H5. Discrete Probability Distributions

Discrete Uniform

$$\mu = \frac{1}{k}\sum_{i=1}^{k} x_i \text{ , and } \sigma^2 = \frac{1}{k}\sum_{i=1}^{k}(x_i - \mu)^2$$

Binomial Distribution

1. Experiment consists of n repeated trials
2. Outcome either **success** or **failure**
3. Prob. of success, p, remains constant trial to trial
4. Repeated trials are independent, **with replacement**

Success with prob p, failure with q = *1 - p*; number of successes in n independent trials.

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, .., n$$

$$\mu = np \text{ , and } \sigma^2 = npq$$

Multinomial Distribution

A given trial results in k outcomes, $E_1, ..., E_k$ with probabilities $p_1, ..., p_k$, then $X_1, ...X_k$, represents number of occurrences for $E_1, ...E_k$

$$f(x_1, .., x_n; p_1, .., p_k, n) = \binom{n}{x_1, ..., x_k} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$$

with

$$\sum_{i=1}^{k} x_i = n, \text{ and } \sum_{i=1}^{k} p_i = 1.$$

Hypergeometric Distribution

1. A random sample size n **without replacement** from N items.
2. K of N items are successes, $N - k$ are failures.

The number of successes X in a rand. sample size of n from N items, k are **success**, $N - k$ are **failure**:

$$h(x; N, n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \text{ ,}$$

$$max\{0, n - (N - k)\} \leq x \leq min\{n, k\}.$$

$$\mu = \frac{nk}{N} \text{ , and } \sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}\left(1 - \frac{k}{N}\right).$$

Multivariate Hypergeometric Distribution

N items partititioned in k cells $A_1, ..., A_k$ with $a_1, ..., a_k$ elements, then $X_1, ..., X_k$, selected from $A_1, ...A_k$ with sample size n:

$$f(x_1, ..., x_k; a_1, ...a_k, N, n) = \frac{\binom{a_1}{x_1}...\binom{a_k}{x_k}}{\binom{N}{n}}$$

$$\text{with } \sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} a_i = N.$$

Negative Binomial Distribution

Repeated ind. trials result in success with p and failure with $q = 1 - p$, then the number of the trial on which the kth success occurs is X:

$$b^*(x; k, p) = \binom{x-1}{k-1} p^k q^{x-k} \text{ , } x = k, k+1, ...$$

Geometric Distribution

Repeated ind. trials result in success with p and failure with $q = 1 - p$, then the number of the trial on which the first success occurs is X:

$$g(x; p) = pq^{x-1} \text{ , } x = 1, 2, ...$$

$$\mu = \frac{1}{p} \text{ , and } \sigma^2 = \frac{1-p}{p^2}.$$

Poisson Distribution

1. Number of outcomes is independent on time interval or region of space. **(no memory)**
2. Probability of single outcome is proportional to length of time interval / size region.
3. Probability of more than 1 outcome of small region is negligible.

Number of outcomes is computed from $\mu = \lambda t$, where t is the used "time","distance", etc. The number of outcomes occurring in a give time interval or region denoted as t, is X:

$$p(x; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!} \text{ , } x = 0, 1, ...$$

$$\mu = \sigma^2 = \lambda t$$

Let X be binomial random variable with $b(x; n, p)$. When $n \to \infty, p \to 0$, and $np \xrightarrow{n\to\infty} \mu$ remains constant:

$$b(x; n, p) \xrightarrow{n\to\infty} p(x; \mu)$$

H6. Continuous Probability Distributions

Continuous Uniform Distribution

The density function of X on interval $[A, B]$ is:

$$f(x; A, B) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B \\ 0, & \text{elsewhere.} \end{cases}$$

$$\mu = \frac{A+B}{2}, \text{ and } \sigma^2 = \frac{(B-A)^2}{12}$$

Normal Distribution

Density of normal random var. X, with mean μ and variance σ^2 is:

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \; -\infty < x < \infty,$$

Transform any observations of X to Z with mean 0 and variance 1 by:

$$Z = \frac{Z - \mu}{\sigma} \text{ equivalent to: } X = \sigma Z + \mu$$

Approximate Binomial Dist. with Normal

X is a binomial rand. var. with mean $\mu = np$ and variance $\sigma^2 = npq$, as $n \to \infty$, is stand. norm. dist $n(z; 0, 1)$:

$$Z = \frac{X - np}{\sqrt{npq}}$$

X is a binom. rand. var. with n and p. X has a approx. norm. dist. with $\mu = np$ and $\sigma^2 = npq = np(1-p)$:

$$\begin{aligned} P(X \leq x) &= \textstyle\sum_{k=0}^{x} b(k; n, p) \\ &\approx \text{area under normal curve to left of } x + 0.5 \\ &= P\left(Z \leq \frac{x+0.5-np}{\sqrt{npq}}\right) \end{aligned}$$

Good if np and $n(1-p) \geq 5$

Gamma Distribution

The **gamma function** is defined by:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \text{ for } \alpha \geq 0.$$

$$\text{equivalent to: } \Gamma(n) = (n-1)!$$

General: Time occurring until α ***specified number of Poisson events occur****.* X has **gamma distribution**, with parameters α and β if density is given by:

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, & x \geq 0 \\ 0, & \text{elsewhere.} \end{cases}$$

where $\alpha \geq 0, \beta \geq 0$

$$\mu = \alpha\beta, \text{ and } \sigma^2 = \alpha\beta^2$$

X has **exponential distribution**, with parameters β ($\alpha = 1$) if density is given by:

$$f(x; \beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}}, & x \geq 0 \\ 0, & \text{elsewhere.} \end{cases}$$

where $\beta \geq 0$

$$\mu = \beta, \text{ and } \sigma^2 = \beta^2$$

Using λ:

$$f(x) = \lambda e^{-\lambda x}$$

Incomplete gamma function, to be used to solve integral of gamma function:

$$F(x; \alpha) = \int_0^x \frac{y^{\alpha-1} e^{-\frac{x}{b}}}{\Gamma(\alpha)} dy$$

Chi-Squared Distribution

A special case gamma distribution,
when $\alpha = v/2, \beta = 2$.
X then has a chi-squared distribution, with v degrees of freedom:

$$f(x; v) = \begin{cases} \frac{1}{2^{\frac{v}{2}} \Gamma(\frac{v}{2})} x^{\frac{v}{2}-1} e^{-\frac{x}{2}}, & x \geq 0 \\ 0, & \text{elsewhere.} \end{cases}$$

$$\mu = v \text{ and } \sigma^2 = 2v$$

H8. Fundamental Sampling Distributions and Data Descriptions

Definitions

Sample A subset of a population

Biased Sampling procedure which consistently over-, or underestimates some characteristic of a population.

Statistic Any function of the random variables constituting a random sample

Sample Mean If $X_1, ..., X_n$ represent a sample of size n, then sample mean is: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Sample Variance If $X_1, ..., X_n$ represent a sample of size n, then sample variance is:
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ or
$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]$

Sample Standard Deviation Denoted by S, is the positive square root of the sample variance S^2

Box-and-whisker plot A plot depicting the minimum, maximum, mean, 1st and 3rd quantile (25% & 75%) ranges.

Outlier A observation outside a multiple of the interquantile range, usually 1.5 times.

Quantile Denoted $q(f)$, a value for which a specified fraction f of the data values is $\leq q(f)$. Sample median is $q(0.5)$.

Quantile Plot Plots the **data values** on the vertical axis, against the **fraction** of observations with this value on the horizontal axis. The equation for this fraction: $f_i = \frac{i-\frac{3}{8}}{n+\frac{1}{4}}$

Normal Quantile-Quantile Plot The plot of $y_{(i)}$ (ordered observations) against the normal distribution, $q_{(\mu,\sigma)}$, with $\mu = 0, \sigma = 1$ so: $q_{0,1}(f_i)$

Sampling Distribution The probability distribution of a statistic.

Central Limit Theorem

Uses Sampling Distribution of Means

Formula $\bar{X}$ is the mean of rand. sample, size n, taken from population with mean μ, variance σ^2, then:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

as $n \to \infty$, is stand. norm. distr. $n(z; 0, 1)$

Sampling Distribution of the Difference between Two Averages

Uses Find the difference in mean between two populations

Formula Independent samples, size n_1 and n_2, drawn from different populations, with means μ_1, μ_2, variances σ_1^2, σ_2^2, then the sampling distribution $\bar{X}_1 - \bar{X}_2$ is norm. distr. with:

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ , and } \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Hence

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

Sampling Distribution of S^2

Uses Find the variances of a population.

Formula If S^2 is variance in rand. sample size n, taken from normal population with σ^2.

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

chi-squared distribution with $v = n - 1$ degrees of freedom.

Degrees of Freedom n, if taken from normal dist. $n-1$ if μ is unknown.

t-Distribution

Uses Inference on population mean, μ; or cases that involve comparative samples. (significant different means)

Formula Z is a stand. norm. rand. var., V is a chi-sq. rand. var. with v degrees freedom. If Z and V are independent, then $T = \frac{Z}{\sqrt{V/v}}$ is given by:

$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}}\left(1 + \frac{t^2}{v}\right)^{-(v+1)/2} \quad , \; -\infty < t < \infty$$

Or, let $X_1, .., X_n$ be ind. rand. var. , all normal, with mean μ and standard dev σ.

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

then $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$, is t-distribution with $v = n - 1$ degrees freedom.

f-Distribution

Uses Used in two-sample situations inferencing on population variances.

Formula U and V are two ind. rand. var. with chi-sq. dist. with v_1, v_2 respectively. $F = \frac{U/v_1}{V/v_2}$ is given by:

$$h(f) = \begin{cases} \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & , f > 0 \\ 0 & , f \leq 0 \end{cases}$$

Writing $f_\alpha(v_1, v_2)$ with v_1 and v_2 degrees freedom:
$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_1, v_2)}$

With 2 sample variances: S_1^2 and S_2^2 of rand. samples of size n_1, n_2, from variances σ_1^2,σ_2^2

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2/\sigma_2^2}{S_2^2/\sigma_1^2}$$

has F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees freedom.