Lecture 1: Data Handling

Sequences

Mean	Variance
$\bar{x} = \frac{1}{n}\Sigma x$	$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$

Percentiles

- P_y is simply the (100p)th percentile when $p = \frac{y}{100}$
- Q_1, Q_2 and Q_3 are equivalent to P_{25}, P_{50} and P_{75}, respectively

$q(p) = x_k + \alpha(x_{k+1} - x_k)$	$k = \lfloor p(n+1) \rfloor$	$\alpha = p(n+1) - \lfloor p(n+1) \rfloor$

Lecture 2: Probability Distributions

Discrete Random Variables

Mean	Variance
$\mu = E(X) = \Sigma x \mathrm{P}(\mathrm{X} = x) = \Sigma x f(x)$	$\sigma^2 = Var(X) = \Sigma(x-\mu)^2 f(x)$
$E(X^2) = \Sigma x^2 \mathrm{P}(\mathrm{X} = x) = \Sigma x^2 f(x)$	$\sigma^2 = E(X^2) - [E(X)]^2$

Bernoulli Random Variables

X	PDF	Notation
$X = 1$, if "success" $X = 0$, if "failure"	$f(x) = p^x(1-p)^{1-x}$	$X \sim B(p)$

Mean	Variance
$E(X) = p$	$Var(X) = p(1-p)$

Binomial Random Variables

X	PDF	Notation
Number of "successes" in a Bernoulli process of n trials	$f(x) = \binom{n}{x} p^x(1-p)^{n-x}\ 0 \le x \le n$	$X \sim B(n, p)$

Mean	Variance
$E(X) = np$	$Var(X) = np(1-p)$

Poisson Random Variables

X	PDF	Notation
Number of certain events occurring in a time interval or region	$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}\ 0 \le x \le n$	$X \sim P(\lambda)$

Mean	Variance
$E(X) = \lambda$	$Var(X) = \lambda$

Y	Notation
Number of certain events occurring in t units of time	$Y \sim Poisson(\lambda t)$
Mean	**Variance**
$E(Y) = \lambda t$	$Var(Y) = \lambda t$

Continuous Random Variables

Mean	Variance
$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$	$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$
$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx$	$Var(X) = E(X^2) - [E(X)]^2$

Calculating the CDF of a PDF

If y is in $[-\infty, a]$, then	If y is in $[a, b]$, then	If y is in $[b, \infty]$, then
$\begin{aligned} F(y) &= \int_{-\infty}^{y} f(x)\,dx \\ &= 0 \end{aligned}$	$\begin{aligned} F(y) &= \int_{-\infty}^{y} f(x)\,dx \\ &= \int_{-\infty}^{a} f(x)\,dx + \int_{a}^{y} f(x)\,dx \\ &= 0 + \int_{a}^{y} f(x)\,dx \\ &= [F(x)]_a^y \\ &= F(y) - F(a) \end{aligned}$	$\begin{aligned} F(y) &= \int_{-\infty}^{y} f(x)\,dx \\ &= \int_{-\infty}^{a} f(x)\,dx + \int_{a}^{b} f(x)\,dx + \int_{b}^{y} f(x)\,dx \\ &= 0 + \int_{a}^{b} f(x)\,dx + 0 \\ &= [F(x)]_a^b \\ &= F(b) - F(a) \\ &= 1 \end{aligned}$

Normal Distribution

X		Notation
…is considered to have Normal Distribution if its PDF has the form of the Normal Probability Density Function		$X \sim N(\mu, \sigma^2)$
Mean	**Variance**	**Standard Score**
$E(X) = \mu$	$Var(X) = \sigma^2$	$z = \frac{x - \mu}{\sigma}$

Lecture 3: Sampling Distribution & Estimation

Mean and Standard Deviation of a Sample Mean

- If $\bar{x}$ is the mean of an SRS of size n from a population with mean μ and standard deviation σ, then:

$\mu_{\bar{x}} = \mu$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	${\sigma^2}_{\bar{x}} = \frac{\sigma^2}{n}$

Distribution of $\bar{x}$

- If a population X has $N(\mu, \sigma^2)$ distribution, then, for the sample mean, $\bar{x}$, of n independent values…

$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$	$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$

General Confidence Interval for μ

- The 100(1 - α)% CI for μ when σ is known is given by:

$\bar{x} \pm E$	$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$\alpha = 1 - \frac{CI}{100}$

- The general conclusion is written as "We are approximately CI confident that the population mean lies between $\bar{x} - E$ and $\bar{x} + E$"

Lecture 4: Estimation, Hypothesis & Testing

Selecting Sample Size

- We can select the sample size, n, that will guarantee a desired confidence level for a fixed margin of error, E.

$E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$	$\sqrt{n} = \frac{z_{\alpha/2}\sigma}{E}$	$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$

- Note: round up to the nearest whole.

Test Statistic for Samples with Known Population Standard Deviations

- To test the validity of an alternative hypothesis H_A we standardise $\bar{x}$ and obtain z-score that tells us how many standard errors $\bar{x}$ is from μ

$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$	μ: Supposed population mean denoted by H_0	σ: Known population standard deviation

Test Statistics Samples with Unknown Population Standard Deviations

$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$	μ: Supposed population mean denoted by H_0	s: Known sample standard deviation

Confidence Intervals for Samples with Unknown Population Standard Deviations

$\bar{x} \pm E$	$E = t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$	$\alpha = 1 - \frac{CI}{100}$

When to Use Standard Scores and Standard Errors

- If σ is given and $n \geq 30$, then we should use z-scores, otherwise, t-scores
- If s is given and $n \geq 100$, then we should use z-scores, otherwise, t-scores

P-Values

- From the tables, we find the probability of obtaining a value or more extreme than the test statistic. This probability is referred to as the p-value
- If the p-value is relatively small, we say that we have enough evidence to reject H_0. i.e. there is a relatively low probability that H_0 is true, so we are willing to promote the H_A

Conclusions at the α Level of Significance

- "Since the p-value of _____ is more / less than α at the (100α)% level of significance, we should reject H_0 and state that, on average, the _____ _____ more / less than μ"

Validity of Conclusions

- For our conclusions to be valid, the chosen sample should…
    1. Be approximately normal
    2. Not be skewed
    3. Not contain any outliers