Descriptive Statistics

Let n denote the number of sample observations and N denote the number of population observations.

Statistic	Sample Formula	Population Formula
Mean	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\mu_X = \frac{1}{N}\sum_{i=1}^{N} X_i$
Variance	$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$	$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)^2$
St. Dev.	$\sqrt{S_X^2}$	$\sqrt{\sigma_X^2}$
Covariance	$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$	$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)$
Correlation	$r_{XY} = \frac{S_{XY}}{S_X S_Y}$	$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

Frequency and Statistics from Grouped Data

Let n denote the total number of observations and n_k denote the number of observations in category k, relative frequency $f_k = n_k/n$. There are J groups or bins.

Statistic	Sample Formula	Population Formula
Mean	$\bar{X} = \sum_{k=1}^{J} f_k X_k$	$\mu = \sum_{k=1}^{J} f_k X_k$
Variance	$S_X^2 = \frac{n}{n-1}\sum_{k=J}^{J} f_k(X_k - \bar{X})^2$	$\sum_{k=1}^{J} f_k(X_k - \mu_X)^2$

Weighted Mean

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

where ω_i is the weight of the i^{th} observation.

Z-Score

$$Z = \frac{X - \mu}{\sigma}$$

Single Random Variable (RV)

Quantity	Discrete RV	Continuous RV
Probability/frequency	$f(x) = Pr(X = x)$	$Pr(a \leq X \leq b) = \int_a^b f(x)\partial x$
Cumulative prob., $Pr(X \leq x_0)$	$\sum_{x \leq x_0} f(x)$	$\int_{-\infty}^{x_0} f(x)\partial x$
Expected value, $E(X)$ or μ_X	$\sum_x xf(x)$	$\int_x xf(x)$ dx

$$Var(X) = \sigma_X^2 = E[(X - \mu)^2]$$

Multiple Random Variables

Conditional probability (discrete):

$$f(y|x) = Pr(Y = y|X = x) = \frac{Pr(X = x, Y = y)}{Pr(X = x)} = \frac{f(x, y)}{f(x)}$$

Conditional expectation (discrete):

$$E(Y|X = x) = \sum_y yPr(Y = y|X = x) = \sum_y yf(y|x)$$

Covariance and correlation:

$$\sigma_{XY} = Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$\rho_{XY} = corr(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad \text{where } -1 \leq \rho \leq 1$$

Two random variables are *uncorrelated* if $\rho_{XY} = 0$.

Two random variables are *independent* if $f(y|x) = f(y)$ for all x and y or equivalently $E(Y|X) = E(Y)$.

Sample Mean Distribution

$$E(\bar{X}) = \mu \qquad Var(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

Sample mean is normally distributed:

1. When the underlying population is normal, **or**
2. If the sample size is large, say $n \geq 100$ by the Central Limit Theorem (CLT)

Confidence Intervals

Known population variance:
$1 - \alpha$ confidence interval for the population mean μ:

$$\bar{x} \pm \underbrace{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\text{Margin of Error}}$$

where $z_{\alpha/2}$ is the z-value that leaves area $\alpha/2$ in the upper tail of the standard normal distribution.

Unknown population variance:
$1 - \alpha$ confidence interval for the population mean μ:

$$\bar{x} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

where $t_{n-1,\alpha/2}$ is the t-value that leaves area $\alpha/2$ in the upper tail of the t-distribution. $n - 1$ is the degrees of freedom. Since t distribution looks just like the standard normal for large n, for $n \geq 100$ continue using the standard normal table.

Hypothesis Testing

Test null hypothesis $H_0 : \mu = \mu_0$ against alternative hypothesis $H_1 : \mu \neq \mu_0$. Construct test statistic Z if true population variance is known, else use T-statistic.

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \quad \text{and} \quad t_0 = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$

Under the null if $\bar{X} \sim N(\mu_0, \sigma^2/n)$, then $Z \sim N(0, 1)$ and $T \sim t_{n-1}$. In case of known population variance, reject the null if $|z_0| > z_{\alpha/2}$. In the case of unknown population variance, reject the null if $|t_0| > t_{n-1,\alpha/2}$. When $n \geq 100$ you can reject the null if $|t_0| > z_{\alpha/2}$.

---

p-value:

Known variance: $p = 2Pr(Z > |z_0|)$
Unknown variance, $n < 100$: $p = 2Pr(T > |t_0|)$
Unknown variance, $n \geq 100$: $p = 2Pr(Z > |t_0|)$

Note: *You will not need to refer to the t-table for the exam, only the standard normal table.*