# Statistics review

**via Spiegel, Schiller, Srinivasan (2009) Schaum's Probability and Statistics (third edition)**

Ian Stewart

July 26, 2017

These notes represents my study of statistics during the summer of 2017, drawing on the excellent statistics primer *Schaum's Probability and Statistics (third edition)* (Spiegel, Schiller, Srinivasan (2009)) particularly for the sample problems. I'm going to skip a lot of useful details like derivations and proofs in order to focus on application rather than theory, but I recommend Wasserman's (2004) *All of Statistics* for an in-depth mathematical treatment of statistical theory. Some would say that without all the proofs, there cannot be a complete understanding of statistics. I agree, but I also think that a surface-level understanding of statistics is often sufficient to do the hypothesis testing and analysis involved in social science research. You can never learn too much, but you also have to determine for yourself how far down the rabbit hole you want to go.

The examples that I use come from social science and natural language processing contexts, because that's what I do.

## 1 Useful distributions

Here I outline the basic ideas of some useful probability distributions and explain their application to sample problems.

### 1.1 Binomial

The binomial distribution models the joint probability of $X$ successes in $n$ Bernoulli trials, where the probability of success is static over time. We say that $p$ is the probability of success in a single trial, and $q$ is the probability of failure in a single trial.

#### Sample problem 4.68

Let's say the probability of survival within 30 years for 30-year-olds in population P is $\frac{2}{3}$. Not sure why statisticians are so morbid, but we'll go with it for now. What is the

| | |
|---|---|
| PDF | $P(X = x) = \binom{n}{x} p^x q^{n-x}$ |
| $\mu$ | $np$ |
| $\sigma$ | $\sqrt{npq}$ |

Table 1: Useful facts about the binomial distribution.

probability that, out of a sample of 5 30-year-olds: (a) 5 survive, (b) more than 3 survive, (c) 2 survive, (d) at least one survives?

(a) $P(X = 5) = \binom{5}{5}(\frac{2}{3})^5(\frac{1}{3})^0 = \frac{32}{243} = 13.2\%$

(b) $P(X \geq 3) = P(X = 4) + P(X = 5) = \binom{5}{4}(\frac{2}{3})^4(\frac{1}{3})^1 + \binom{5}{5}(\frac{2}{3})^5(\frac{1}{3})^0 = 79.0\%$

(c) $P(X = 2) = \binom{5}{2}(\frac{2}{3})^5(\frac{1}{3})^0 = 16.5\%$

(d) $P(X \geq 1) = \sum_{1 \leq i \leq 5} P(X = i) = 1 - P(X = 0) = \binom{5}{0}(\frac{2}{3})^0(\frac{1}{3})^5 = 99.6\%$

**Application** The binomial distribution is often used to test whether an event is occurring more frequently than by chance. In the context of the sample problem, a study may find that the number of 30-year-olds in a certain subpopulation far exceeds the number expected by previous findings, by comparing the actual number of survivals with the expected survival count ($\mu = np$) generated by the binomial distribution.

## 1.2 Multinomial

The multinomial distribution models the probability of $i$ separate independent random variables $X_i$ occurring $n_i$ times over $n$ Bernoulli trials, such as drawing differently colored marbles out of a (large) bag. We say that $p_i$ is the probability of success for variable $X_i$ in a single trial, and that $q_i$ is the probability of failure.

| | |
|---|---|
| PDF | $P(X_i = n_i) = n! \Pi \frac{p_i^{n_i}}{n_i!}$ |
| $\mu_i$ | $np_i$ |
| $\sigma_i$ | $\sqrt{np_iq_i}$ |

Table 2: Useful facts about the multinomial distribution.

**Sample problem 4.96**

Let's say the population P of a large fictional country is geographically divided such that the ratio of South (S) to North (N) to East Coast (E) to West Coast (W) is 4:3:2:1. Drawing a sample of size $n = 10$, what is the probability of drawing: (a) 4 South, 3 North, 2 East and 1 West, (b) 8 South, 2 West?

For both (a) and (b), we compute the probabilities of drawing a single member of one of the subpopulations as:

$p_S = \frac{4}{4+3+2+1} = \frac{4}{10} = 0.4, p_N = 0.3, p_E = 0.2, p_W = 10\%$
(a) $P(n_S = 4, n_N = 3, n_E = 2, n_W = 1) = 10! \frac{0.4^4 0.3^3 0.2^2 0.1^1}{4!3!2!1!} = 3.48\%$
(b) $P(n_S = 8, nW = 2) = 10! \frac{0.4^8 0.1^2}{8!2!} = 0295\%$

**Application** The multinomial distribution is frequently used in topic modeling situations (e.g., Mei, Shen and Zhai (2007)) because each document in a collection can be said to be drawn from a multinomial distribution of a fixed number of topics.

## 1.3 Hypergeometric

The hypergeometric distribution models the probability of drawing $X$ successes in $n$ Bernoulli trials from a population of size $N$, without replacement. For instance, drawing sample individuals from a population without replacement. Just like the binomial distribution, we say $p$ is the probability of observing success in one trial, and $q$ is the probability of observing failure.

| | |
|---|---|
| PDF | $P(X = x) = \frac{\binom{Np}{x}\binom{Nq}{n-x}}{\binom{N}{n}}$ |
| $\mu$ | $np$ |
| $\sigma$ | $\sqrt{\frac{npq(N-n)}{N-1}}$ |

Table 3: Useful facts about the hypergeometric distribution.

**Sample problem 4.98**
The population $P$ contains 15 individuals, with 5 vegetarians and 10 omnivores. In a random sample of 8 individuals, what is the probability of drawing: (a) 4 vegetarians? (c) at least 1 vegetarian?
(a) $P(X = 4) = \frac{\binom{5}{4}\binom{10}{8-4}}{\binom{15}{8}} = \frac{70}{429} = 16.3\%$
(b) $P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{5}{0}\binom{10}{8}}{\binom{15}{8}} = 1 - \frac{1}{13*11} = 99.3\%$

## 1.4 Geometric

The geometric distribution models the probability of undergoing $X$ Bernoulli trials until the first success. Just like the binomial distribution, we say that $p$ is the probability of success and $q$ is the probability of failure.

**Sample problem 4.52**

| | |
|---|---|
| PDF | $P(X = x) = pq^{x-1}$ |
| $\mu$ | $\frac{1}{p}$ |
| $\sigma$ | $\sqrt{\frac{q}{p^2}}$ |

Table 4: Useful facts about the geometric distribution.

We have a text corpus of pizza toppings (each a unigram token), of which there are 6 unique types in equal proportions. Sampling repeatedly from the corpus, what is the probability of drawing "anchovies" on the fifth sample?

$P(X = anchovies) = (\frac{1}{6})(\frac{5}{6})^4 = 0.0804$

## 1.5 Negative binomial

The negative binomial distribution models the probability of observing $X$ Bernoulli trials until the $r^{th}$ success. The geometric distribution is a special case of this distribution in which $r = 1$.

| | |
|---|---|
| PDF | $P(X = x) = \binom{x-1}{r-1} p^r q^{x-r}$ |
| $\mu$ | $\frac{r}{p}$ |
| $\sigma$ | $\frac{\sqrt{rq}}{p}$ |

Table 5: Useful facts about the negative binomial distribution.

**Sample problem** Using the same text corpus as in the geometric distribution problem (6 toppings in equal proportions), what is the probability of drawing 20 tokens until we observe 5 "anchovies" tokens?

$p = \frac{1}{6}, r = 5, x = 20$

$P(X = 20) = \binom{19}{4} \frac{1}{6}^5 \frac{5}{6}^{15} = 3876 \frac{5^{15}}{6^{20}} = 3.24\%$

## 1.6 Poisson

The Poisson distribution models the probability of observing $X$ successes in a sample occurring at a fixed rate $\lambda$ in a given timespan. The Poisson distribution can approximate the binomial distribution in cases with a large $n$ ($n \geq 50$) and a low success rate ($p < 0.1$), which means that $\lambda$ can be approximated as $np$.

**Sample problem 4.93**

Over a corpus of pizza topping words, the average rate of seeing the word "mushroom" is 3 per 100,000 tokens. In a sample of 200,000 tokens, what is the probability of observing (a) zero "mushroom" tokens, (e) between 4 and 8 "mushroom" tokens?

| | |
|---|---|
| PDF | $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$ |
| $\mu$ | $\lambda$ |
| $\sigma$ | $\sqrt{\lambda}$ |

Table 6: Useful facts about the Poisson distribution.

$p = \frac{3}{100000}, \lambda = np = 200,000(\frac{3}{100000}) = 6$
(a) $P(X = 0) = \frac{6^0 e^{-6}}{0!} = e^{-6} = 0.248\%$
(b) $P(4 \leq X \leq 8) = \sum_{4 \leq x \leq 8} \frac{\lambda^x e^{-6}}{x!} = e^{-6} \sum_{4 \leq x \leq 8} \frac{\lambda^x}{x!} = 69.6\%$

**Application** The Poisson distribution has been used to model the average word frequency rate of individual users online (Altmann, Pierrehumbert and Motter 2011), which makes sense if you consider each individual user to have a different but constant rate of using a particular word.

## 1.7 Normal

The normal distribution is the most widely used continuous probability function due in part to its ubiquity in natural phenomena. After estimating a fit to a normal distribution, it is common to convert value $x$ to its standard normal form $Z = \frac{x-\mu}{\sigma}$, which is normally distributed with $\mu = 0$ and $\sigma = 1$.

| | |
|---|---|
| PDF | $P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$ |
| $\mu$ | $\lambda$ |
| $\sigma$ | $\sqrt{\lambda}$ |

Table 7: Useful facts about the normal distribution.

**Sample problem 4.81**

A population P uses the word "y'all" for the second-person plural, at an average rate $\mu = 61.4\%$ with standard deviation $\sigma = 0.25\%$. What is the proportion of the population that uses "y'all" (a) between 61% and 61.8% of the time? (b) exactly 61.5% of the time?

Let's say $x$ is the proportion of "y'all"-ers.

(a) $P(0.61 \leq x \leq 0.618) = P(\frac{0.61-0.614}{0.0025} \leq Z \leq \frac{0.618-0.614}{0.0025}) = P(-1.6 \leq Z \leq 1.6) = P(Z \geq -1.6) - P(Z \geq 1.6) = 89.04\%$

(b) $P(x = 0.615) = P(Z = \frac{0.615-0.614}{0.0025}) = P(Z = 0.4) = 36.8\%$

**Applications** The normal distribution shows up in a lot of places, due in part to the Central Limit Theorem which shows that the sum of a set of independent and identi-

cally distributed (IID) variables tends toward a normal distribution. Furthermore, many statistical analysis techniques such as a linear regression rely on assumptions of normality.

## 1.8 Chi-square

The chi-square distribution models the distribution of a sum $\chi^2$ of the squares of independent normally distributed variables $\{\forall v X_v\}$:
$\chi^2 = \sum_v X_v^2$
The parameter $v$ is known as the degrees of freedom, which makes sense if you think about higher $v$ indicating more variability in the underlying $X$ values and thus more "freedom" in the distribution of $\chi^2$.

| | |
|---|---|
| PDF | $P(X = x) = \frac{1}{2^{v/2}\Gamma(v/2)} x^{(v/2)-1} e^{-x/2}$ |
| $\mu$ | $v$ |
| $\sigma$ | $2v$ |

Table 8: Useful facts about the chi-squared distribution.

**Sample problem 4.111**
Assume $\{\forall v X_v\}$ is a set of independent normally distributed random variables. If we look at the $99^{th}$ percentile of the distribution of the sum of $X_v$, what is the value of $\chi^2$ that we would observe with: (a) 8 degrees of freedom? (b) 28 degrees of freedom?

Just to be clear, we are looking up chi-squared values in a table of percentile values versus degrees of freedom, such that the area under the distribution curve is at the $99^{th}$ percentile.
(a) $\chi^2_{0.99,8} = 20.1$
(b) $\chi^2_{0.99,28} = 48.3$

**Applications** The chi-squared distribution is often used in hypothesis testing to compare combinations of normal distributions, such as Chancellor, Mitra and De Choudhury (2016) in which the difference between two predictive models' fit to data ("deviance") follows a chi-squared distribution with degrees of freedom equal to the difference in the number of predictive variables.

## 1.9 Student's t

Student's t distribution is used to approximate a normal distribution with a known mean and unknown standard deviation (over sample $\{\forall_i x_i\}$, sample mean $\bar{X}$, sample variance $S$, sample size $n$):

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

where $n-1=v$ is the degrees of freedom. When $v>30$, the t distribution approximates the normal distribution, so the t distribution is often used in cases where the sample size is small. The distribution is symmetric, so $t_{v,p}=-t_{v,1-p}$ and $P(c\leq t)=1-P(-c\leq t)$ for a constant $c$.

| | |
|---|---|
| PDF | $P(X=x)=\frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})}(1+\frac{x^2}{v})^{-\frac{v+1}{2}}$ |
| $\mu$ | 0 |
| $\sigma$ | $\frac{v}{v-2}$ |

Table 9: Useful facts about student's t distribution.

**Sample problem** *4.119*

If variable $U$ is drawn from student's distribution with $v=10$, find constant $c$ such that: (a) $P(U>c)=0.05$, (b) $P(-c\leq U\leq c)=0.98$, (c) $P(U\leq c)=0.20$

(a) $c=t_{8,1-0.05}=t_{8,0.95}=1.86$

(b)

$$P(-c\leq U\leq c)=P(c\leq U)-P(-c\leq U)=1-2(P(-c\leq U))=0.98, P(-c\leq U)=0.01, P(c\leq U)=0.99 t_{10,0.}$$

(c)

$$P(U\leq c)=0.20, P(c\leq U)=1-0.20=0.80, t_{10,0.80}=0.879, t_{10,0.20}=-0.879$$

**Applications** This distribution is often used in the comparison of two sample means whose standard deviations are assumed to be identical, such as Pavalanathan and Eisenstein (2016) who use a paired t-test to compare the distribution of emoticon use between a control and treatment population.

## 1.10 F

The F distribution models the distribution of the ratio of chi-squared variables $V_1$ and $V_2$, with respective degrees of freedom $v_1$ and $v_2$, such that

$$\frac{V_1/v_1}{V_2/v_2}=V\sim F_{v_1,v_2}$$

Note the use of multiple degrees of freedom $v_1, v_2$.

The F distribution is not symmetric but the reciprocal statistic value can be computed as follows: $F_{1-p,v_2,v_1}=\frac{1}{F_{p,v_1,v_2}}$

| | |
|---|---|
| PDF | $P(X = x) = \frac{\Gamma(\frac{v_1+v_2}{2})}{\Gamma(\frac{v_1}{2})\Gamma(\frac{v_2}{2})} v_1^{v_1/2} v_2^{v_2/2} u^{(v_1/2)-1} (v_2 + v_1 x)^{(v_1+v_2)/2}$ |
| $\mu$ | $\frac{v_2}{v_2-2}$ $(v_2 > 2)$ |
| $\sigma$ | $\sqrt{\frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-4)(v_2-2)^2}}$ $(v_2 > 4)$ |

Table 10: Useful facts about the F distribution.

#### Sample problem

**Applications** The F distribution is used in the analysis of variance when a sample's standard deviation is unknown, such as Prabhakaran, Reid and Rambow (2014) which used an ANOVA test to compare the group means of different predictive features to assess which features significantly differentiated different subgroups.

## 2 Sampling