$
3. If $A_1, A_2, \ldots \in \Sigma$ are pairwise disjoint then $P(\cup_{i=1}^{\infty} A_i) = \cup_{i=1}^{\infty} P(A_i)$

Probabilities can be thought of as a frequency of occurrence.

Conditional Probabilities

$$P(A|B) \; = \; \frac{P(A \cap B)}{P(B)}, \qquad P(B) > 0$$

Let $A_1, A_2, \ldots, A_n$ be a partition of $\mathcal{S}$ then $P(B) \; = \; \sum_{i=1}^{n} P(B|A_i)P(A_I$.

Theorem 1 ***Bayes Theorem:***
Let $A_1, A_2, \ldots,$ be a partition of the space space $\mathcal{S}$ then,

$$P(A_i|B) \; = \; \frac{P(B|A_i)P(A_i)}{P(B)} \tag{1.1}$$

$$= \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)} \tag{1.2}$$

Two events A and B are **statistically independent** if

$$P(A \cap B) \; = \; P(A)P(B) \qquad \text{or} \quad P(A|B) = P(A) \tag{1.3}$$

1.2 Random Variables

Random Variable: A function $X : \mathcal{S} \to \mathbb{R}$
Example: The sum of a roll of two die.

Probabilities can be induced by a random variable.

$$P_X(X = X_i), \qquad \text{Discrete}$$

or

$$p(x), \qquad \text{Continuous}$$

Cumulative Distribution Function: $F_X(x) \;=\; P_X(X < x)$

Two random variables X, Y are **identically distributed** if

$$\forall A \in \Sigma : \quad P(X \in A) \;=\; P(Y \in A)$$

Expectation Of Random Variables:

$$\mathrm{E}(X) \quad = \quad \sum_i X_i \, P(X_i) \tag{1.4}$$

$$\mathrm{E}(X) \quad = \quad \int x \, p(x) \, dx \tag{1.5}$$

Variance Of Random Variables: (w/ mean μ_X)

$$\mathrm{Var}(X) \;=\; E[(X \;-\; \mu_X)^2] \;=\; E[X^2] - (E[X])^2 \tag{1.6}$$

Covariance: (of X and Y)

$$\mathrm{cov}(X, Y) \;=\; E[(X - \mu_X)(Y - \mu_Y)]$$

If X and Y are statistically independent then,

$$\mathrm{cov}(X, Y) \;=\; E_X[(X - \mu_X)] E_Y[(Y - \mu_Y)] \tag{1.7}$$

$$= \; 0 \tag{1.8}$$

However, the converse is not true.

Correlation: (of X and Y)

$$\mathrm{corr}(X, Y) \;=\; \frac{\mathrm{cov}(X, Y)}{\sigma_X^2 \sigma_Y^2}$$

Note that, $|\mathrm{corr}(X, Y)| \leq 1$. We also remark that,

$$\mathrm{Var}(aX \pm bY) \;=\; a^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(Y) \pm 2\, a\, b\, \mathrm{cov}(X, Y)$$

so that if X and Y are statistically independent,

$$\mathrm{Var}(aX \pm bY) \;=\; a^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(Y)$$

Conditional Expectation Of Random Variables:

$$\mathrm{E}(X|Y = y) \quad = \quad \sum_i X_i \, P(X_i|Y = y) \tag{1.9}$$

Moment Generating Function

$$M_X(t) \;=\; E_X[e^{tX}] \tag{1.10}$$

Note: $M_X^{(n)}(0) \;=\; E[X^n]$. The MGF has issues with existence.

Characteristic Function

$$\phi_X(t) \;=\; E_X[e^{itX}] \tag{1.11}$$

Note: $(-i)^n \, \phi_X^{(n)}(0) \;=\; E[X^n]$ and for X_i independent,

$$\phi_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n \phi_{X_i}(t)$$

This characteristic function always exists!

Theorem 2 ***Chebyshev Inequality***
Let X be a random variable and $g(X)$ be a non-decreasing function of X then $\forall r > 0$,

$$P\left(g(X) > r\right) \quad \leq \quad \frac{E[g(X)]}{r} \tag{1.12}$$

Convex Function: A function f is convex if $\forall x_1, x_2 \in X$, $\forall t \in [0,1]$ then

$$f\left(tx_1 + (1-t)x_2\right) \quad \leq \quad tf(x_1) + (1-t)f(x_2) \tag{1.13}$$

Theorem 3 ***Jensen's Inequality***
If X is a random variable and ϕ is a convex function then,

$$\phi\left(E[X]\right) \quad \leq \quad E[\phi(X)] \tag{1.14}$$

2 Probability Distributions

Probability distributions defined for both continuous and random variables. Random variables with discrete values have **discrete distributions**, while random variables with continuous values have **continous distributions**.

2.1 Discrete Distributions

Bernoulli Distribution
A Bernoulli random variable binary outcome,

$$x \;=\; \begin{cases} 1, & \text{prob. } p \\ 0, & \text{prob. } 1-p \end{cases}$$

Distribution:

$$P(x \,|\, p) \;=\; p^x (1-p)^{(1-x)} \tag{2.1}$$

Mean: $E[x] \;=\; p$
Variance: $\mathrm{Var}(x) \;=\; p(1-p)$

Binomial Distribution
A binomial random variable y is defined as the sum of n independent Bernoulli random variables, x_i all with prob. p:

$$y \;=\; \sum_{i=1}^n x_i$$

Distribution: The distribution is given as a function of $y = k$ (success), where $k \leq n$ where (n is the number of trials).

$$P(y = k|\, n, p) \;=\; \frac{n!}{(n-k)!k!}\, p^k\, (1-p)^{(n-k)} \tag{2.2}$$

Note: The product comes from the independence of the trials.

Mean: $E[x] \;=\; np$
Variance: $\mathrm{Var}(x) \;=\; n\, p(1-p)$

Note: These results can come from the definition of i.i.d property of the n Bernoulli trials and linearity of Var.

Geometric Distribution
A geometric random variable x is defined as the number $x = k$ of i.i.d Bernoulli trials *until* a success.
Distribution:

$$P(x = k) \;=\; p\,(1-p)^{k-1} \tag{2.3}$$

Mean: $E[x] \;=\; \frac{1}{p}$
Variance: $\mathrm{Var}(x) \;=\; \frac{(1-p)}{p^2}$

Note: A geometric random variable is **memoryless**, i.e. if $s > t$ then $(P(x > s|x > t) \;=\; P(x > s - t)$.

Poisson Distribution
A Poisson random variable used to x is defined as the number of occurrences within a fixed time interval, given that the "average" number of occurrences is λ.
Distribution:

$$P(x|\lambda) \;=\; \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x \;=\; 1, 2, \ldots \tag{2.4}$$

Mean: $E[x] \;=\; \lambda$
Variance: $\mathrm{Var}(x) \;=\; \lambda$

2.2 Continuous Distributions

Beta Distribution
Distribution:

$$P(x|\alpha, \beta) \;=\; \frac{1}{B(\alpha, \beta)} x^{\alpha}(1-x)^{\beta-1} \tag{2.5}$$

For $0 < x < 1$, $\alpha > 0$ and $\beta > 0$.

Mean: $E[x] \;=\; \frac{\alpha}{\alpha+\beta}$
Variance: $\mathrm{Var}(x) \;=\; \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Where,

$$B(\alpha, \beta) \;=\; \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \tag{2.6}$$

$$\;=\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{2.7}$$

Note: Then Beta distribution is useful for deriving other distributions.

Normal Distribution

Distribution:

$$P(x|\mu,\sigma^2) \;=\; \frac{1}{2\pi\sigma^2}e^{-(x-\mu)^2/\sigma^2} \tag{2.8}$$

Mean: $E[x] \;=\; \mu$
Variance: $\mathrm{Var}(x) \;=\; \sigma^2$

Note: Then,

$$P(|x-\mu| \;\le\; \sigma^2) \;\simeq\; 0.67 \tag{2.9}$$
$$P(|x-\mu| \;\le\; 2\sigma^2) \;\simeq\; 0.95 \tag{2.10}$$
$$P(|x-\mu| \;\le\; 3\sigma^2) \;\simeq\; 0.99 \tag{2.11}$$
$$\tag{2.12}$$

Student t Distribution
Arises from estimating mean of $N(\mu,\sigma^2)$ population, but where sample size is small and σ^2 is unknown. The **degrees of freedom** ($df > 2$) is

$$df \;=\; n-1$$

Distribution:
Mean: $E[t] \;=\; 0$
Variance: $\mathrm{Var}(t) \;=\; \frac{df}{df-2}$

Note: The $t-$ distribution has fatter tails than normal distribution and is used for statistical significance between sample means and confidence intervals:

$$t \;=\; \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

χ^2 Distribution
Let $Z_1,\ldots,Z_k$ be indep. normally distributed random variables then,

$$Q \;=\; \sum_{i=1}^{k} Z_i^2 \tag{2.13}$$

is chi^2 distributed with k degrees of freedom.
Distribution:

$$P(x,k) \;=\; \frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}, \qquad x\in[0,\infty) \tag{2.14}$$

Mean: $E[x] \;=\; k$
Variance: $\mathrm{Var}(x) \;=\; 2k$

Note: Is used for chi^2 "Goodness of fit" test.

2.3 Approximation Theorems

Theorem 4 *Normal Approx. To Binomial Random Variables*
When the number of trials, n, is sufficiently large then the probability of x success in n trials each having probability p can be approximated with,

$$Binomial(n,p)(x) \quad \simeq \quad \frac{1}{2\pi\, n\, p\, (1-p)} e^{-(x-np)^2/np(1-p)} \tag{2.15}$$

Or $N_{\mu,\sigma^2}(x)$*, where,* $\mu = np$ *and* $\sigma^2 = np(1-p)$*.*

Theorem 5 *Normal Approx. To Poisson Random Variables*

When the average number of occurrences, λ*, , is sufficiently large then the probability of* k *occurrences can be approximated with,*

$$Poisson(x, \lambda) \quad \simeq \quad \frac{1}{2\pi\lambda} e^{-(x-\lambda)^2/\lambda} \tag{2.16}$$

Or $N_{\mu,\sigma^2}(x)$*, where,* $\mu = \lambda$ *and* $\sigma^2 = \lambda$*.*

3 Statistics & Estimators

3.1 Statistics

A **statistic** is any function of a sample. An **Estimator is any function of a sample *that is used to estimate a population parameter***.

Sample mean:

$$\bar{X} \; = \; \frac{1}{n} \sum_{i=1}^{n} X_i$$

Sample variance

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

Theorem 6 *Let* $X_1, \ldots, X_n$ *be a random sample from* $N(\mu, \sigma^2)$ *and* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *and* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$*. Then,*

1. $\bar{X}$ *and* S^2 *are independent random variables.*
2. $\bar{X} \; \sim \; N(\mu, \sigma^2/n)$
3. $(n-1)\frac{S^2}{\sigma^2} \; \sim \; \chi^2_{n-1}$

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \; = \; \frac{(\bar{X} - \mu)(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} \; = \; \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} \quad \text{(Distributionally!)}$$

That is,

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \text{Student t-distribution with n-1 deg. of freedom}$$

Unbiased Estimator: An estimator $T(X)$ for a parameter (Θ) is unbiased if,

$$E[T(X)] \; = \; \Theta$$

Theorem 7 *Let $X_1, \ldots, X_n$ be a random sample with mean μ and variance S^2, then*

1. $E[\bar{X}] = \mu$
2. $E[S^2] = \sigma^2$
3. $Var(\bar(X) = \frac{\sigma^2}{n}$

Sufficient Statistic: A statistic $T(\mathbf{x})$ is **sufficient statistic** for $\boldsymbol{\theta}$ if the conditional distribution of the sample $\mathbf{x}$ given $T(\mathbf{x})$ does not depend on $\boldsymbol{\theta}$.

Theorem 8 *Sufficiency Principle If $T(\boldsymbol{x})$ is a sufficient statistic for $\boldsymbol{\theta}$, then any inference about $\boldsymbol{\theta}$ should depend on the sample $\boldsymbol{x}$ only through $T(\boldsymbol{x})$.*

Theorem 9 *Factorization Theorem Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf or pm of sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{x})$ is a sufficient statistic for θ iff $\exists\, g(\boldsymbol{t}|\theta)$ and $h(\boldsymbol{x})$ s.t. $\forall \boldsymbol{x}$ and θ,*

$$f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)\, h(\boldsymbol{x})$$

3.2 Convergence Theorems

Convergence In Probability Let $X_1, \ldots, X_n$ be a sequence of random variables then $X_i \rightarrow^{\mathcal{P}} X$. If $\forall \epsilon > 0,\ \lim_{n\to\infty} P(|X_i - X| > \epsilon) = 0$.

Absolute In Convergence Let $X_1, \ldots, X_n$ be a sequence of random variables then $X_i \rightarrow^{A.S.} X$. If $\forall \epsilon > 0,\ P(\lim_{n\to\infty} |X_i - X| > \epsilon) = 0$.

Absolute convergence is equivalent to pointwise convergence.

Convergence In Distribution Let $X_1, \ldots, X_n$ be a sequence of random variables then $X_i \rightarrow^{\mathcal{D}} X$. If $\lim_{n\to\infty} F_{X_n}(X) = F(X)$.

Convergence Relationships

$$X_n \rightarrow^{A.S.} X \Rightarrow X_n \rightarrow^{\mathcal{P}} X$$

$$X_n \rightarrow^{\mathcal{P}} X \Rightarrow X_n \rightarrow^{\mathcal{D}} X$$

Consistent Estimator: An estimator $T(\mathbf{x})$ for $\boldsymbol{\theta}$ is **consistent** if it converges in probability.

Consistency is the minimum requirement for an estimator!

Theorem 10 *Let $X_1, \ldots, X_n$ be a sequence of random variables such that $X_n \rightarrow^{\mathcal{P}} X$ and $h(X)$ is continuous. Then $h(X_n) \rightarrow^{\mathcal{P}} X$.*

Theorem 11 *S^2 is a consistent estimator (if $S_n \to 0$)*
Proof:

$$\begin{aligned}
\lim_{n\to\infty} P(|S_n^2 - \sigma^2|^2 > \epsilon) &\leq \lim_{n\to\infty} \frac{E\left(|S_n^2 - \sigma^2|^2\right)}{\epsilon} && (3.1)\\
&= \lim_{n\to\infty} \frac{Var(S_n^2)}{\epsilon} && (3.2)\\
&= 0 && (3.3)
\end{aligned}$$

by Chebshev's theorem.

Theorem 12 *Weak Law Of Large Numbers*
Let $X_1, \ldots, X_n$ be i.i.d of random variables with $E[x_i] = \mu$ and $Var(x_i) = \sigma^2 < \infty$ then $x_n \to^{\mathcal{P}} \mu$ ***Proof:***

$$\begin{aligned}
\lim_{n\to\infty} P(|\bar{x} - \mu| > \epsilon) &\leq \lim_{n\to\infty} \frac{E\left(|\bar{x} - \mu|^2\right)}{\epsilon^2} &(3.4)\\
&= \lim_{n\to\infty} \frac{Var(\bar{x})}{\epsilon^2} &(3.5)\\
&= \lim_{n\to\infty} \frac{\sigma^2_n}{n\,\epsilon} &(3.6)\\
&= 0 &(3.7)
\end{aligned}$$

Theorem 13 *Central Limit Theorem*
Let $X_1, , X_2, \ldots,$ be i.i.d of random variables with $E[x_i] = \mu$ and $0 < Var(x_i) = \sigma^2 < \infty$ then,

$$\frac{X_n - \mu}{\sigma/\sqrt{n}} \to^{\mathcal{D}} N(0,1)$$

3.3 Maximum Likelyhood Esimators

The **maximum likelyhood estimator** is the value of a population distribution θ that maximizes the probability of observing the sample. We can find the MLE from a random sample$x_1, x_2, \ldots x_n$ from $f(x|\theta)$ then the **likelyhood** function is defined as,

$$L(\theta\,|x_1,\, x_2,\, \ldots,\, x_n) = \prod_{i=1}^{n} f(x_i\,|\,\theta)$$

Then we can find the MLE $\widehat{\theta}$ such that,

$$\frac{\partial L(\theta\,|x_1, x_2, \ldots, x_n)}{\partial\theta} = 0 \qquad \text{or} \qquad \frac{\partial \log(L(\theta\,|x_1, \ldots))}{\partial\theta} = 0$$

Theorem 14 *Invariance Property Of MLE*
If $\widehat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\widehat{\theta})$.

Example: $X_1, , X_2, \ldots, X_n$ i.i.d. Bernoulli(ϕ). Find the MLE of ϕ.
The likelyhood function is,

$$\begin{aligned}
L(\phi\,|x_1,\, x_2,\, \ldots,\, x_n) &= \prod_{i=1}^{n} f(x_i\,|\,\phi) &(3.8)\\
&= \prod_{i=1}^{n} \phi^{x_i}\,(1-\phi)^{1-x_i} &(3.9)\\
&= \phi^{\sum_{i=1}^{n} x_i}\,(1-\phi)^{\sum_{i=1}^{n}(1-x_i)} &(3.10)\\
&= \phi^{n\bar{x}}\,(1-\phi)^{n\,(1-\bar{x})} &(3.11)
\end{aligned}$$

Or,

$$\log(L(\phi\,|\ldots)) = n\,\bar{x}\,\log(\phi) + n\,(1-\bar{x})\,\log(1-\phi) \qquad (3.12)$$

So,

$$\frac{\partial \log(L)}{\partial \theta} = \frac{n\,\bar{x}}{\phi} - \frac{n\,(1-\bar{x})}{(1-\phi)} = 0 \tag{3.13}$$

$$\Rightarrow n\,\bar{x}\,(1-\widehat{\phi}) - n\,\widehat{\phi}\,(1-\bar{x}) = 0 \tag{3.14}$$

$$\Rightarrow \widehat{\phi} = \bar{x} \tag{3.15}$$

This is the average of overall positive outcomes. We can test that this is the maximum, by taking the second derivative:

$$\frac{\partial^2 \log(L(\widehat{\phi})}{\partial \theta^2} = \frac{-n\,\bar{x}}{\widehat{\phi}^2} - \frac{n\,(1-\bar{x})}{(1-\widehat{\phi})^2}|_{\widehat{\phi}=\bar{x}} \tag{3.16}$$

$$= \frac{-n\,\bar{x}}{\bar{x}^2} - \frac{n\,(1-\bar{x})}{(1-\bar{x})^2} \tag{3.17}$$

$$= \frac{-n\,(1-\bar{x})\,-n\,\bar{x}}{\bar{x}\,(1-\bar{x})} \tag{3.18}$$

The MLE has issues with existence.

Theorem 15 *The MLE is a consistent estimator*
It also has optimal variance, but can be difficult to compute.

3.4 Bayesian Estimators

In the **classical approach or frequentist**, θ is known, but fixed. $x_1, x_2, \ldots, x_n$ are drawn from a population index by *theta* and knowledge about the value of is θ is obtained. In a **Bayesian approach**, θ is a quantity whose variation can be described by a probability distribution called a prior, $P(\theta)$. A sample is taken from a population and used to update the prior distribution , now called the posterior distribution, $P(\theta\,|\,\mathbf{x})$.

Let $f(\theta\,|\,\mathbf{x})$ be the sampling distribution then,

$$P(\theta\,|\,\mathbf{x}) = \frac{P(\mathbf{x}\,|\,\theta)\,P(\theta)}{m(\mathbf{x})}, \quad \text{where} \quad m(\mathbf{x}) = \int P(\mathbf{x}\,|\,\theta)P(\theta)d\theta$$

The **Bayesian estimator** could then be taken the be the expected value:

$$\widehat{\theta} = E(\theta\,|\,\mathbf{x})$$

This requires us to calculate the full posterior distribution. Instead, one could use a **Bayesian estimator** that is called the **Maximum A-Posteriori (MAP)**:

$$\widehat{\theta} = \max_{\theta} P(\theta\,|\,\mathbf{x})$$

Note: Bayesian estimators are ALWAYS biased due to their choice of prior, however, they can reduce the variance in our estimators.

Example: Let $y \sim bin(n,p)$ and $p \sim beta(\alpha,\beta)$, then

$$P(p\,|\,y) = \frac{P(y\,|\,p)P(p)}{m(y)} \tag{3.19}$$

$$= \binom{n}{k} p^y\,(1-p)^{n-y}\,\frac{\Gamma(\alpha+\beta}{\Gamma(\alpha)\,\Gamma(\beta)}\,p^{\alpha-1}\,(1-p)^{\beta-1} \tag{3.20}$$

$$= beta(y+\alpha, n-y+\beta) \tag{3.21}$$

This means the Bayesian estimate of p is,

$$\widehat{p}_{\text{Bayes}} = \frac{y+\alpha}{y+\alpha+(n-y+\beta} \tag{3.22}$$

$$= \frac{y+\alpha}{\alpha+\beta+n} \tag{3.23}$$

$$= \left(\frac{n}{\alpha+\beta+n}\right)\frac{y}{n} + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right) \tag{3.24}$$

Note that this is a linear combination of the sample mean and the prior mean. However, as the sample size grows the contribution of the prior mean grows smaller and we get more confident in the sample mean.

In the limit as the sample size $n \to \infty$ Bayesian and classical estimators should converge to the same estimator.

Indeed, the maximum likelyhood estimator is the same thing as a maximum a-posteriori estimator with uniform prior!

3.5 Evaluating Estimators

For continuous random variables

$$\text{MSE} = E_\theta(\widehat{\theta}-\theta)^2 \tag{3.25}$$

$$= \tag{3.26}$$

4 Confidence Intervals

5 Statistical Testing

5.1 Hypothesis Testing

5.2 χ^2 "Goodness Of Fit" Tests

5.3 AB Testing